https://dx.doi.org/10.20880/kler.2026.61.1.103

전이(Transfer) 평가를 위한 세트형 서·논술형 과제 설계 체계 연구

— 구인·과제·증거의 정렬(Alignment) 메커니즘을 중심으로

박고운 한국교원대학교 국어교육 박사과정

I. 서론

2022 개정 교육과정은 역량 중심 교육의 심화를 표방하며, 단편적 지식의 재생산을 넘어 자료를 비판적으로 해석·평가하고 근거를 조직하여 공적 담화를 구성하는 고차적 수행을 강조한다(교육부, 2022). 국어과 '독서와 작문'은 이러한 지향을 교과 구조 차원에서 구체화한 대표적 맥락으로, 다자료 기반 이해와 판단, 읽기–쓰기 통합 수행, 논증적 표현을 통해 지식을 재구성하고 새로운 맥락으로 적용하는 전이(transfer) 능력을 요구한다(교육부, 2022; 최숙기, 2021; 최숙기·박종임, 2023ㄱ). 따라서 선다형 중심 평가가 포착하기 어려운 고차 사고와 전이 수행을 학교 평가에서 직접 관찰할 수 있는 서·논술형 평가의 역할이 중요해지고 있다(김경희, 2020; 장성민, 2025).

그러나 서·논술형 평가의 정책적 확대에도 불구하고 학교 현장에서는 채점 공정성 논란과 업무 부담이 지속되며, 그 결과 사고의 심화를 유도하기보다 정답이 비교적 명확한 단답형 서술로 회귀하거나 수행을 형식화하는 '최적화 행동'이 반복된다(김경희·이명진, 2021; 남가영·김호정, 2023; 박종임, 2024). 이러한 현상은 교사 개인의 역량 문제로만 환원되기 어렵다. 한국교

육과정평가원의 최근 연구는 서·논술형 평가의 양적 확대에도 불구하고 실제 학교 현장에서는 평가가 단답화·형식화되는 경향이 지속되며, 이것이 평가 설계가 학습자의 수행을 안정적으로 관찰·판정하기 어렵게 만드는 구조적 조건과 결부되어 있음을 확인하였다(김수진·김희경·나우열 외, 2025). 교사 대상 질적 연구들 역시 평가 설계 과정에서 '무엇을 측정할 것인가'에 대한 합의는 존재하나, '어떻게 설계하고 점검할 것인가'에 대한 구체적 절차 지식이 부족함을 반복적으로 지적해왔다(김형성, 2024; 박종임, 2024). 이는 서·논술형 평가의 질 제고가 개별 교사의 역량 문제가 아니라 설계 체계 수준의 지원을 요구하는 구조적 문제임을 시사한다.

본 연구는 이 구조적 결함을 평가의 구성요소가 하나의 목표를 향해 정합적으로 연결되지 못하는 문제로 파악한다. 평가가 포착하려는 구인(construct)은 관찰 가능한 증거(evidence)로 전환되어 과제(task)를 통해 산출되어야 하며, 산출된 증거는 루브릭(rubric)에 의해 판정되어야 한다. 이 대응 관계가 느슨해지면 정렬(alignment)이 흔들리고, 고차 수행을 겨냥할수록 공정성 불신이 강화되는 역설이 나타난다(김경희·이명진, 2021; Mislevy, Steinberg, & Almond, 2003). 본 연구는 이러한 상태를 '정렬 붕괴(alignment collapse)'로 개념화하고, 서·논술형 평가의 반복되는 난맥을 설계 문제로 진단하고자 한다(김수진 외, 2025; 박혜영·김성숙·김경희 외, 2019). 이하에서는 이를 본 연구의 분석어로 사용하여, 구인–증거–과제–루브릭의 대응 관계가 약화되며 나타나는 설계 수준의 정합성 약화 상태를 지칭한다.

주목할 점은 이 문제가 서·논술형 평가에 필요한 요소의 '부재'에서 비롯되지 않는다는 점이다. 선행연구는 고차 사고, 자료 기반 판단, 지식의 재구성 및 전이를 목표로 제시해 왔고, 성취기준 정합성, 실제적 맥락 제공, 충분한 자료 제시, 분석적 루브릭 개발 등을 설계 조건으로 누적해 왔다(김경희, 2020; 박혜영 외, 2019). 그럼에도 실행이 흔들리는 이유는 "무엇을 측정할 것인가"의 합의 부족이라기보다, 합의된 요소들을 "어떤 규칙과 절차

로 정렬하여 과제로 구현하고, 정렬의 성공 여부를 어떻게 점검할 것인가"에 대한 설계 지식이 충분히 명료화되지 않았기 때문이다(김수진 외, 2025; 박종임, 2024). 따라서 서·논술형 평가 논의는 구성요소의 나열을 넘어 정렬 붕괴를 예방하는 설계 체계(design framework) 수준으로 이동할 필요가 있다.

이에 본 연구는 서·논술형 평가를 '문항 유형'의 문제가 아니라 '정렬 가능한 설계 체계'의 문제로 재정식화한다. 전이 수행은 맥락 변화 속에서 개념이 재적용되고 조정되는 과정에서 드러나므로, 단일 문항의 결과물만으로 수행의 전개를 안정적으로 포착하기 어렵고(송슬기, 2024; Bransford & Schwartz, 1999), 단일 산출 중심 설계는 '아는 것을 쓰는' 수준의 산출(knowledge telling)로 수렴하여 지식의 재구성과 논증적 정교화(knowledge transforming)를 관찰하기 어렵다(Scardamalia & Bereiter, 1987). 따라서 평가 도구는 학습자의 사고 전개가 누적되도록 과정적 구조(sequential structure)를 갖추어야 하며, 본 연구는 이를 위한 설계·해석 단위로서 세트형 서·논술형 과제 구조를 제안한다. 세트형 구조는 단순한 문항 묶음이 아니라 수행 요구를 단계적으로 조직하여 사고 전개 경로가 축적되도록 하는 설계 장치이다.

본 연구는 정렬 붕괴를 예방하는 설계 논리를 제시하기 위해 증거중심설계(Evidence-Centered Design, ECD)와 개념기반 탐구학습(Concept-Based Inquiry Learning, CBIL)의 논리를 결합한다. ECD는 구인–증거–과제–루브릭의 정렬 절차를 제공하여 채점이 사후적 감각이 아니라 설계된 증거 수집임을 논증하는 기반을 마련하며(Mislevy et al., 2003), CBIL의 '사실(자료)–개념(원리)–일반화(전이)' 경로는 과제 내부에서 사고가 심화되는 흐름을 조직하는 원리로 활용될 수 있다(송슬기, 2025; Erickson, Lanning, & French, 2017/2019). 또한 전이 수행이 평가에서 무엇으로 관찰되는지(증거화)를 검토할 때 학문 문식성(disciplinary literacy) 논의가 제시해 온 수행의 특징을 참고한다(김영란, 2021; 편지윤, 2021;

Shanahan & Shanahan, 2008). 각 이론의 역할과 통합의 필연성은 Ⅱ장에서 논증한다.

본 연구는 설계 프레임워크의 이론적 타당성과 작동 가능성(feasibility) 탐색에 초점을 둔 이론적 설계 제안(design proposition) 연구이다. 학습자 수행 자료나 전문가 평정에 기반한 경험적 타당화, 학습 효과 검증, 채점자 간 신뢰도 확보 등은 본 연구의 범위를 넘어서며 후속 연구 과제로 남겨 둔다.

본 연구의 고유한 기여는 다음과 같다. 첫째, 서·논술형 평가의 반복적 난맥을 단순한 실행상의 문제가 아니라 구인–증거–과제–루브릭 간 대응 약화로 인한 '정렬 붕괴(alignment collapse)'의 문제로 재개념화하였다는 점이다. 둘째, ECD·CBIL·학문 문식성을 병렬적으로 소개하는 데 머물지 않고, 이를 Layer 1의 구인 운영화, Layer 2의 과제 구조화, Layer 3의 증거 판정으로 연결되는 3층위 설계 프레임워크로 재구성하였다는 점이다. 셋째, 전이를 평가 장면에서 직접 판정 가능한 수행으로 전환하기 위해 '조건부 근거 판단'을 운영화 구인으로 제안하였다는 점이다. 다만 본 연구가 다루는 전이는 전이 일반 전체를 포괄하는 범용 모형이 아니라, 자료 비평과 논증적 판단이 결합된 맥락에서 중등 이상 학습자의 판단 전이를 설계 가능한 수행으로 번역하기 위한 제한적 모형이다.

본 연구의 Ⅱ장은 정렬 붕괴의 진단과 통합 설계 논리를 제시하고, Ⅲ장은 세트형 설계 프레임워크와 작동 메커니즘을 제안한다. Ⅳ장은 2022 개정 '독서와 작문' 맥락에서의 구체적 설계 사례를 제시하며, Ⅴ장은 연구의 기여와 한계, 후속 연구 과제를 논의한다.

II. 정렬 붕괴의 진단과 통합 설계 논리

본 장은 서·논술형 평가가 반복적으로 단답화·형식화되는 현상을 '정

렬 붕괴'라는 구조적 문제로 진단하고, 이를 복원하기 위한 통합 설계 논리를 논증하는 데 목적이 있다. 먼저 이 현상을 타당도 논증 구조의 해체로 개념화하고(Ⅱ장 1절), 이어서 측정 논리·인지 논리·증거화 원리라는 세 가지 이론적 조건이 왜 개별적으로는 불충분하며 통합이 논리적으로 필연적인지를 밝힌다(Ⅱ장 2절). 본 장은 Ⅲ장에서 제시될 3층위 설계 체계(Layer 1 구인의 운영화–Layer 2 과제의 구조화–Layer 3 증거의 판정)의 이론적 정당성을 확보하는 것을 목적으로 하며, 구체적 설계 절차와 과제 사례는 다음 장에서 다룬다.

1. 정렬 붕괴의 구조

1) 역량 평가의 필요조건과 서·논술형 평가

앞서 확인한 바와 같이, 2022 개정 교육과정과 '독서와 작문'은 전이 수행을 핵심 목표로 요청한다(교육부, 2022; 최숙기, 2021; 최숙기·박종임, 2023ㄱ). 문제는 전이가 단일 정답 산출로 환원되기 어려우므로, 평가가 전이를 주장하려면 학습자가 자료를 해석하며 개념을 적용·조정하고 그 조정을 정당화하는 과정이 응답에서 관찰 가능해야 한다는 점이다(송슬기, 2024). 따라서 서·논술형 평가는 선다형 중심 평가가 포착하기 어려운 수행을 직접 관찰하는 방식으로 요청된다(김경희, 2020; 장성민, 2025).

또한 전이 수행은 영역 고유의 기준을 활용해 판단을 갱신하고 공적 설명을 조직하는 읽기·쓰기 관행과 연결되며, 학문 문식성 논의는 이를 평가에서 관찰 가능한 수행의 표지로 정교화하는 관점을 제공해 왔다(김영란, 2021; 편지윤, 2021; Shanahan & Shanahan, 2008). 따라서 전이를 평가한다는 것은 결과의 적절성 확인을 넘어, 수행이 드러나도록 증거를 설계하고 판정하는 문제로 전환된다.

2) 현장 실행의 구조적 딜레마

그러나 학교 현장에서는 평가의 질적 제고가 충분히 이루어지지 못하고 있다. 교사들은 서·논술형 평가를 도입하더라도 채점 공정성 우려와 업무 부담으로 인해 문항 요구를 단순화하고 채점 기준을 축소하는 경향이 있으며(박종임, 2024), 이러한 단답화·형식화는 실행자의 무성의가 아니라 제도적 조건 하에서 안정성을 확보하기 위한 합리적 선택으로 해석될 수 있다(남가영·김호정, 2023). 역설적으로, 고차 수행을 겨냥할수록 답안의 다양성과 해석 여지가 증가하여 공정성 불신이 커지고, 공정성을 강화하려고 기준을 세분화하고 절차를 표준화할수록 평가가 다시 지식 인출이나 서식 채우기로 회귀하여 본래의 수행 목표를 약화시킨다(김수진 외, 2025; 박종임, 2024).

이러한 딜레마는 교사 개인의 역량 문제로 환원되기 어렵다. 문항의 질은 성취기준 정합성과 수행 관찰 가능성에 달려 있으나(정민주·서수현·남민우 외, 2022), 이러한 조건을 충족시키는 설계 절차와 정렬 기준은 충분히 명료화되지 않았다. 성취기준, 실제적 맥락 제공, 충분한 자료 제시, 분석적 루브릭 개발 등이 설계 조건으로 제시되어 왔으나(박혜영 외, 2019), 이들 요소가 실제로 어떤 규칙과 절차로 정렬되어야 하는지에 대한 설계 지식은 제공되지 못했다. 평가 문항 개발과 채점 기준 설정의 어려움은 개인 역량을 넘어 구조적 지원의 문제와 결부되어 있다(김형성, 2024).

3) 정렬 붕괴: 타당도 논증 구조의 해체

본 연구는 이 현상을 '정렬 붕괴'로 개념화한다. 타당도는 측정 도구 자체의 속성이 아니라 측정 결과의 해석과 사용에 대한 논증이며(Messick, 1989), 타당도 논증은 점수 생성에서 일반화, 외삽, 함의로 이어지는 추론 사슬의 각 단계를 명시적으로 검토해야 한다(Kane, 2013). 평가가 타당하다는 것은 구인에서 출발하여 증거, 과제, 채점으로 이어지는 각 단계가 논리적으로 연결되고 각 연결이 검증 가능해야 함을 의미한다(American Education-

al Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

정렬은 구인–증거–과제–루브릭의 대응 관계를 뜻하며, 무엇을 측정하려는지(구인)가 무엇으로 관찰되고(증거), 어떤 과제 요구로 산출되며(과제), 어떤 규칙으로 판정되는지(루브릭)가 연역적으로 연결되는 상태를 가리킨다. 정렬 붕괴란 이 대응 관계가 느슨해져 고차 수행을 목표로 할수록 오히려 타당도 논증이 취약해지는 구조를 말한다. 예컨대 평가 문항이 비판적 사고를 목표로 표방하더라도 실제 채점 기준이 특정 정보의 포함 여부로 환원되는 경우, 구인과 증거·판정의 연결이 약화된다(김경희, 2020).

단답화·형식화는 정렬 붕괴에 대한 '증상 관리'로 기능한다. 교사들은 타당도 논증이 취약한 상태에서 평가를 실행해야 하는 압력 속에서 논증 사슬을 복원하기보다, 평가 요구를 단순화하여 불확실성을 줄이는 전략을 선택한다. 문제의 핵심은 무엇을 측정할 것인가의 합의 부족이 아니라, 합의된 요소들을 어떤 규칙과 절차로 정렬하여 과제로 구현하고 정렬의 성공 여부를 점검할 것인가에 대한 설계 지식이 부재하다는 점에 있다(김수진 외, 2025). 따라서 필요한 것은 구인에서 채점까지의 각 단계를 명시적으로 연결하고 그 연결을 점검할 수 있는 설계 체계이다.

2. 정렬 복원을 위한 통합 설계 논리

정렬 붕괴를 복원하기 위해서는 (1) 구인–증거–과제–루브릭을 논증 구조로 연결하는 측정 논리, (2) 학습자의 사고가 단계적으로 심화되도록 과제를 조직하는 인지 논리, (3) 전이와 같은 추상적 구인을 관찰 가능한 담화적 증거로 전환하는 증거화 원리가 함께 요구된다. 이하에서는 각 논리가 제공하는 기여와 함께 독립적으로는 불충분한 이유를 밝히고, 통합이 논리적으로 필연적임을 논증한다.

1) 측정 논리: 증거중심설계(ECD)

ECD는 평가를 구인에 대한 주장을 증거로 뒷받침하는 논증 구조로 이해하며, 구인 모형에서 증거 모형으로, 다시 과제 모형으로 이어지는 연역적 관계를 통해 타당도 논증 사슬을 절차화한다(Mislevy et al., 2003). 성취기준–문항–채점 기준 간 정합성이라는 원칙은 중요하지만(박혜영 외, 2019), 그 정합성을 설계 단계에서 어떻게 구현하고 점검할 것인지에 대한 절차는 충분히 구체화되지 못했다. ECD는 바로 이 설계 지식의 공백을 메우는 틀로 기능한다.

무엇을 보려는가와 무엇을 실제로 보는가 사이의 간극을 좁히기 위해서는 구인을 측정 가능한 하위 요소로 분해하고 각 요소가 답안에서 어떤 형태로 관찰되는지를 사전에 명시해야 하며(김경희, 2020), 고차 사고를 측정한다고 주장하면서 실제로는 표면적 정보 재생산만을 판정하는 모순을 방지하려면 채점 기준이 구인 정의와 대응하는지 설계 단계에서 검토해야 한다(김수진 외, 2025). 다만 ECD는 측정 논리를 정교하게 제공하더라도, 학습자의 사고가 과제 내에서 어떤 인지 경로를 따라 점진적으로 심화되는지에 대해서는 상대적으로 제한적인 설명을 제공한다. 서·논술형 평가의 난점은 채점 기준의 부재뿐 아니라, 사고 전개가 드러나도록 과제 구조를 설계하는 데에도 존재한다(송슬기, 2024).

2) 인지 논리: 개념기반 탐구학습(CBIL)

이 공백을 보완하기 위해 사고가 단계적으로 심화되도록 과제를 조직하는 인지 논리가 필요하다. CBIL은 학습이 사실의 축적에서 개념의 이해를 거쳐 일반화와 적용으로 심화되는 인지 경로를 따른다고 본다(Erickson et al., 2017/2019). 이는 핵심 개념을 중심으로 학습 경험이 점진적으로 추상화되고 확장된다는 나선형 교육과정 원리를 계승한 것이다(Bruner, 1960).

평가가 사실 인출에 머물지 않고 개념 이해와 전이를 포착하려면 사고의 단계적 누적이 가능한 구조여야 한다(Erickson et al., 2017/2019). 단

일 문항으로 복합 수행을 요구할 때 학습자는 사고의 출발점과 경로를 구성하기 어려워하며, 단일 산출 과제는 '아는 것을 쓰는' 수준의 산출로 수렴하기 쉬워 지식의 재구성과 논증적 정교화가 드러나기 어렵다(Scardamalia & Bereiter, 1987). CBIL의 단계적 심화 경로는 학습 설계뿐 아니라 평가 과제 배열에도 적용 가능하다. 즉, 사실 식별 → 개념 적용 → 일반화/전이로 이어지는 인지 자원의 단계적 재활용 구조를 제공할 때, 학습자는 초기 문항에서 형성한 개념적 틀을 후속 문항에서 자율적으로 활용할 수 있게 된다(Fisher & Frey, 2013). 이는 본 연구에서 제안하는 세트형 구조의 인지적 정당성을 제공하는 근거가 된다. 다만 CBIL은 사고가 심화되는 경로를 제시하더라도, 평가 장면에서 그 사고가 무엇으로 관찰·판정되어야 하는지(증거화·판정 논리)에 대해서는 충분히 설명하지 않는다. 따라서 CBIL의 인지 경로는 담화적 수행 증거로 번역될 필요가 있다.

3) 증거화 원리: 학문 문식성

이를 위해 전이와 같은 추상적 구인을 관찰 가능한 담화적 증거로 전환하는 증거화 원리가 필요하다. 학문 문식성 논의는 읽기와 쓰기가 영역 중립적 기능이 아니라 특정 학문 영역의 지식 생산 방식과 타당성 기준에 따라 달라지는 실천임을 강조한다(Shanahan & Shanahan, 2008). 예컨대 역사 영역에서는 출처 확인, 맥락화, 상호 검증이, 과학 영역에서는 변인 통제와 조건–결과 연결의 명시가 핵심 추론 절차로 작동하므로(Wineburg, 1991; Fang, 2012), 학문 문식성 평가는 학습자에게 이러한 영역 고유의 추론 절차를 실제로 수행하도록 요구해야 한다. 이는 자료의 조건 정보를 식별하고 그 조건이 결론에 미치는 영향을 추론하며 그 추론을 정당화하는 담화 구성으로 드러난다. 다만 학문 문식성 논의는 '무엇을 관찰할 것인가'를 정교화하는 데 강점이 있으나, 그 증거를 안정적으로 산출하도록 과제를 조직하고 일관되게 판정하는 설계 규칙을 직접 제공하지는 않는다. 따라서 증거화 원리는 ECD의 측정 논리 및 CBIL의 인지 논리와 결합될 필요가 있다.

다만 이 결합이 실질적으로 작동하려면, 학문 문식성의 증거 표지는 선험적으로 고정된 목록이 아니라 각 영역의 인식론적 관행에 따라 도출 절차를 거쳐 구체화되어야 한다. 따라서 증거 표지의 생성 규칙은 ① 해당 영역에서 타당성 판단에 사용되는 조건·규범을 확인하고, ② 그 관행의 언어적 실현 방식을 분석하며, ③ 이를 관찰 가능한 담화 증거와 과제 최소 요구로 번역하는 절차를 거쳐 도출될 필요가 있다. 본 연구의 Ⅳ장 사례는 사회과학적 자료 읽기 맥락에서 이러한 절차를 제한적으로 구체화한 예시로 이해할 수 있다(Shanahan & Shanahan, 2008).

요컨대 ECD는 구인 – 증거 – 과제 – 루브릭의 정렬 논리를 제공하지만, 학습자의 사고가 과제 내부에서 어떤 인지 경로를 따라 단계적으로 심화되어야 하는지까지 직접 규정하지는 않는다. 반대로 CBIL은 사실 – 개념 – 일반화/전이로 이어지는 인지 심화의 경로를 제공하지만, 그러한 사고를 평가 장면에서 무엇으로 관찰하고 어떤 규칙으로 판정할 것인지에 대한 증거화·판정 절차를 완결하지는 않는다. 학문 문식성은 영역 고유의 추론 절차와 담화 표지를 정교화함으로써 '무엇을 증거로 볼 것인가'를 구체화하는 강점을 지니지만, 그 자체로 과제를 조직하고 수준 차이를 일관되게 판정하는 설계 절차를 제공하지는 않는다. 따라서 세 이론은 서로를 대체하는 관계가 아니라 서로 다른 설계 레이어의 공백을 메우는 보완 관계에 있으며, 본 연구의 3층위 프레임워크는 바로 이 교차 지점의 공백을 메우기 위해 구성된 재구성 체계이다.

3. 설계 실행 체계

앞 절에서 논증한 바와 같이, ECD·CBIL·학문 문식성의 통합은 정렬 복원을 위한 이론적 기반을 제공한다. 그러나 이러한 통합이 실제 평가 설계로 작동하기 위해서는 추상 개념을 설계에서 적용 가능한 요소로 전환하는 과정이 필요하다. 본 연구는 이 전환 작업 자체를 핵심 재구성 지점으로 보

고 이를 '운영화(operationalization)'로 규정한다. 운영화는 추상적 개념을 설계 가능한 형태(결정 규칙·산출물·점검 기준)로 번역하는 작업을 의미한다(Hevner, March, Park, et al., 2004).

운영화의 관점에서 볼 때, Ⅱ장 2절의 세 핵심 이론은 정렬의 '골격'을 제공하지만, 설계가 단답화·형식화로 회귀하지 않도록 통제하기 위해서는 설계 실행의 미시적 원리들이 필요하다. 본 절에서 제시하는 '설계 실행 원리'는 핵심 이론을 대체하는 독립 이론이 아니라, 핵심 이론이 제시하는 정렬 구조가 실제 과제 요소(지문, 문항, 루브릭)로 번역되는 과정에서 결정의 기준을 제공하는 운영화 메커니즘을 뜻한다. 이하에서는 설계 실행 원리를 개념과 필요성의 수준에서만 정리하고, 구체적 설계 프레임워크와 작동 메커니즘은 다음 장에서 제시한다. 이때 다음 장의 3층위 체계는 이론 자원을 병렬적으로 반복 제시하는 것이 아니라, 각 이론이 제공하는 기능과 남는 공백을 설계 산출물 수준에서 재배치한 결과로 이해되어야 한다.

1) 단계화된 책임 이전: '단일 산출의 한계'를 통제하는 문항 연계 원리

비계(scaffolding)는 학습자가 단독으로 수행하기 어려운 과제에 대해 일시적 지원을 제공하되, 지원이 점진적으로 소거되고 수행의 책임이 학습자에게 이전되는 방식으로 작동한다(Wood, Bruner, & Ross, 1976; van de Pol, Volman, & Beishuizen, 2010). 단계화된 책임 이전은 이 원리를 "학습의 진행"뿐 아니라 "평가 과제의 배열"에도 적용 가능한 논리로 재해석한 것으로, 초기에는 사고 경로를 가시화하는 구조화된 요구를 제공하고 후속에서는 동일한 인지 자원을 학습자가 자율적으로 조직·정당화하도록 요구를 조정하는 방향을 제시한다(Fisher & Frey, 2013). 서·논술형 평가에서 이는 단일 문항이 복합 수행을 한 번에 요구할 때 발생하는 경로 불명확성과 산출의 피상화를 설계 차원에서 통제하기 위한 원리로 기능한다(Scardamalia & Bereiter, 1987). 이 원리는 Ⅲ장에서 Layer 2의 문항 간 인지 연계 원리를 구성하는 준거로 전환된다.

2) 생산적 마찰: '조건–결론 제약 추론'을 필수화하는 인지 부담 조정 원리

전이 수행은 개념의 단순 적용이 아니라 맥락 변화 속에서 개념을 재조정(adaptation)하고 그 조정을 정당화하는 과정에서 드러난다(Bransford & Schwartz, 1999). 따라서 평가 과제가 지나치게 평탄하면 학습자는 개념을 기계적으로 대입하거나 표면 정보 재진술로 회귀할 수 있으며, 반대로 과제가 과도하게 난삽하면 추론을 포기하고 형식적 답안으로 수렴할 가능성이 커진다. 생산적 마찰은 이러한 양극단을 피하면서 조건 분석과 제약 추론이 응답에서 필수적으로 나타나도록 과제의 인지 부담을 조정하는 설계 원리이다(Bjork & Bjork, 2011; Kapur, 2008; Kapur & Bielaczyc, 2012). 본 연구에서는 생산적 마찰을 "혼란의 증대"가 아니라 "조건–결론 제약 추론의 필수화"로 개념화하여 사용하며, 이 원리는 Ⅲ장에서 Layer 2의 지문 기능 분업 및 요구 배치 메커니즘을 구성하는 준거로 전환된다.

3) 분석적 채점: '타당도 논증'을 지지하는 판정 규칙의 원리

서·논술형 평가에서 채점자 간 신뢰도는 타당도 논증의 핵심 조건이며(American Educational Research Association et al., 2014), 고차 수행일수록 답안 다양성이 증가하여 판정의 일관성을 확보하기 어렵다(Jonsson & Svingby, 2007). 분석적 루브릭은 답안을 단일 점수로 환원하기보다 복수 준거를 분리하여 점검함으로써, 판정의 기준을 구인에 더 밀착시키고 채점 일관성을 확보하는 방식으로 논의되어 왔다(Brookhart, 2013/2022; Panadero & Jonsson, 2013). 본 연구에서 분석적 채점은 '정렬 붕괴'가 채점 단계에서 재발하는 것을 통제하기 위한 원리로 위치하며, 이 원리는 Ⅲ장에서 Layer 3의 판정 규칙과 수준 경계 설정을 구성하는 준거로 전환된다.

이상의 논의를 요약하면, ECD·CBIL·학문 문식성은 서로를 대체하는 이론이 아니라, 서·논술형 평가 설계에서 서로 다른 층위의 공백을 메우는 상보적 이론 자원이다. 본 연구는 이들을 병렬적으로 절충하거나 나열하는 데 머물지 않고, 정렬 붕괴를 통제하기 위한 3층위 설계체계의 논리 안에

서 기능별로 재배치한다. 즉, 본 연구의 관심은 개별 이론의 단순 결합이 아니라, 각 이론의 기여와 한계를 교차 검토하여 서·논술형 평가 설계에 필요한 최소 설계 논리로 재구성하는 데 있다. 〈표 1〉은 이러한 재구성의 관점에서 각 이론의 핵심 기여와 한계, 그리고 3층위 설계체계 안에서의 재구성 원리를 정리한 것이다. 다음 장에서는 이 재구성 논리가 실제 설계 산출물 수준에서 어떻게 구체화되는지를 제시한다.

〈표 1〉 3층위 설계체계 구성을 위한 이론별 핵심 기여, 한계, 재구성 원리

이론	핵심 기여	한계	재구성 원리	적용 층위
ECD	구인 – 증거 – 과제 – 판정 간 정렬을 타당도 논증의 구조로 조직하는 틀을 제공함	사고의 단계적 심화나 세트형 과제 배열의 원리를 직접 제시하지는 않음	구인의 운영화와 증거 판정의 기준을 정렬하는 설계 논리로 재구성함	Layer 1, Layer 3
CBIL	사실 – 개념 – 일반화/전이로 이어지는 인지 심화의 경로를 제시함	인지 경로가 평가 장면에서 어떤 증거로 관찰되고 어떻게 판정되어야 하는지는 직접 제시하지 않음	지문 기능의 분화와 문항 간 단계적 연계 원리를 구성하는 설계 논리로 재구성함	Layer 2
학문 문식성	영역 고유의 추론 절차와 담화 수행의 특징을 구체화함	과제 조직과 수준 판정을 일관되게 연결하는 설계 절차로는 충분히 조직되어 있지 않음	운영화 구인과 판정 가능한 담화적 증거를 정교화하는 설계 자원으로 재구성함	Layer 1, Layer 3

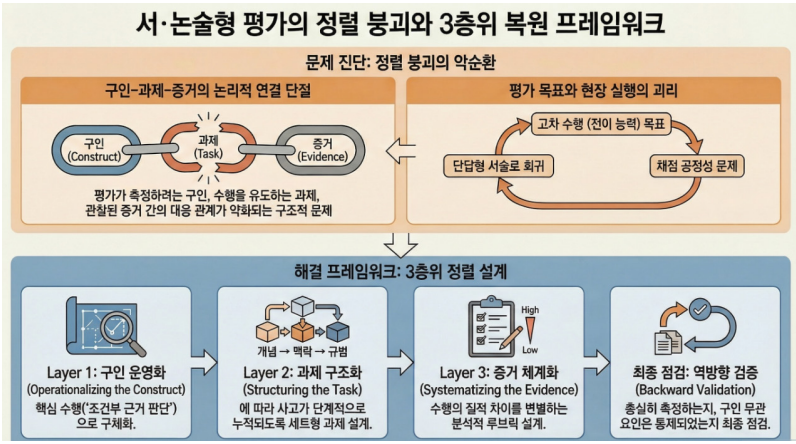
III. 세트형 서·논술형 평가의 설계 프레임워크와 작동 메커니즘

본 장은 Ⅱ장에서 논증한 정렬 붕괴(alignment collapse)의 구조를 설계 수준에서 통제하기 위한 대안으로서, 세트형 서·논술형 평가의 설계 프레임워크와 작동 메커니즘을 제시하는 데 목적이 있다. 여기서 '세트형'은

문항을 여러 개 제시하는 형식 자체를 의미하지 않는다. 세트형은 구인–증거–과제–루브릭의 대응 관계가 느슨해지는 문제를 최소한의 설계 산출물로 고정하고, 정렬 상태를 점검·조정할 수 있도록 하는 구조적 장치를 의미한다(Messick, 1989; Kane, 2013; American Educational Research Association et al., 2014). 본 장은 프레임워크의 논리와 절차를 일반 모델로 제시하며, 특정 소재나 실제 과제 사례의 내용은 다음 장에서 구체화한다.

본 프레임워크는 Ⅱ장에서 검토한 ECD·CBIL·학문 문식성의 핵심 논리를 병렬적으로 병치하는 데 머물지 않고, 각 이론이 제공하는 기능과 각 이론만으로 남는 공백을 3층위 설계 체계의 서로 다른 Layer에 재배치함으로써 서·논술형 평가에 특화된 최소 설계 산출물 체계로 재구성한다. ECD는 구인–증거–과제–판정의 정렬 골격을 제공하고, CBIL은 사실 → 개념 → 일반화/전이의 인지 심화 경로를 제공하며, 학문 문식성은 영역 고유의 추론 절차와 담화 표지를 통해 운영화 구인과 판정 표지를 정교화한다. 이에 본 연구는 이들 이론 자원을 Layer 1의 구인 운영화, Layer 2의 과제 구조화, Layer 3의 증거 판정으로 번역하고, 역방향 검증을 통해 정렬 상태를 점검하는 3층위 설계 프레임워크를 제안한다. 이때 학문 문식성은 독립된 설계 절차의 완결 체계라기보다, 운영화 구인과 판정 가능한 담화적 표지를 정교화하는 증거화 자원으로 기능한다.

〈그림 1〉은 본 연구의 설계 체계를 3층위 정렬(Layer 1–3)과 역방향 검증(Backward Validation)의 순환 구조로 제시한다. Layer 1은 구인의 운영화, Layer 2는 과제의 구조화, Layer 3은 증거의 판정(루브릭 판정 규칙)을 담당한다. 역방향 검증은 최종 산출물(루브릭·예시 반응)이 원래의 구인 정의를 충실히 반영하는지와 구인 무관 분산(construct-irrelevant variance; CIV)의 영향을 최소화했는지를 점검하는 절차이다(American Educational Research Association et al., 2014; Kane, 2013).



〈그림 1〉 서·논술형 평가의 정렬 붕괴와 복원 프레임워크

〈표 2〉는 층위별 핵심 산출물과 설계 질문을 요약한 것으로, 이후 절의 서술은 〈표 2〉를 기준 좌표로 삼아 전개된다.

〈표 2〉 층위별 핵심 산출물과 설계 질문

층위	핵심 산출물	설계자가 답해야 할 질문	최소 점검 기준(요지)
Layer 1(구인)	구인 정의, 하위 차원, 관찰 가능한 수행 표지	무엇을 전이 수행으로 볼 것인가? 그 수행은 어떤 담화적 증거로 드러나는가?	운영화 구인이 채점 가능한 수행 요소로 분해되었는가
Layer 2(과제)	지문 기능 배치, 문항 단계(책임 이전), 마찰 설계	해당 수행을 유도하는 자료(지문)와 문항 연계를 어떤 순서로 설계할 것인가?	토대 – 마찰 – 규범 기능 요건이 누락 없이 구현되는가
Layer 3(증거)	루브릭 준거, 수준 경계, 판정 규칙	무엇을 증거로 간주하고, 수준 차이를 어떤 기준으로 변별할 것인가?	조건 – 추론 – 정당화가 변별의 중심에 고정되는가
정렬 점검	역방향 검증 절차	루브릭이 구인을 충실히 반영하는가? 구인 무관 요인은 통제되는가?	구인 충실성 + CIV 통제의 이중 점검이 가능한가

본 장은 설계 결과물(지문·문항·루브릭) 자체보다, 그 결과물이 일관되

게 산출되도록 하는 설계 원리와 작동 메커니즘, 그리고 최소 점검 기준을 제시한다. 또한 Ⅱ장 3절에서 개념 수준으로 정리한 설계 실행 원리(책임 이전, 생산적 마찰, 분석적 채점)는 핵심 이론 틀을 대체하는 독립 이론이 아니라, Layer별 설계 결정과 점검 기준을 구체화하기 위한 운영화 메커니즘(operationalization mechanisms)으로 사용한다. 본 장에서는 설계 원리와 최소 점검 기준을 중심으로 논의를 전개한다.

1. Layer 1: 구인의 운영화

Layer 1의 목적은 교육과정이 요구하는 상위 목표(전이)를 평가 장면에서 채점 가능한 수행 구인으로 운영화하고, 이를 관찰 가능한 담화적 증거로 구체화하는 데 있다. 여기서 본 연구가 다루는 전이는 전이 일반 전체를 포괄하는 범용 모형이 아니라, 자료 비평과 논증적 판단 맥락에서 학습자가 조건 정보를 식별하고, 그 조건이 결론의 타당성 및 일반화 범위를 어떻게 제약하는지 추론하며, 그 판단을 공적 준거로 정당화하는 수행에 초점을 둔 판단 전이의 설계 모형이다. 구인의 운영화는 측정하려는 추상 개념을 관찰·판정 가능한 하위 요소로 분해하는 절차로서(Messick, 1989), ECD는 이를 구인 모형－증거 모형－과제 모형의 연역적 관계로 체계화한다(Mislevy et al., 2003). 전이는 그 자체로 추상 구인이므로, Layer 2·3의 설계가 일관되게 작동하기 위해서는 우선 Layer 1에서 구인 구조가 명시되어야 한다.

1) 구인의 이중 구조: 전이 → 조건부 근거 판단 → 담화적 표지

본 연구는 전이를 평가 응답에서 확인 가능한 수행으로 전환하기 위해 ‘조건부 근거 판단’을 운영화 구인으로 설정한다. 이는 전이 일반의 모든 하위 유형을 대체하는 정의가 아니라, 자료 비평과 논증적 판단이 결합된 과제 맥락에서 전이가 실제 답안에 어떻게 드러나는지를 포착하기 위한 제한적 운영화이다. 곧 학문 문식성 논의가 제시한 영역 고유의 추론 절차, 즉 자

료의 조건 정보를 식별하고 그 조건이 결론에 미치는 영향을 추론하며 정당화하는 담화 구성을 평가 구인으로 전환한 것이다(Shanahan & Shanahan, 2008; 편지윤, 2021). 따라서 본 연구는 공적 준거를 활용한 판단 전이의 설계 가능성을 중심으로 논의를 전개한다.

〈표 3〉은 상위 구인(전이)과 운영화 구인(조건부 근거 판단), 그리고 이를 판정하기 위한 담화적 표지를 위계적으로 정리한 것이다. 〈표 3〉은 Layer 2에서 "어떤 지문 기능과 어떤 문항 요구가 필요한가"를, Layer 3에서 "무엇이 점수 차이를 만들어야 하는가"를 규정하는 기준틀로 작동한다. 즉, Layer 1은 이후 층위에서 발생할 수 있는 설계 임의성(예: 구인과 무관한 지문 선택, 표현 중심의 채점 기준)을 사전에 제약하는 선결정 장치로 기능한다.

〈표 3〉 구인의 이중 구조: 전이 → 조건부 근거 판단 → 담화적 표지

위계	정의(핵심 의미)	평가에서의 역할	대표 관찰 단위(예)
상위 구인: 전이	맥락 변화 속에서 개념·원리를 재적용·조정하여 판단·표현하는 능력	평가가 포착해야 하는 목표 수행	세트 전체 수행 (종합 응답)
운영화 구인: 조건부 근거 판단	조건 정보를 식별하고, 조건이 결론의 타당성·일반화 범위를 어떻게 제약하는지 추론하며, 공적 기준으로 정당화하는 수행	채점 가능한 핵심 수행 구인	문항 간 누적 응답 (특히 종합 응답)
담화적 표지(증거)	조건 명시·추론, 조건–결론 연결의 논증 구조, 기준·규범 동원 정당화의 언어적 흔적	루브릭이 확인할 증거 항목	문장/단락 수준 근거 진술

2) 조건부 근거 판단의 3단계 구조와 설계 함의

운영화 구인인 조건부 근거 판단은 단일 능력이 아니라 세 단계의 인지·담화 과정이 연쇄적으로 작동하는 복합 수행이다. 〈표 4〉는 이 3단계 구조를 '관찰 가능한 증거'와 '과제 설계 요구(최소 조건)'로 연결하여 제시한다. Layer 1의 핵심은 〈표 4〉의 '과제 설계 요구'를 Layer 2에서 설계 원리(지문 기능 분업·문항 단계 누적·마찰 강도 조정)로 전환하는 데 있다.

〈표 4〉 조건부 근거 판단의 3단계 구조와 설계 함의

단계	인지·담화 과정(요지)	답안에서 관찰 가능한 담화적 증거(예)	과제 설계 요구 (최소 조건)
1단계: 조건 식별	자료 생산 조건·제약을 포착하고 의미를 부여	출처·표본·시점·문항 방식·누락 정보 등을 '조건'으로 명명	조건 정보가 드러나거나, 조건 누락의 중요성을 추론할 단서 제공
2단계: 제약 추론	조건이 결론의 타당성/일반화 가능성을 제한함을 논리적으로 연결	"A 조건이므로 B 결론은 과대/과소 추정될 수 있음"과 같은 제약 진술	조건 변화/불완전성이 결론을 흔들도록 설계(비교·대조·반례 가능)
3단계: 규범 정당화	공적 기준을 근거로 판단을 조직	준거 제시 → 판단 → 근거 조직 → 반론 처리/대안 제시	공적 준거가 자료 또는 과제 조건에 포함(기준 제시 또는 준거 선택 요구)

〔**핵심 결정 사항 | Layer 1**〕
- 전이를 어떤 운영화 구인으로 고정할 것인가(본 연구: 조건부 근거 판단)
- 운영화 구인을 판정 가능한 담화적 표지로 분해했는가(조건·추론·정당화)
- 상 수준 수행이 '정보 포함'이 아니라 '조건–결론 제약 추론'과 '규범 기반 정당화'를 필수로 요구하도록 정의되어 있는가

2. Layer 2: 과제의 구조화(지문 기능 분업과 문항 간 인지 연계)

Layer 2는 〈표 3〉과 〈표 4〉에서 규정한 수행이 실제 응답으로 산출되도록 지문(정보 자원)과 문항(수행 요구)을 구조화하는 단계이다. 본 연구에서 Layer 2의 핵심은 (1) 지문 기능의 분업을 통해 토대–마찰–규범을 누락 없이 제공하고, (2) 문항 단계에서 책임을 점진적으로 이전하여 수행을 누적시키며, (3) 생산적 마찰을 조정하여 단답화·형식화를 방지하는 데 있다. 이하에서는 이 결정을 지배하는 설계 원리와 작동 메커니즘, 그리고 최소 점검 기준을 제시한다.

1) 지문 기능 분업 원리: '토대–마찰–규범'의 설계적 분리

지문은 단순 정보 제시가 아니라 사고 단계의 자원을 제공해야 한다. 이는 CBIL의 사실–개념–일반화 경로를 평가 과제 구조에 적용한 것으로(Er-

ickson et al., 2017/2019), 토대는 조건 식별과 기본 렌즈를 마련하고, 마찰은 조건–결론 제약 추론이 회피되지 않도록 결론을 흔들며, 규범은 판단을 공적 준거에 의해 정당화할 자원을 제공한다. 이 기능 분업이 확보될 때 〈표 4〉의 1~3단계(조건 식별–제약 추론–규범 정당화)를 과제 구조에 구현할 수 있다.

2) 문항 단계 누적 원리: 책임의 점진적 이전(구조화 → 자율)

세트형 과제의 문항은 동일 난이도의 병렬 배열이 아니라, 초기에는 수행을 유도하는 구조화가 제공되고 후속에는 통합 수행을 학습자가 조직하도록 요구가 조정되어야 한다. 이는 비계 이론의 점진적 책임 이전(gradual release of responsibility) 원리를 평가 과제 배열에 적용한 것으로(Fisher & Frey, 2013; Wood et al., 1976), Ⅱ장 3절에서 제시한 설계 실행 원리를 Layer 2에서 문항 요구의 형식·범위·통합 수준을 조정하는 구체적 메커니즘으로 전환한 것이다.

3) 생산적 마찰 조정 원리: 과잉·결손의 동시 통제

단답화·형식화의 중요한 원인은 '추론이 필요 없는 과제' 혹은 '추론을 포기하게 만드는 과제'의 양극단이다. 따라서 Layer 2의 마찰은 "혼란의 양"이 아니라, 조건–결론 제약 추론을 필수적 경로로 만드는 강도로 조정되어야 한다. 이는 생산적 실패(productive failure) 연구에서 제시된 적정 난이도 설계 원리, 즉 학습자가 기존 지식으로는 해결하기 어려우나 새로운 개념 적용을 통해 해결 가능한 과제 조건을 설정하는 논리를 참조한 것이다(Kapur, 2008; Kapur & Bielaczyc, 2012). 마찰은 토대(렌즈)와 규범(준거) 없이 단독으로 제시될 수 없으며, 토대 제시가 절차 안내의 과잉으로 기울 경우 마찰이 약화되는 문제가 발생할 수 있다. 따라서 토대–마찰–규범의 균형은 설계 단계에서 조정되어야 할 핵심 결정 사항이다. 위 세 원리를 설계자가 적용할 때 필요한 결정을 〈표 5〉로 요약한다.

〈표 5〉 Layer 2 핵심 의사결정 요약표

결정 축	설계 질문	최소 점검 기준(요지)	대표 조정 방향(요지)
토대 (지문 기능)	조건을 '무엇'으로 볼지 정할 렌즈가 제공되는가	조건 범주·용어·기본 관계가 확보됨	토대 지문에 핵심 개념·조건–결론 관계의 기본틀 보강
마찰 (지문/문항)	조건 분석과 추론을 회피할 수 없게 만드는가	조건 변이/불완전성이 결론을 흔들어 추론을 강제	비교 기준을 명확화하고 단서는 제공하되, 정답 경로는 열어두지 않음
규범 (지문/문항)	공적 기준으로 정당화할 근거가 제공되는가	준거가 제시되거나 준거 선택·적용이 요구됨	규범 지문에 준거 제공 또는 문항에서 준거 선택·적용 요구 강화
책임 이전 (문항 단계)	후속 문항에서 자율 통합이 발생하는가	종합 문항에서 조건·추론·정당화가 통합됨	안내 축소 + 통합 요구 강화(반론/대안 포함 가능)

〔**핵심 결정 사항 | Layer 2**〕
- 토대–마찰–규범 중 어느 기능이 누락되었는가(누락은 추론 부재로 귀결될 가능성이 큼)
- 문항 진행에서 책임이 실제로 이전되는가(재진술 반복은 이전 실패 신호가 될 수 있음)
- 마찰이 추론 필수화로 작동하는가, 아니면 혼란 과잉/정답화로 이탈하는가
- 종합 문항이 3단계(조건 식별–제약 추론–규범 정당화)를 통합하도록 요구하는가

3. Layer 3: 증거의 판정(루브릭 준거·수준 경계·판정 규칙)

Layer 3의 목적은 Layer 2에서 산출되는 응답을 〈표 3〉~〈표 4〉의 구인 구조에 비추어 일관되게 판정하는 데 있다. 서·논술형 응답은 다양성을 전제하므로, 판정 규칙은 표현의 유려함이나 근거의 양 같은 구인 무관 요소(CIV)에 의해 점수가 좌우되지 않도록 설계되어야 한다. 따라서 변별의 근거는 조건 식별–제약 추론–규범 정당화의 수행 요소에 고정되어야 한다.

1) 관점과 기준의 이원화 원리

판정의 첫 조건은 '무엇이 점수 차이를 만들어야 하는가'를 고정하는 일이다. 본 연구에서는 조건부 근거 판단의 구성 요소가 곧 변별 축이 되며, 판단의 준거는 채점자의 주관적 인상이 아니라 공적 기준의 적용 여부로 제한

된다. 이는 분석적 루브릭의 핵심 원리, 즉 복수 준거를 독립적으로 점검하여 구인과 판정 규칙의 대응을 밀착시키는 방식을 따른 것이다(Brookhart, 2013/2022; Panadero & Jonsson, 2013). 이를 통해 논리 없는 결론이나 형식 중심 서술이 상 수준으로 판정되는 오류를 억제한다.

2) 근거의 질적 차등 원리

수준 경계는 근거의 개수나 문장량이 아니라, 〈표 4〉의 3단계가 답안에서 질적으로 구현되는 방식(조건 식별의 정교함, 제약 추론의 명시성, 규범 정당화의 적용 수준)에 의해 설정되어야 한다. 고차 수행일수록 답안 다양성이 증가하므로, 판정의 일관성을 확보하기 위해서는 수행의 질적 수준을 명시적으로 기술한 루브릭이 필요하다(Jonsson & Svingby, 2007). CBIL의 사고 경로는 여기에서 수준 구획의 보조적 준거로만 사용되며, 수준 정의의 중심은 운영화 구인(조건부 근거 판단)이다. 아래 〈표 6〉은 Layer 3에서 유지해야 할 최소 준거 축과 수준 경계 설정 원리를 요약한 것이다.

〈표 6〉 루브릭 최소 준거 축과 수준 경계 설정 원리

준거 축	상 수준의 필수 요건(요지)	중/하 수준의 대표 결손(요지)	CIV 통제 포인트
조건 식별	조건 범주를 다각도로 식별하고 핵심 조건을 선별함	조건을 일부만 식별하거나 조건으로 인식하지 못함	조건 언급 '유무'가 아니라 조건의 기능화(변수화)를 본다
제약 추론	조건 – 결론 제약을 명시적 논리로 연결함	연결이 약하거나 암묵적/연결 부재	표현의 유려함보다 추론의 명시성을 우선 판정
규범 정당화	공적 준거 적용으로 판단·반론·대안을 조직함	준거 언급이 피상적/준거 없이 상식·감정 의존	배경지식 과잉이 점수를 좌우하지 않도록 준거 적용 근거를 지문에 고정

〔핵심 결정 사항 | Layer 3〕
- 상 수준 답안이 3단계(조건 식별+제약 추론+규범 정당화)를 모두 요구하는가
- 수준 차이가 근거의 양/문장량이 아니라 추론·정당화의 질로 변별되는가
- CIV(표현·형식·배경지식)가 점수 차이를 좌우하지 않도록 판정 규칙이 고정되어 있는가

4. 정렬 점검: 역방향 검증(Backward Validation)

본 프레임워크는 설계 산출물이 구인 논증을 지지하는지 점검하는 역방향 검증 절차를 포함한다. 역방향 검증은 평가 점수의 해석과 사용이 의도한 구인을 충실히 반영하는지, 그리고 구인 무관 분산(CIV)이 통제되었는지를 설계 단계에서 점검하는 타당도 논증의 핵심 절차이다(Kane, 2013; Messick, 1989). 본 연구에서는 효과 검증이나 일반화 주장과 구분하여, 설계 정합성 점검의 최소 절차로 제시된다. 본 절에서는 복수 표의 총망라가 아니라, 실행 단계에서 사용할 수 있는 간소화 체크리스트를 〈표 7〉로 제시한다.

〈표 7〉 역방향 검증 체크리스트

점검 차원	점검 질문	수정 필요 신호(예)	조정 원리(요지)
구인 충실성	상 수준이 3단계를 모두 요구하는가	상 수준인데 조건 추론 없이 결론만 제시	조건–결론 연결 추론의 명시를 상 수준 필수 요건으로 상향
CIV 통제	표현·형식·배경지식이 점수 차이를 좌우하는가	형식 완벽하나 추론 결여가 상 수준	형식 요소를 요건/감점으로 제한, 구인 점수는 추론에 고정
수준 변별	상–중–하 경계가 질적으로 명확한가	상·중 차이가 근거 개수로만 갈림	도달 단계(추론·정당화의 질)로 수준 경계 재정의
증거 산출성	과제가 해당 증거를 실제로 산출하게 하는가	규범 정당화 준거가 지문/문항에 없음	준거 제시 또는 준거 선택·적용 요구를 과제 조건에 삽입
층위 간 정렬	Layer 1 →2 →3 연결이 연역적으로 유지되는가	지문은 풍부하나 문항이 단답 요구	문항 요구를 조건–제약–정당화 구조로 재구성
역추적 가능성	수정 시 어느 층위를 조정해야 하는지 드러나는가	문제 발생 시 조정 위치가 불명확	결손이 '조건/추론/규범' 중 어디인지 표시한 뒤 해당 Layer를 조정

이상에서 제시한 3층위 설계 체계는 (1) 전이 수행을 운영화 구인으로 고정하고(Layer 1), (2) 그 구인이 응답으로 산출되도록 지문 기능과 문항 단

계를 구조화하며(Layer 2), (3) 산출된 증거를 구인 충실성과 CIV 통제의 관점에서 일관되게 판정하도록 루브릭 규칙을 고정하는 방식(Layer 3)으로 구성된다. 또한 역방향 검증은 이러한 정렬 상태가 실제로 유지되는지 점검·조정하기 위한 최소 절차로 기능한다. 다음 장에서는 본 장의 일반 원리가 2022 개정 '독서와 작문'의 자료 비평 맥락에서 구체적으로 어떻게 구현되는지를 사례로 제시한다.

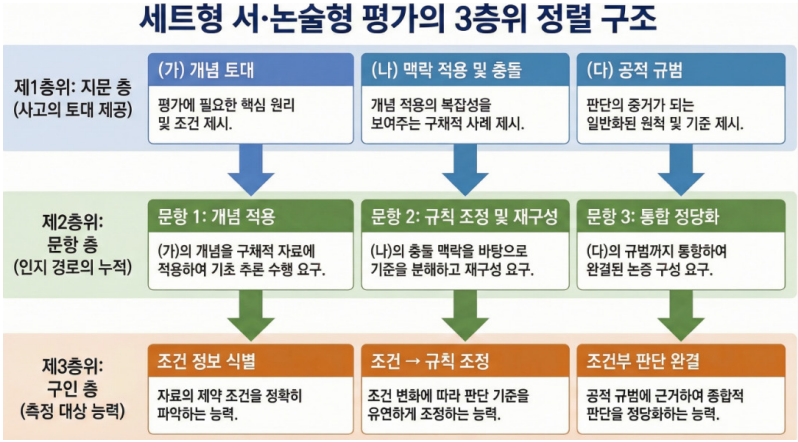
IV. 설계 프레임워크의 구체화: 여론조사 자료 비평 세트 설계 사례

본 장은 Ⅲ장에서 제안한 정렬 복원 프레임워크를 2022 개정 '독서와 작문' 맥락에서 구체화한 설계 사례를 제시한다. 여론조사 자료 비평이라는 영역은 조건 정보(표본, 응답률, 질문 표현)에 따라 해석이 달라지는 특성으로 인해, Ⅲ장 1절에서 정의한 '조건부 근거 판단' 구인을 관찰 가능한 수행으로 전환하기에 적합한 맥락이다. 특히 이 사례는 전이 일반 전체를 대표하려는 것이 아니라, 사회적 자료를 비판적으로 해석하고 공적 기준으로 판단을 정당화하는 판단 전이가 어떻게 설계 가능한 수행으로 번역될 수 있는지를 보여주는 구체적 예시라는 점에서 의의를 가진다. 2022 개정 교육과정의 '독서와 작문'은 다자료 기반 이해와 판단, 읽기–쓰기 통합 수행, 논증적 표현을 통한 지식 재구성을 핵심 성취기준으로 제시하며(교육부, 2022; 최숙기·박종임, 2023ㄱ), 이는 Ⅲ장 2절에서 제안한 지문 기능 분업과 문항 간 인지 연계 설계가 작동할 수 있는 교육과정적 토대를 제공한다.

본 설계는 Ⅱ장에서 진단한 정렬 붕괴 문제를 Ⅲ장의 3층위 프레임워크로 통제한 구체적 사례(design case)이다. 지문·문항·루브릭이라는 설계 산출물을 통해 구인–증거–과제의 정렬 상태가 어떻게 고정·점검되는지 제

시하며, 학습자 수행 기반 경험적 타당화는 후속 연구 과제로 남겨 둔다.

〈그림 2〉는 Ⅲ장의 3층위 정렬 체계가 본 과제 세트에서 구현되는 양상을 요약한 도식이다. Layer 1은 '조건부 근거 판단' 3단계를 문항 요구로 전환하고, Layer 2는 지문 A~C와 문항 1~3을 CBIL 경로와 책임 이전 원리에 따라 배열하며, Layer 3은 관점/기준 이원화와 질적 차등 원리를 적용한 루브릭으로 증거를 판정한다.



〈그림 2〉 세트형 서·논술형 평가의 3층위 정렬 구조

이어서 Ⅳ장 1절은 설계 맥락과 입력 조건을, Ⅳ장 2절은 Layer 1의 구현(구인 3단계 분산 배치)을, Ⅳ장 3절은 Layer 2의 구현(지문 기능 분업과 문항 인지 경로 연계)을, Ⅳ장 4절은 Layer 3의 구현(루브릭 변별 메커니즘)을, Ⅳ장 5절은 역방향 검증을 각각 제시한다. 지문·문항 전문은 〈부록 1〉, 루브릭 및 예상 답안 전문은 〈부록 2〉에 제시하였다.

1. 설계 맥락 및 설계 요구: 입력 조건의 명시

본 절은 본 세트 설계가 전제로 삼는 교육과정 맥락, 과제 영역 특성, 그리고 설계 제약 조건을 명시한다.

1) 교육과정 맥락: '독서와 작문' 성취기준과의 정렬

본 세트는 2022 개정 교육과정 '독서와 작문'의 성취기준과 직접 정렬되도록 설계되었다. 구체적으로 정보 추론(01-03), 비판적 평가(01-04), 표현 전략 활용(01-05), 논증적 글쓰기(01-11), 주제 통합적 독서(01-13) 성취기준이 문항1-2-3에 각각 배치되며, 이는 Ⅲ장 1절에서 정의한 조건부 근거 판단의 3단계(조건 식별 → 조건-결론 제약 추론 → 정당화 구조 조직)와 대응된다. 문항별 성취기준의 구체적 정렬 구조는 〈표 8〉과 같다.

〈표 8〉 문항–성취기준–구인–과제 조건의 정렬

문항	배점	대응 성취기준	측정 구인 단계	문항 요구 조건 (구인 → 과제 구성)
문항 1	4점	[12독작01-03] 글에 드러난 정보를 바탕으로 내용을 파악하고 드러나지 않은 정보를 추론하며 읽는다.	1단계: 조건 정보 식별 접속률·응답률·반영률 개념을 활용해 〈보기〉 수치(접속률 20%, 응답률 15%, 반영률 3.0%)를 해석하고, 응답자 편향 가능 및 전체 대표성 제약을 서로 다른 2가지 관점에서 식별	"(가)의 '접속률'과 '응답률(및 반영률)' 개념을 사용해 〈보기〉 수치를 해석할 것" "서로 다른 2가지 관점(기준)에서 설명" → 조건 차원 분해 유도
문항 2	6점	[12독작01-04] 글의 내용·관점·표현 방법을 평가하며 읽는다. [12독작01-05] 글의 내용 조직 방법과 표현 전략을 찾고 글쓰기에 활용한다.	2단계: 조건-결론 제약 추론 질문 문장의 유도성과 선택지 구성의 공정성을 서로 다른 기준으로 분리하여 중립성을 판단하고, 편향 문항을 중립적으로 재구성	"중립 질문 선택의 근거를 (가) 또는 (나)에서 찾아 서로 다른 2가지 기준으로 제시" "편향된 질문을 중립 문장으로 고치고, 선택지 3개 이상으로 균형 제시" → 조건 변화의 효과 인식

문항 3	10점	〔12독작01-11〕 타당한 근거를 수집하고 효과적인 설득 전략을 활용하여 논증하는 글을 쓴다. 〔12독작01-13〕 다양한 글을 주제 통합적으로 읽고 학습 목적을 고려하여 글을 쓴다.	3단계: 정당화 구조 조직 개인 경험을 문제의식으로 형성하고, (가)–(나)–(다)를 통합하여 근거 3가지 이상 제시, 그 중 (다)의 표본오차–격차 비교 논리를 적용하여 조건부 근거 기반 논증 완성	"개인 경험/상황을 '조건 정보 없이 단정적 공유 → 혼란/오해'로 재맥락화하여 문제의식 제시" "(다)의 '표본오차–격차 비교' 논리 적용 필수" "반론·재반박 포함" → 규범 기반 정당화 구조 완결

2) 과제 영역 특성: 여론조사 자료의 조건–결론 의존성

여론조사 자료 비평은 다음과 같은 영역 특성으로 인해 조건부 근거 판단 구인을 수행으로 전환하기에 적합하다고 판단하였다. 첫째, 조건 정보의 명시적 식별 가능성이다. 표본 크기, 응답률, 질문 표현, 선택지 구성 등이 수치 또는 텍스트로 명시되므로, 학습자는 '조건 정보 식별' 단계를 구체적으로 수행할 수 있다. 둘째, 조건–결론 제약의 인과적 추론 요구이다. 동일한 주제라도 응답률이나 질문 표현에 따라 결론의 타당성이 달라지므로, 조건이 결론을 어떻게 제약하는지에 대한 추론이 필수적이다. 셋째, 공적 규범의 준거 제공 가능성이다. 중앙선거관리위원회의 여론조사 공표 기준, 통계 윤리, 조사 방법론 등 영역 고유의 판단 기준이 존재하므로, 학습자가 개인 의견이 아니라 공적 기준에 근거해 정당화를 수행하도록 유도할 수 있다.

아울러 본 과제는 통계적 사실의 '정확한 계산'을 평가하기 위한 것이 아니라, 수치와 용어가 텍스트 내에서 증거로 기능하는 방식과 그 한계를 비판적으로 읽어내는 수행을 평가하기 위한 설계이다. 따라서 요구되는 수리 처리는 비율·백분율 해석 등 최소 수준에 한정되며, 점수 차이는 계산 숙련이 아니라 조건 식별 – 제약 추론 – 정당화의 담화적 수행에서 발생하도록 구성하였다.

3) 주요 설계 제약 조건

본 설계는 2022 개정 교육과정 '독서와 작문'의 성취기준(정보추론, 비판

적 평가, 논증적 글쓰기)과 직접 정렬된다. 또한 본 사례는 고 2~3학년 수준의 기본적인 통계 문해력(비율, 백분율 해석)과 공적 담화 판단 경험을 전제한 설계이며, 이러한 점에서 본 연구의 1차 적용 범위 역시 중등 이상 학습자로 한정된다.[1] 설계의 주요 입력 조건과 제약 사항은 〈표 9〉와 같다.

〈표 9〉 설계 입력 조건 및 제약

조건 범주	설정값	설계 함의
학습자 수준	고 2~3학년	기본적인 통계 문해력(비율, 백분율) 전제
과제 분량	50분, 3지문-3문항	지문 총량 약 2,500자, 서술형 2문항+논술형 1문항 구조
평가 목적	다목적 평가 설계	교수·학습 맥락에 따라 총괄평가(of learning), 형성평가(for learning), 자기평가(as learning)로 유연하게 활용 가능하도록 구조화
채점·피드백	분석적 루브릭	CBIL 경로 기반의 질적 수준 진단이 가능한 채점 기준 적용

본 설계는 특정 평가 목적(형성/총괄/자기평가)을 사전에 고정하지 않으며, 교사의 교수·학습 맥락에 따라 재맥락화될 수 있도록 구조화되었다. 평가 목적에 따라 루브릭 활용 방식(총괄평가 시 앵커 예시, 형성평가 시 피드백 대화, 자기평가 시 해석 훈련 등)은 조정이 필요하다.

2. Layer 1의 적용: 구인(조건부 근거 판단)의 과제화

Layer 1은 추상적 구인을 과제 수행 요구로 변환하는 단계이다. Ⅲ장 1절에서 정의한 '조건부 근거 판단'의 3단계 위계는 문항 1-2-3에 분산 배치되어, 학습자가 개별 문항을 해결하는 과정에서 자연스럽게 구인 전체를 수

1) 초등 또는 초기 중등 단계에 적용할 경우에는 '규범 정당화'의 수준과 준거 제시 방식에 대한 추가 조정이 필요하다.

행하도록 유도 의도를 갖고 배치되었다. 〈표 10〉은 구인의 하위 요소가 각 문항의 구체적 요구 조건으로 매핑되는 논리를 보여준다.

〈표 10〉 구인 단계 – 문항 요구 조건 매핑

구인 단계	문항 (유형)	과제화 논리 (Mapping Logic)	구체적 수행 요구 (〈부록 1〉 참조)
1단계 조건 정보 식별	문항1 (서술형)	〔자료 내 단서 포착〕 텍스트에 명시된 통계적 조건(수치, 비율)을 식별하고, 이를 개념적 틀(지문 A)로 해석하도록 요구	- 제시된 표의 '접속률/응답률' 수치 식별 - '반영률 3.0%'의 의미를 지문 A의 개념으로 해석
2단계 조건–결론 제약 추론	문항2 (서술형)	〔조건 변화에 따른 판단〕 동일 사안에 대해 조건(질문 표현, 선택지)이 달라질 때 결론이 어떻게 왜곡되는지 비교·추론하도록 요구	- 편향된 질문(문항 1번)과 중립적 질문(문항 2번)의 비교 - 질문 표현 변경 시 응답 결과의 차이 추론
3단계 정당화 구조 조직	문항3 (논술형)	〔통합적 논증 구성〕 식별한 조건과 추론 결과를 종합하여, 공적 규범(지문 C)을 근거로 타당성을 판정하고 반론을 방어하도록 요구	- 뉴스 보도의 타당성 판정(타당/미흡) - 지문 C의 '공표 기준'을 근거로 반론 제기 및 재반박 - 구체적 개선 방안 제안

이는 문항 1의 '식별' 수행이 문항 2의 '추론'을 거쳐 문항 3의 '정당화'로 누적되도록 설계되었음을 나타낸다.

3. Layer 2의 구현: 지문–문항의 인지 경로 누적 설계

Layer 2는 학습자의 사고가 심화되도록 정보 자원(지문)과 수행 절차(문항)를 배치하는 단계이다. 이는 Ⅲ장 2절에서 제시한CBIL(사실–개념–일반화) 경로와 책임 이전(responsibility transfer) 원리를 따른다.

1) 지문 기능의 전략적 분업(지문 A → B → C)

지문 (가)~(다)는 단순한 정보 나열이 아니라, 〈표 11〉과 같이 사고의 단계를 유도하는 기능적 위계로 설계되었다.

〈표 11〉 지문 기능 분업과 인지적 유도 전략

지문	기능 (Function)	설계 의도 및 인지적 유도 전략
지문 A (가)	개념 토대 제공 (ConceptualLens)	- 여론조사의 기본 원리(표본, 오차)와 한계 제공 - 전략: 이후 제시될 사례(B, C)를 분석할 수 있는 '분석적 틀(lens)'을 학습자에게 사전 제공함
지문 B (나)	맥락 적용·충돌 (ContextualApplication)	- 교실 상황에서의 설문 오류 사례 제시(불완전 정보, 편향) - 전략: 지문 A의 개념이 실제 맥락에서 어떻게 왜곡되는지 보여주며, 학습자가 문제의식을 구체화하도록 유도
지문 C (다)	공적 규범 제시 (PublicNorms)	- 선거여론조사 공표 기준 및 통계 보도 윤리 제시 - 전략: 개인적 비판을 넘어, 사회적으로 합의된 기준(Norm)에 근거해 판단을 정당화하도록 '전이의 준거' 제공

2) 문항 간 책임 이전과 생산적 마찰

문항 1 → 3으로 진행될수록 비계는 축소되고 학습자의 통합 책임은 확대된다. 문항 1~2는 "(가)의 개념을 사용하여", "서로 다른 2가지 관점으로"와 같이 사고의 틀을 명시적으로 제공하여 수행을 유도한다. 문항 3은 "논증적 글쓰기 방식으로 작성하시오"라는 포괄적 지시를 제시하되, 지문 (가)~(다)와 문항 1~2에서 구축한 논리를 학습자가 자율적으로 선택·통합하도록 요구한다. 이는 Ⅲ장 2절 2항에서 제시한 '책임의 점진적 이전' 원리가 문항 배열에 구현된 지점이다.

4. Layer 3의 구현: 루브릭 변별 메커니즘

Layer 3은 학습자 반응을 일관되게 해석하는 증거 모형이다. 〈부록 2〉에 제시된 루브릭은 점수의 단순 합산이 아니라, 사고의 질적 도약 여부를 판정하기 위해 다음 두 가지 메커니즘을 적용한다.

1) 관점/기준의 이원화 원리

분석적 채점의 신뢰도를 높이기 위해, 하나의 문항 내에서 서로 다른 인

지적 차원을 분리하여 채점한다. 예컨대 문항 2는 단순히 "타당한가?"를 묻지 않고, ① 질문 문장의 유도성과 ② 선택지 구성의 공정성이라는 두 가지 독립된 기준을 각각 적용했는지 평가한다. 이는 학습자가 복합적인 문제를 단일 인상으로 처리하지 않고, 다각도로 분석했는지를 검증하는 장치다.

2) CBIL 경로 기반의 질적 차등(상-중-하 변별)

루브릭의 등급(상/중/하)은 정답 개수가 아니라 사고의 도달 깊이에 따라 결정된다. 〈표 12〉는 루브릭이 '조건부 근거 판단'의 수준을 어떻게 구획하는지 보여준다.

〈표 12〉 루브릭의 수준 변별 논리(CBIL 경로 적용)

수준	CBIL 단계	판정 기준 (Criteria Logic)	대표적 수행 양상 (〈부록 2〉 예시답안 참조)
상	일반화·전이 (Generalization)	- 규범 기반 통합: 지문 C의 공적 기준(오차범위 등)을 근거로 판단하고, 대안을 제시함 - 조건부 추론: "조건 X가 결여되었으므로 결론Y는 제한적이다"라는 인과 구조가 명시됨	"표본오차가 제시되지 않았으므로, 2%p 격차를 우세로 단정하는 것은 통계적으로 타당하지 않다."
중	개념 적용 (Conceptual Application)	- 개념적 연결: 지문 A의 개념(응답률 등)을 언급하고 문제점을 지적하나, 추론이 불완전함 - 단순 매칭: 조건과 결론의 관계보다 조건 자체의 문제 지적에 그침	"응답률이 낮아서 문제다. 사람들이 참여를 안 해서 믿을 수 없다."
하	사실적 지식 (Factual Knowledge)	- 인상 비평: 지문의 논리보다 개인적 경험이나 상식("그냥 이상하다")에 의존 - 정보 나열: 지문 내용을 베껴 쓰거나 단편적 사실만 언급함	"설문조사는 원래 잘 틀린다. 조심해야 한다."

5. 정렬 점검 및 역방향 검증

설계의 마지막 단계로, Ⅲ장 4절 〈표 7〉의 체크리스트를 활용하여 역방

향 검증(Backward Validation)을 수행하였다. 이는 설계 단계에서 루브릭이 구인을 충실히 반영하는지(구인 충실성)와 구인 무관 요인이 통제되었는지(CIV)를 점검하는 절차이다.

이러한 점검 절차는 다음 4단계로 수행되었다. ① 구인 충실성 점검(문항 요구와 구인 정의의 대조), ② CIV 통제 점검(가상 답안 시뮬레이션을 통한 구인 무관 요인 탐지), ③ 수준 간 변별력 점검(CBIL 경로 관점에서 상/중/하 경계의 질적 기준 재검토), ④ 조정 및 재점검(루브릭 수정 후 구인과의 재정렬 확인). 실제 채점자 간 신뢰도 및 학습자 수행 패턴 검증은 후속 연구 과제로 남겨 둔다. 〈표 13〉은 이 절차에서 탐지된 조정 필요 지점과 설계 조정의 대표 사례를 제시한다.

〈표 13〉 역방향 검증을 통한 설계 조정의 대표 사례

점검 영역	탐지된 문제	조정된 설계	정렬 효과
구인 충실성 (문항 1)	문항은 "서로 다른 2가지 관점" 요구하나, 루브릭에 "같은 말 반복" 판정 규칙 부재 → 응답자 편향/무응답자 누락을 동일 논리로 서술해도 만점 가능	루브릭에 "2가지처럼 보이나 실질적으로 같은 말 반복 시 1점" 규칙 추가	Ⅲ장 1절 구인 1단계 "조건 차원의 분해"가 판정 규칙으로 고정됨
CIV 통제 (문항 3)	초기 설계에서 "표현의 유려함"이 배점에 과도하게 영향 → 형식 완결성이 높으나 조건부 추론이 결여된 답안이 상 수준 획득 가능	표현 요소를 '형식 요건(감점제, 최대 -2점)'으로 분리하고, 배점 핵심을 '조건부 추론의 명시성'에 집중	점수 차이가 "글쓰기 능력"이 아니라 "조건부 근거 판단 능력"에서 발생하도록 통제
수준 간 변별력 (문항 3)	상/중 차이가 "근거 개수" (3개 vs 2개)로만 변별 → 근거는 많으나 지문 C 규범 미적용 답안이 상 수준 획득 가능	상 수준 필수 조건에 "Text C의 표본오차 – 격차 비교 논리를 적용하여 우열 단정 가능/불가능 판단" 명시	Ⅲ장 3절 2항 "CBIL 일반화 단계 도달"이 상 수준의 질적 기준으로 작동

이 외에도 수치 일관성 점검(응답률 15%와 반영률 3.0%의 개념 구분), 정수 운영 원칙(소수점 채점 배제), 하 수준 판정 규칙 명확화(지문 근거 없이 일반론만 제시 시 0점) 등의 조정이 이루어졌으며, 상세 내용은 〈부록 2〉의 루

브릭에 반영되었다. 실제 채점자 간 신뢰도 및 학습자 수행 패턴 분석은 후속 연구에서 수행될 예정이다.

V. 논의 및 제언

본 연구는 서·논술형 평가가 고차 사고력을 표방하면서도 실제로는 단답형으로 회귀하는 현상을 '정렬 붕괴(alignment collapse)'로 개념화하고, 이를 설계 수준에서 통제하기 위한 하나의 시도로 3층위 설계 체계(Layer 1 구인의 운영화 – Layer 2 과제의 구조화 – Layer 3 증거의 판정)와 역방향 검증(Backward Validation)을 제안하였다. 선행 연구들은 교사들이 채점 공정성 우려와 업무 부담 속에서 평가 요구를 단순화하는 행동을, 개인의 무성의가 아니라 제도적 조건 하에서 불확실성을 낮추기 위한 합리적 선택(최적화 행동)으로 해석해 왔다(남가영·김호정, 2023; 박종임, 2024). 이러한 논의를 고려할 때, 본 연구의 3층위 설계는 타당도 논증의 핵심 연결(구인 → 증거 → 과제 → 판정)을 개인의 암묵적 판단이 아닌 명시적 설계 산출물로 드러내어, 교사가 단답형 회귀를 통해 불확실성을 줄이려는 경향을 완화하는 하나의 설계적 대안으로 이해될 수 있다(Messick, 1989; Kane, 2013; 김수진 외, 2025). 다만 이러한 효과가 실제 실행 맥락에서 어느 정도 나타나는지는, 다양한 학교 조건에서의 적용 연구를 통해 경험적으로 검토될 필요가 있다.

본 연구의 학술적 의의는 "전이가 중요하다"는 수준을 넘어서, 전이를 무엇으로 관찰하고 어떻게 판정할 것인가를 설계 언어로 구체화하려 했다는 점에 있다(Mislevy et al., 2003; Messick, 1989). 이를 위해 전이를 '조건부 근거 판단'으로 운영화하고, 조건 식별 – 제약 추론 – 규범 정당화의 3단계 수행이 응답에서 담화적 표지로 나타나도록 구인 구조를 제시하였다(Shanahan & Shanahan, 2008; 편지윤, 2021). 또한 단일 문항 중심 설계가 복합

수행의 전개를 안정적으로 포착하기 어렵다는 논의(Bransford & Schwartz, 1999; Scardamalia & Bereiter, 1987)를 바탕으로, 사고 경로가 누적되도록 하는 세트형 구조를 설계·해석 단위로 활용하였다. 이때 본 설계는 '사회과 지식'이나 '통계 계산 능력' 자체를 측정하기보다, 텍스트에 제시된 수치·용어·조건이 주장의 타당성을 어떻게 제약하는지를 비판적으로 읽어내고, 그 제약을 언어로 명시하며, 공적 준거에 근거해 판단을 조직하는 '문식성 수행(literacy practice)'을 평가 대상으로 삼도록 구성되었다는 점에서 국어과 평가 구인으로 위치를 갖는다(Shanahan & Shanahan, 2008; 김영란, 2021; 편지윤, 2021). 설계 사례에서도 '정확한 계산'보다는 최소한의 수리 처리를 전제로 하고, 점수 차이가 계산 숙련이 아닌 조건 식별–제약 추론–정당화의 담화 구성에서 발생하도록 조정했음을 밝힌 바 있다.

현장 연구에 따르면, 교사가 매번 구인–과제–루브릭 정렬을 처음부터 구현하기는 쉽지 않고, 설계 절차 지식의 부족이 반복적으로 지적되어 왔다(박혜영 외, 2019; 김수진 외, 2025; 김형성, 2024). 따라서 정책적·실천적 차원에서는 '서·논술형 확대'라는 양적 목표를 넘어서, 본 연구와 유사한 설계 논리를 반영한 표준화된 설계 모델(템플릿)과 점검 체크리스트를 교육청·평가원 차원에서 개발·보급하는 방안을 검토할 수 있다.

그러나 표준 모델만으로 정교한 설계의 인지적 부담이 충분히 해소되기 어려울 수 있으므로, 후속 연구에서는 생성형 AI를 '지능 증강(IA)' 관점에서 결합하는 가능성을 탐색할 필요가 있다. 예를 들어 본 연구의 3층위 체계를 활용하여, 교사가 Layer 1의 구인과 Layer 3의 판정 규칙을 설정하고, AI가 Layer 2의 지문 조합과 문항 초안을 생성하며, 교사가 역방향 검증으로 적합성을 점검하는 '인간-AI 협력 설계(HITL)' 모델을 구상해 볼 수 있다(박고운, 2025). 평가 실행 단계에서도 AI를 보조적 피드백 도구로 활용하여 채점 부담과 피드백 시의성 문제를 완화하려는 시도들이 논의되고 있으며(최숙기·박종임, 2023ㄴ), 본 연구의 설계 문법은 이러한 논의를 구체화하는 하나의 기준틀로 활용될 수 있다. 다만 교사 지원 맥락에서의 AI 활용은 적용

조건, 책임 소재, 산출물의 품질 관리에 대한 논의와 함께 이루어져야 할 것이다.

본 연구는 정렬 붕괴를 통제하기 위한 설계 논리를 구성하는 이론적 설계 제안(design proposition) 연구로 위치 지어진다(Hevner et al., 2004). 이러한 연구 한계를 고려하여 제안된 체계의 타당성과 유용성은 경험적 검증을 통해 보완되어야 한다. 또한 타당도는 검사 자체가 아니라 점수 해석과 사용에 대한 논증이라는 관점에서, 향후 연구는 본 설계가 전제로 하는 해석·사용 논증을 어떤 경험적 증거로 지지할 것인지까지 포함해 계획될 필요가 있다(Kane, 2013).

후속 연구는 다음과 같은 방향을 포함할 수 있다. 첫째, 개발 루브릭을 실제 학생 답안에 적용해 채점자 간 신뢰도(inter-rater reliability)와 판정 일관성을 검증하고, 분석적 루브릭이 신뢰도 및 교육적 활용도에 미치는 영향을 점검하는 연구가 필요하다(Jonsson & Svingby, 2007). 둘째, 역방향 검증을 넘어 학습자의 사고 구술(think-aloud) 자료를 통해, 설계자가 의도한 인지 경로(조건 식별–제약 추론–규범 정당화)가 실제 수행에서 어떻게 나타나는지, 또는 어느 지점에서 어긋나는지를 확인하는 인지적 타당도(cognitive validity) 연구가 요구된다. 셋째, 본 연구는 '여론조사 비평'이라는 특정 맥락을 중심으로 설계되었으므로, 이 프레임워크를 문학 비평, 과학적 탐구, 역사적 사료 분석 등 다양한 영역으로 확장하여, 교과별 지식 구조에 따라 '토대–마찰–규범'의 지문 기능이 어떻게 변주되어야 하는지 탐색하는 연구가 필요하다.

종합하면, 본 연구가 제안한 설계 체계는 정렬 붕괴 문제에 접근하는 하나의 설계 논리적 응답으로 볼 수 있으며, 이후 경험적 검증과 교과 간 확장을 통해 그 타당성과 실용성이 보다 정교하게 평가될 수 있을 것이다.

* 본 논문은 2026.01.26. 투고되었으며, 2026.02.08. 심사가 시작되어 2026.03.07. 심사가 종료되었음.

참고문헌

교육부(2022), 『2022 개정 국어과 교육과정(교육부 고시 제2022-33호[별책5])』, 세종: 교육부.

김경희(2020), 「서·논술형 평가의 평가학적 의미 탐색」, 『교육평가연구』 33(4), 839-862.

김경희·이명진(2021), 「교수학습과 학생평가 개선을 위한 서·논술형 평가 지침 활용 및 피드백 효과 제고 방안」, 『교육과정평가연구』 24(3), 27-51.

김수진·김희경·나우열·민호기·백승주·성경희·이미숙·이민형·이영미·한금영(2025), 『서·논술형 평가에 대한 쟁점 및 요구 분석(KICE 이슈페이퍼ORM 2025-41-8)』, 진천: 한국교육과정평가원.

김영란(2021), 「학문 문식성(disciplinary literacy)의 의미와 중등교육에의 시사점」, 『리터러시 연구』 12(1), 367-401.

김형성(2024), 『국어 교사의 논술형 평가 전문성 신장 방안 연구』, 한국교원대학교 박사학위논문.

남가영·김호정(2023), 「서술형·논술형 평가 실행에 관한 국어 교사의 최적화 행동 분석」, 『교과교육학연구』 27(1), 31-50.

박고운(2025), 「GenAI-HITL 기반 '독서와 작문' 연계 서술형 평가 과제 개발 및 타당성 검토」, 『국어교육학연구』 60(4), 129-174.

박종임(2024), 「국어과 서·논술형 평가의 도입 현황 및 실행 상의 쟁점 탐색 연구」, 『청람어문교육』 101, 273-307.

박혜영·김성숙·김경희·이명진·김광규·김지영(2019), 『수업-평가 연계 강화를 통한 서·논술형 평가 내실화 방안(연구보고RRE 2019-6)』, 진천: 한국교육과정평가원.

송슬기(2024), 「깊이 있는 학습을 위한 필요조건으로서의 논술형 평가의 특징과 지원 방향에 관한 탐색」, 『교육문화연구』 30(4), 149-172.

송슬기(2025), 「개념적 학습을 유도하는 확장형 논술의 운영 조건 탐색」, 『열린교육연구』 33(1), 79-98.

장성민(2025), 「학문 문식성 기반의 논술형 평가 방향 탐색: 수능 서·논술형 평가 도입의 맥락에서」, 『리터러시 연구』 16(5), 623-657.

정민주·서수현·남민우·최숙기·이상일·남가영(2022), 「좋은 국어과 평가 문항 특성에 관한 질적 분석 연구: 국어과 평가 문항 양호도 분석틀 개발 연구(2)」, 『청람어문교육』 89, 7-42.

최숙기(2021), 「서·논술형 수능 도입을 대비한 2022 개정 국어과 교육과정의 개정 방향 탐색」, 『청람어문교육』 83, 129-156.

최숙기·박종임(2023ㄱ), 「2022 개정 국어과 교육과정 「독서와 작문」 교육과정 개발의 원리와 방향」, 『작문연구』 57, 165-199.

최숙기·박종임(2023ㄴ), 「인공지능 시대의 작문 평가를 위한 ChatGPT 활용 방안 연구」, 『청람어문교육』 95, 65-109.

편지윤(2021), 「학문 문식성 교육 내용으로서 지식에 대한 시론」, 『새국어교육』 129, 9-48.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014), *Standards for educational and psychological testing*, American Educational Research Association.

Bjork, E. L. & Bjork, R. A. (2011), "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning", In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz(Eds.), *Psychology and the Real World: Essays illustrating fundamental contributions to society* (pp. 56-64), Worth Publishers.

Bransford, J. D. & Schwartz, D. L. (1999), "Rethinking transfer: A simple proposal with multiple implications", *Review of Research in Education 24*, 61-100.

Brookhart, S. M. (2013), 『루브릭, 어떻게 만들고 사용할까?』, 장은경 · 김민아 · 남예지 · 양하늬 · 조은비 · 주혜란 · 차혜경(역), 서울: 우리학교, 2022.

Bruner, J. S. (1960), *The process of education*, Cambridge, MA: Harvard University Press.

Erickson, H. L., Lanning, L. A., & French, R. (2017), 『생각하는 교실을 위한 개념기반 교육과정 및 수업』, 온정덕 · 윤지영(역), 서울:학지사, 2019.

Fang, Z. (2012), "Language correlates of disciplinary literacy", *Topics in Language Disorders 32*(1), 19-34.

Fisher, D. & Frey, N. (2013), *Better learning through structured teaching: A framework for the gradual release of responsibility* (2nd ed.), ASCD, Alexandria.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004), "Design science in information systems research", *MIS Quarterly 28*(1), 75-105.

Jonsson, A. & Svingby, G. (2007), "The use of scoring rubrics: Reliability, validity and educational consequences", *Educational Research Review 2*(2), 130-144.

Kane, M. T. (2013), "Validating the interpretations and uses of test scores", *Journal of Educational Measurement 50*(1), 1-73.

Kapur, M. (2008), "Productive failure", *Cognition and Instruction* 26(3), 379-424.

Kapur, M. & Bielaczyc, K. (2012), "Designing for productive failure", *Journal of the Learning Sciences 21*(1), 45-83.

Messick, S. (1989), "Validity", In R. L. Linn(Ed.), *Educational measurement* (3rd ed., pp. 13-103), NY: American Council on education and Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003), "On the structure of educational assessments", *Measurement: Interdisciplinary Research and Perspectives 1*(1), 3-62.

Panadero, E. & Jonsson, A. (2013), "The use of scoring rubrics for formative assessment purposes revisited: A review", *Educational Research Review 9*, 129-144.

Scardamalia, M. & Bereiter, C. (1987), "Knowledge telling and knowledge transforming

in written composition", In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, Vol. 1. Disorders of first-language development; Vol. 2. Reading, writing, and language learning* (pp. 142-175). Cambridge University Press.

Shanahan, T. & Shanahan, C. (2008), "Teaching disciplinary literacy to adolescents: Rethinking content-area literacy", *Harvard Educational Review 78*(1), 40-59.

van de Pol, J., Volman, M., & Beishuizen, J. (2010), "Scaffolding in teacher-student interaction: A decade of research", *Educational Psychology Review 22*(3), 271-296.

Wineburg, S. S. (1991), "Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence", *Journal of Educational Psychology 83*(1), 73-87.

Wood, D., Bruner, J. S., & Ross, G. (1976), "The role of tutoring in problem solving", *Child Psychology & Psychiatry & Allied Disciplines 17*(2), 89-100.

〈부록 1〉 문항 설계

제시된 자료(가)~(다)를 꼼꼼히 읽고, 수치와 텍스트 이면에 숨겨진 논리적 허점을 분석하여 각 문항에 답하시오.

(가) 여론 조사의 원리와 한계

민주 사회에서는 선거를 통해 *민의(民意)를 반영한다. 그러나 선거는 몇 년에 한 번씩 치러지므로, 일상에서 나타나는 민의가 곧바로 정책에 반영되기 어렵다. 이러한 간격을 메우기 위해 여론이 활용된다. 여론은 어떤 공공 사안에 대해 여러 사람의 지지를 받고 있다고 인정되는 공통의 의견을 뜻하며, 여론 조사는 이를 수치로 파악하려는 장치다. 그런데 여론 조사는 조사 방법, 응답률, 질문 표현에 따라 결과가 달라질 수 있어, 수치만으로 판단하기는 어렵다.

여론 조사는 대체로 '표본 조사'로 이루어진다. 예를 들어 인구 100만 명이 사는 도시에서 전 시민을 조사하는 것은 현실적으로 불가능하므로, 일부 시민만 선정하여 전체의 특성을 추정한다. 이때 지역·성별·연령 등을 고려해 대표성을 확보하려 하지만, 응답률이 낮으면 문제가 생긴다. 전화 조사에서 '접속률'은 전화를 받는 비율을, '응답률'은 전화를 받아 답변을 완료한 비율을 말한다. 예를 들어 200명에게 전화했을 때 10명이 전화를 받고(접속률 5%) 그 중 1명만 답변을 완료한다면(응답률 10%) 실제 참여자는 1명에 불과하다. 응답자는 시간과 노력을 투자해야 하므로, 조사 주제에 관심이 많고 적극적인 사람들의 의견이 더 많이 반영될 가능성이 있다. 반대로 응답하지 않은 사람들은 해당 주제에 무관심하거나 반대 의견을 가졌을 가능성이 있어, 이들의 목소리는 결과에서 누락된다. 따라서 응답률이 낮으면 응답자 집단이 전체를 대표한다고 보기 어렵고, 여론 조사 결과를 '다수의 목소리'와 동일하다고 단정하기 어렵다.

질문 표현과 선택 항목도 결과에 영향을 미칠 수 있다. 2003년 1월 미국의 퓨 리서치 센터 조사에서는 '이라크의 사담 후세인 정권을 막기 위한 군사 행동에 찬성하는지' 묻자68%가 찬성했다. 그런데 같은 시기에 '미군 사상자 수천 명의 희생을 감수하고서도'라는 표현을 추가하여 묻자4 3%만 찬성했고 48%가 반대했다. 선택 항목도 마찬가지다. 갤럽은 사형에 대해 '찬성/반대/잘 모르겠다'로 묻던 때에는 60% 이상이 찬성했지만, '가석방 없는 종신형'을 항목에 추가하자 종신형(48%)이 사형(47%)보다 높게 나타났다. 질문의 표현이나 선택지 구성이 달라지면 응답도 달라질 수 있다.

따라서 여론 조사 결과를 해석할 때는 누가 의뢰했는지, 어떤 조사 방식이 사용되었는지, 응답률은 어느 정도인지 같은 외적 요소를 확인할 필요가 있다. 더 나아가 질문지의 표현과 선택 항목 같은 내적 요소도 결과를 바꿀 수 있음을 고려해야 한다. 여론 조사는 대중의 의견을 반영하기만 하는 것이 아니라 대중의 의견에 영향을 미친다. 잘못된 여론 조사는 편파적 결과를 만들고, 그 결과가 다시 대중에게 영향을 주어 그릇된 여론을 형성하게 할 수도 있다. 따라서 조사 과정과 결과에 영향을 미치는 요소들이 공정하게 작용하는지 감시하고, 주도적으로 자신의 목소리를 내는 태도는 민주 시민에게 필요한 자질이자 태도이다.

*민의(民意): 사회 구성원의 공적 의사(여론)를 뜻함. 다만 설문·여론조사 수치가 민의를 대표한다고 말하려면 표본, 응답률, 질문 문장 등 '조건 정보'가 함께 제시되어야 함

(나) 교실 토론: "그 설문, 진짜 '우리 학년 여론' 맞아?"

〔국어 수업 중 교실 토론 발췌〕

교사: 우리 학년에서 "휴대전화 수거 방식"설문을 했지? 결과가 "대부분 찬성"(찬성 72%)이라고 공유됐는데, 이게 정말 믿을 만한 결과일까? 숫자만 보고 판단하면 놓치는 게 있을 수 있어.

학생 A: 네. 단톡방에 링크 올라왔고, 질문이 "수업 중 휴대전화 전면 수거 찬성?"이런 식이었어요. 찬성이 72%라서 다들 "그럼 끝난 거네"라고 했고요.

학생 B: 근데 저희 반만 해도 링크 못 본 애들 많아요. 알림 꺼 둔 애도 있고, 시험기간이라 바빠서

그냥 넘긴 애도 있고요.

교사: 바로 그거야. B가 말한 게 핵심이야. "누가 참여했고, 누가 참여하지 않았는지"를 먼저 봐야 해. 설문에 응답한 사람들이 전체를 대표한다고 바로 말하긴 어렵거든. 특히 관심 있는 학생들만 많이 참여하면 결과가 한쪽으로 쏠릴 수 있어.

학생 C: 그리고 질문이 좀 애매했어요. "전면 수거"는 싫지만 "수업 시간엔 잠깐 모아 두는 건 괜찮다"는 애들도 있거든요. 근데 선택지가 찬/반만 있으니까, 자기 생각이 딱 맞는 데가 없어요.

교사: 맞아. 질문 문장이랑 선택지가 결과를 바꿀 수 있어. 같은 주제라도 단어를 어떻게 쓰느냐, 조건을 어떻게 붙이느냐에 따라 사람들 생각이 달라지거든.

〔학년 단체 채팅방 일부 대화 화면〕
학생 D: "찬성 72%" 떴대, 이제 전면 수거 확정 아님?
학생 E: 근데 참여 인원이 90명 중 25명이라던데... 그럼 72%는 25명 중 18명이라는 거잖아.
학생 F: 질문 순서도 좀 그랬대, "휴대전화 때문에 수업 집중 깨진 적 있어요?" 먼저 묻고, 바로 "전면 수거 찬성?" 물었대. 그럼 답이 더 한쪽으로 갈 수도 있지 않나?.

교사: 지금 나온 얘기들을 정리해 보면, 몇 가지 점검할 게 보이네. 첫째, 응답한 사람들이 전체를 대표하는가. 둘째, 응답률이 어느 정도인가. 셋째, 질문 설계가 공정한가. 넷째, 결과를 퍼뜨릴 때 "확정"처럼 말해서 분위기를 만들고 있는 건 아닌가, 이런 것들이야.

학생 A: 그럼 설문을 제대로 하려면 어떻게 해야 해요?

교사: 최소한 이런 게 필요할 것 같아. 링크를 한 곳에만 올리지 말고 여러 경로로 알려서 참여 기회를 넓히는 것, "전면 수거"처럼 포괄적으로 묻지 말고 상황을 나눠서 묻는 것, 응답자 수랑 응답률을 함께 공개하는 것, 그리고 결과를 "결론"이 아니라 "참고 자료"로 표현하는 것. 이 정도는 해야 설문이 제대로 기능할 수 있지.

(다) 뉴스 해설: "여론조사, 조건부 근거로 읽어야"

〔앵커〕 선거철이면 여론조사 결과가 연일 보도됩니다. 하지만 이 수치는 단순한 '정보'가 아니라, 유권자의 판단과 후보의 전략에도 영향을 주는 사회적 신호이기도 합니다. 그래서 여론조사를 볼 때는 "누가 앞서나"만 확인하기보다, "이 결과가 어떤 절차로 만들어졌는지"를 함께 따져볼 필요가 있습니다.

〔기자〕 문제는 일부 보도가 "A 후보 40%, B 후보3 8%... A 우세"처럼 숫자만 앞세우는 경우가 있다는 점입니다. 중앙선거관리위원회 안내에 따르면, 여론조사 결과를 보도할 때는 결과 수치와 함께 조사의뢰자, 조사기관, 표본 크기, 조사 방법, 표본오차율, 응답률, 질문 내용 등을 함께 제시해야 합니다. 화면에서 보이는 체크리스트가 바로 그것입니다.

여론조사 보도 시 공표 기준(체크리스트) –중앙선거관리위원회 안내 참조 재구성

구분	공표 항목	체크	왜 필요한가(해석 포인트)
1	조사의뢰자	□	이해관계/목적(프레이밍) 점검
2	조사기관	□	조사 수행의 전문성·책임 소재
3	표본 크기(n)	□	표본오차·신뢰도 판단의 기초
4	조사 방법(예: 전화/온라인, 표집 방식 등)	□	대표성·편향 가능성 점검
5	표본오차(신뢰수준 포함)	□	격차가'우열'인지'오차 범위 내 흔들림'인지 판단
6	응답률	□	과소/과대 대표, 응답자 편향 점검
7	질문 문장(주요 문항)	□	유도 표현/전제/선택지 설계에 따른 결과 변형 점검

위7개 항목이 누락되면 여론조사 수치는'정답'이 아니라'조건부 근거'로서의 해석 가능성이 크게 제한된다고 설명합니다.

〔전문가〕 이러한 정보는 결과 해석에 직접 영향을 미치는 핵심 단서입니다. 예를 들어 표본 크기와 표본오차가 빠지면, 격차가 '의미 있는 차이'인지 '오차 범위 안의 흔들림'인지 판단하기 어렵습니다. 즉, 표본오차가 ±3%p인데 후보 간 격차가 2%p라면, 이는 통계적으로 우열을 가릴 수 없는 '오차 범위 내 흔들림'으로 해석해야 합니다. 또한 응답률이 제시되지 않으면 특정 집단이 과소·과대 대표되었을 가능성을 점검하기도 어렵습니다.

특히 질문 표현이 다소만 달라져도 응답이 달라질 수 있기 때문에, 여론조사를 근거로 주장을 펼칠 때는 질문 문장 자체가 어떤 방향의 해석을 유도하는지도 함께 살펴봐야 합니다. 실제로 동일한 정책에 대해 긍정적 결과 표현을 강조하는 질문과 부정적 결과 표현을 강조하는 질문은 찬반 비율을 크게 바꿀 수 있습니다. 여론조사 결과를 보도할 때 질문 문장을 함께 공개하지 않으면, 수치만으로는 그 결과가 중립적 질문에서 나온 것인지, 특정 방향을 유도한 질문에서 나온 것인지 판단할 수 없습니다. 조건 정보가 투명하게 공개되지 않으면 숫자는 민의의 창이 아니라 특정 해석을 유도하는 장치가 될 위험이 있습니다.

〔앵커〕 그렇다면 여론조사 기사를 읽을 때 최소한 무엇을 확인해야 할까요. 첫째, 의뢰자·기관·표본·방법·오차·응답률·질문 같은 설명 정보가 함께 제시되어 있는지 확인합니다. 둘째, 수치 차이가 표본오차 범위를 넘어서는지 점검합니다. 셋째, 질문 문장의 표현과 선택지가 무엇을 전제하는지 살펴봅니다. 넷째, 이 결과를 근거로 사용하려면 반례나 대안 자료도 함께 검토합니다.

결국 여론조사 결과는 '정답'이 아니라 '조건부 근거'입니다. 표본과 방법, 응답률, 질문 내용 같은 조건이 투명하게 공개될 때에만 그 수치를 비판적으로 해석하고 논증의 근거로 사용할 수 있습니다. 반대로 이러한 조건이 생략되면 여론조사는 '민의의 창'이 아니라 특정 해석을 유도하는 장치가 될 위험도 있습니다.

1. (서술형)

아래〈보기〉는 우리 학년 학생회에서 '교내 생활 규정 개정'을 위해 실시한 온라인 설문의 실제 참여 결과 보고서입니다. 이 결과를 보고 "학년 전체의 지배적인 여론이다"라고 단정하기 어려운 이유를 (가)에서 근거로 서로 다른 2가지 관점에서 설명하시오. (4점)

〈보 기〉 여론조사 참여 현황 보고서(Data Analytics)

구분	통계 수치	비고
전체 대상(N)	500명	학교 전체 학생 수
링크 클릭(접속)	100명	온라인 설문 접속자 수(접속률 20%)
설문 완료(응답)	15명	최종 유효 응답자 수(응답률 15%)
최종 반영률	3.0%	전체 대상 대비 유효 응답 비율

〈답안 작성 조건〉

① (가)의 '접속률'과 '응답률(및 반영률)' 개념을 사용해 〈보기〉 수치를 해석할 것.
② 그 해석을 바탕으로, 결과를 "학년 전체의 지배적 여론"으로 단정하기 어려운 이유를 서로 다른 2가지 관점(기준)에서 설명하되, 각 관점이 (가)의 여론조사 해석 원리와 연결되도록 제시할 것.

2. (서술형)

다음 〈보기〉는 학생회에서 휴대폰 관리 방식에 대한 의견을 묻기 위해 작성한 설문 초안입니다. 지문 (가)와 (나)에 제시된 질문 표현 및 선택지 구성의 원리를 참고하여, 가장 중립적인 질문 1개를 선택하고 그 이유를 서로 다른 2가지 관점(기준)에서 제시하시오. 또한 가장 편향된 질문 1개를 골라 중립적인 질문으로 고쳐 쓰시오. (6점)

〈보 기〉 설문 문항 초안 3종

- 문항 1: "학생들의 자유를 위해 휴대폰 수거를 당장 없애야 하지 않나요?" ( ) 찬성 ( ) 반대
- 문항 2: "수업 중 휴대폰 관리 방식에 대해, 다음 중 어디에 더 동의하나요?"
  ( ) 전면 수거 유지 ( ) 부분 허용 ( ) 잘 모르겠다
- 문항 3: "휴대폰을 수거하면 수업 집중이 높아진다는 의견에 동의하나요?"
  (1점: 매우 동의 ~ 4점: 매우 비동의)

〈답안 작성 조건〉

① 중립 질문 선택의 근거를 (가) 또는 (나)에서 찾아 서로 다른 2가지 기준(관점)으로 제시할 것(각각 지문 기호 명시).
② 두 기준은 '질문 문장의 유도성'과 ' 선택지 구성의 공정성'을 각각 적용하되, 지문의 원리를 활용해 자신의 말로 재구성할 것.
③ 가장 편향된 질문 1개를 선정하여 중립적인 1 문장으로 수정하고, 필요 시 선택지 수정도 포함할 것.

3. (논술형)

지문 (다)의 뉴스 해설을 읽고, 아래 〈보기〉의 보도가 여론조사를 '조건부 근거'로 적절히 제시했는지 평가하시오. 단, 답안 작성 시 본인의 학교·학급·온라인 커뮤니티 등에서 '통계 수치나 설문 결과'가 조건 정보 없이 단정적으로 공유되어 혼란이나 오해가 생겼던 경험(또는 그럴 법한 상황)을 출발점으로 삼아 문제의식을 제시한 뒤, 논증적 글쓰기 방식으로 작성하시오. (10점)

〈보 기〉 뉴스 보도 자료(TV 뉴스 방송 화면 캡처)

뉴스 앵커: "최근 실시된 학생회장 후보 선호도 조사 결과입니다. A 후보가 40%의 지지를 얻어 38%를 얻은 B 후보를 앞서고 있는 것으로 나타났습니다. 이번 결과로 선거 판세가 A 후보 쪽으로 기울었다는 분석이 나옵니다.
자막: "A 후보 40%, B 후보 38%...A 앞섰다"

〈답안 작성 조건〉

논증적 글쓰기 방식으로 작성하되, 다음 요소를 모두 포함할 것.

〔필수 포함 요소(순서 무관, 논리적으로 구성할 것)〕

① 개인 경험 또는 상황을 활용하여 문제의식을 제시하고, 〈보기〉 보도에 대해 타당/미흡을 판정할 것
② (가), (나), (다)를 바탕으로 근거 3가지를 제시하되, 그 중 최소 1가지는 (다)의 관점을 적용하여 〈보기〉에서 조건 정보의 제시 여부를 점검하고, 그 점검 결과에 근거해 격차 해석의 타당성을 논할 것.
③ 예상 반론과 재반박을 포함할 것.
④ 개선 방안을 2가지 이상 제시할 것(그 중 1가지는 자신이 지적한 문제점과 직접 연결되도록 구체화할 것).

〔형식 요건〕

① 근거를 제시할 때 지문의 기호((가), (나), (다))를 명시할 것.
② 단락 구분과 논리적 완결성을 갖출 것.

〈부록 2〉 루브릭과 예시 답안

■ 채점 루브릭

1(서술형)

총 배점	채점 요소	배점	내용 요소 및 배점
4점	수치·개념 적용	2점	접속률 20%(100/500), 응답률 15%(15/100), 반영률 3.0%(15/500)를 (가)의 개념으로 해석하고, 응답자 편향 가능 및 전체 대표성 제약(반영률 3.0%)을 논지에 연결함
		1점	수치 언급은 있으나 접속률/응답률/반영률의 개념 구분이 불명확하거나 (가) 개념과의 연결이 부분적임
		0점	"적다/믿을 수 없다" 수준 일반론, 수치·개념 적용 없음
		부분점 규칙	개념+수치 연결 명확(2점)/ 부분 연결(1점)/ 없음(0점)
	관점 2가지 (서로 다름)	2점	(가)의 노리를 근거로 서로 다른 관점(기준) 2가지 제시(각 1점) ① 응답자 편향(관심/적극 집단 과대 반영) ② 무응답자 누락(무관심/반대 집단 과소 반영) 등
		1점	2가지 처럼 보이나 실질적으로 같은 말 반복, 또는 1가지만 명확
		0점	관점 1개 이하 또는 근거 없음
		부분점 규칙	관점 2개 모두 '서로 다름'(2점)/ 1개만 타당(1점)/ 없음(0점)

2(서술형)

총 배점	채점 요소	배점	내용 요소 및 배점
6점	중립 질문 선택	2점	2번 선택
		1점	3번 선택 + 전제/유도 가능 문제를 1가지 이상 지적
		0점	1번 선택 또는 근거 없이 선택
		부분점 규칙	2번(2점)/ 3번 + 문제지적(1점)/ 그 외(0점)
	중립성 이유 (서로 다른 기준 2가지)	2점	(가), (나)를 근거로 서로 다른 기준 2가지 제시(각 1점)
		1점	타당한 기준 1개만 명확, 또는 2개를 제시했으나 동일 기준 반복/추상적 진술
		0점	근거 없이 "중립적"만 진술
		부분점 규칙	이유 2개 각각 1점(2점)/ 추상·반복이면 해당 이유 0점

	편향 문항 수정 (문장+선택지)	2점	질문 문장 중립화(1점) + 선택지 3개 이상으로 균형 제시(1점)
		1점	문장 중립화는 되었으나 선택지 구성이 제한적(2개) 또는 문장에 유도/전제 일부 잔존
		0점	유도/전제 강화 또는 선택지 2개 이하
		부분점 규칙	문장 중립화 1점 + 선택지 3개 이상 1점

3(논술형)

총 배점	채점 요소	배점	내용 요소 및 배점
10점	문제 의식 (경험 기반)	1점	개인 경험/상황 1개를 1~2문장으로 제시하고, '조건 정보 없이 단정적 공유 → 혼란/오해'를 이번 과제(여론조사 보도 평가)와 명시적으로 연결함(1점)
			(1점 요소는 정수 운영을 위해) 연결이 약하거나 나열 수준이면 0점 처리
			경험/상황 제시 없음 또는 과제와 무관(0점)
	판정의 명확성	1점	〈보기〉 보도를 타당/미흡 중 하나로 명확히 판정하며, 판정 기준이 '조건부 근거' 관점임이 드러남(1점)
			(1점 요소는 정수 운영) 모호하면 0점 처리
			판정 부재(0점)
	근거의 충족·적절성	4점	(가), (나), (다)를 활용해 서로 다른 근거 3가지 이상 제시하고, 그 중 최소 1개에서 (다)의 '표본오차-격차 비교' 논리를 적 용해 우열 단정 가능/불가능을 판단함. 근거들이 판정과 논리적 으로 연결됨
			근거가 2~3개 수준이거나 반복/피상적임, 또는 표본오차-격차 비교가 누락/오류/피상적 언급에 그침
			근거 1개 이하 또는 "정보 부족" 같은 일반론만 제시(0점)
	반론·재반박	2점	예상 반론을 구체화하고, (가), (나), (다)의 사례·개념을 활용해 재반박이 논리적으로 연결됨(2점)
			반론/재반박 중 하나가 약함(형식적 반론, 재반박 일반론) 또는 지문 활용 피상적(1점)
			반론·재반박이 사실상 없음(누락 포함)(0점)
	개선 방안의 구체성	2점	개선 방안 2가지 이상 제시하며, 그 중 최소 1가지는 보도에서 '무엇을 어떻게' 바꿀지가 실행 수준으로 드러남(자막/원고/ 공표 항목/표기 방식 등 "형식 예시"는 가능하되, 핵심은 실행가능한 구체성)(2점)
			2가지를 제시했으나 모두 추상적("정확히/공정하게" 수준) 또는 1가지만 제시(1점)
			개선 방안 없음(0점)

- 형식 요건(감점): 정수 운영으로 정리

감점 요소	감점 배점	내용
출처 (지문 기호) 미표기	1점	근거·재반박·개선안의 핵심 주장에 (가)/(나)/(다) 표기가 거의 없어서 출처 식별이 어려운 경우에 적용
단락·논리적 완결성 미흡	1점	단락 구분이 없거나 논지가 뒤섞여 판전–근거–반론–개선 흐름이 식별되지 않는 경우 적용

※ 감점은 최대 -2점, 최저점은 0점

- '근거의 충족·적절성' 항목 정수화 가이드(핵심 변별점)

배점	내용
4점	근거 3개 이상 + (다) 표본 오차–격차 비교를 논리 적용(우열 단정 가능/ 불가능 판단까지) + 판정과 연결 명확
3점	근거 3개 이상이나 (다) 비교가 언급 수준/ 연결 약함, 또는 근거 중 1개가 피상적
2점	근거 2개 수준(또는 3개지만 반복/피상적) + (다) 비교 누락/ 오류
1점	근거 1개 수준(일반론 위주)
0점	일반론

■ 예시 답안

1(서술형)

수준	답안	채점 근거	
상 (4점)	이 결과를 학년 전체의 지배적 여론으로 단정하기 어려운 이유는 다음과 같다. 첫째, (가)에 따르면 응답률이 낮을수록 조사 주제에 관심이 많고 적극적인 집단의 의견이 상대적으로 더 반영될 수 있다. 〈보기〉에서 전체 500명 중 접속은 100명(20%)이지만 그중 응답 완료는 15명으로 응답률이 15%(15/100)에 그쳐, 응답자 집단이 특정 성향으로 편향되었을 가능성이 있다. 둘째, 응답라지 않은 학생들은 무관심하거나 반대 의견을 가졌을 수 있는데, 이들이 제외되면 최종 반영률이 3.0%(15/500)에 불과하므로 학년 전체를 대표한다고 보기 어렵다.	수치·개념 적용	접속률·응답률·반영률을 구분하여 (가)의 개념으로 해석하고 논지에 연결함(응답률 15%, 반영률 3.0%를 정확히 활용)
		관점 2가지	① 응답자 편향 ② 무응답자 누락을 서로 관점으로 제시함

중 (2점)	응답한 학생이 15명이라 전체 학년의 의견이라고 단정하기 어렵다. (가)에서 응답률이 낮으면 관심 있는 학생들 의견이 더 반영될 수 있다고 했으므로, 응답자 편향이 생길 수 있다. 또한 응답하지 않은 학생들의 의견이 빠지면 대표성이 약해질 수 있다.	수치·개념 적용	15명 수치를 언급하고 (가)와 연결하나, 응답률(15%)·반영률(3%) 구분 및 논지 연결이 약함
		관점 2가지	"응답자 편향/의견 누락"을 말했으나 두 관점의 구분이 선명하지 않고 동일 논리로 반복되는 수준임.
하 (0점)	참여자가 너무 적어서 믿을 수 없다. 전체 여론이라고 보기 어렵다.	수치·개념 적용	수치·개념 적용 없음
		관점 2가지	서로 다른 과점 제시 없음

2(서술형)

수준	답안	채점 근거	
상 (6점)	가장 중립적인 질문은 2번이다. 첫째, (가)에 따르면 질문 표현은 응답을 특정 방향으로 유도할 수 있는데, 2번은 "학생들의 자유를 위해", "당장 없애야"처럼 결론을 전제하거나 동의를 유도하는 표현이 없어 유도성이 상대적으로 낮다. 둘째, (나)에서 지적하듯 선택지가 찬/반만 있으면 중간 입장이 배제되는데, 2번은 부분 허용이나 유보 입장까지 포함할 수 있어 선택지 구성이 더 공정하다. 가장 편향된 질문은 1번이다. 이를 중립적으로 수정하면 "수업 중 휴대폰 관리 방식에 대해 어떻게 생각하나요?"로 바꿀 수 있다. 선택지는 ① 전면 수거 유지 ② 부분 허용 ③ 수거하지 않음 ④ 잘 모르겠다로 제시한다.	중립 질문 선택	2번 선택
		중립성 이유	① 유도성(질문 문장) ② 선택지 공정성(스펙트럼)으로 서로 다른 기준 제시
		편향 문항 수정	문장 중립화+ 선택지3개 이상 균형 제시
중 (4점)	중립적인 질문은 2번이다. 유도하는 표현이 적고 선택지가 여러 입장을 담을 수 있어 중립적이라고 생각한다. 1번을 수정하면 "휴대폰 수거에 대해 어떻게 생각하나요?"로 바꿀 수 있고, 선택지는 ① 찬성 ② 반대 ③ 잘 모르겠다로 제시한다.	중립 질문 선택	2번 선택
		중립성 이유	"유도 표현이 적다"는 기준 1개는 성립하나, 다른 기준은 (가), (나) 근거 연결이 약함/추상적임
		편향 문항 수정	문장 중립화 1점은 충족하나, 선택지가 3개이긴 해도 '관리 방식'의 스펙트럼(전면/부분/미수거 등)을 충분히 반영하지 못해 균형 제시로 보기 어려움 (0점)

하 (0점)	중립적인 것은1번이다. 학생들의 자유를 중요하게 생각하기 때문이다. 3번을 수정하면“휴대폰 수거에 찬성하나요?”(선택지: 찬성/반재)이다.	중립 질문 선택	1번 선택
		중립성 이유	개인 가치 진술로 기준 충족 불가
		편향 문항 수정	유도 강화 + 선택지 2개(응답 강요)

3(논술형)

수준	답안
상 (10점)	- (1점)(문제의식) 우리 반 단톡방에서 설문 결과가 “찬성85%”로만 공유되어 이미 결정된 사안처럼 오해가 퍼진 적이 있다. 참여 규모나 질문 문장이 공개되지 않으면 수치가 단정처럼 작동해 혼란이 커진다. - (1점)(판정) 〈보기〉 보도는 여론조사를‘조건부 근거’로 제시하지 못해 미흡하다. - (4점)(근거1 : (다)) (다)는 여론조사 보도에서 조사기관, 표본 규모, 조사 방법, 질문 문장, 표본오차 등 조건 정보를 함께 제시해야 한다고 한다. 그런데 〈보기〉는 결과 수치만 제시해, 시청자가 신뢰성과 대표성을 판단할 근거가 부족하다((다)). - (근거2: (다) 표본오차 – 격차 비교) 특히 표본오차 정보가 제시되지 않으면A(40%)와B(38%)의 2%p 차이를 ‘우세’로 단정할 수 있는지 판단하기 어렵다. (다)의 예시처럼 표본오차가 ± 3%p 수준이라면 2%p 격차는 오차 범위 안에서 흔들릴 수 있어 우열을 단정하기 어렵다. 그럼에도 〈보기〉는“A 앞섰다”처럼 단정적으로 제시해 해석을 과도하게 끌고 간다((다)). - (근거3: (가)) (가)에서처럼 질문 표현과 조건이 달라지면 응답이 크게 달라질 수 있는데, 〈보기〉는 질문 문장을 제시하지 않아 결과가 어떤 방식으로 형성되었는지 점검할 수 없다((가)). - (2점)(반론·재반박) “대략 흐름만 보면 되니 숫자만으로도 충분하다”는 반론이 가능하다. 그러나 (가)의 사례처럼 질문 설계에 따라 응답이 달라질 수 있고, (다)가 요구하는 조건 정보가 빠진 상태에서 수치만 제시되면 ‘근거’가 아니라 ‘결론’처럼 작동해 오해를 확대할 수 있다((가)(다)). - (2점)(개선) 첫째, 자막에 표본오차와 조사 기본 조건(표본 수, 조사기관 등)을 함께 제시하고, 격차가 오차 범위 내일 가능성이 있으면 단정 표현을 피해야 한다. 둘째, 원고/보도 본문에서 질문 문장과 조사 방법, 응답률 등 조건 정보를 명시해 시청자가 조건을 확인한 뒤 판단하도록 해야 한다.
중 (6점)	- (0점)(문제의식) 예전에 설문 결과만 보고 결정된 줄 알아 혼란스러웠던 적이 있다. - (1점)(판정) 이 보도는 미흡하다. - (1점)(근거) (다)에서 여론조사는 조사기관, 표본, 오차, 질문 등을 알려야 한다고 했는데 〈보기〉는 그런 정보가 없다((다)). 또 질문이 없어서(가)처럼 질문에 따라 결과가 달라질 수 있는지도 모른다((가)). - (2점)(반론·재반박) 숫자만 보면 된다고 해도 질문이 달라지면 결과가 달라질 수 있으니 조건 정보가 필요하다. - (2점)(개선) 표본오차를 자막에 넣고, 조사기관과 질문 문장을 원고에 쓰면 좋다.
하 (0점)	- (문제의식) 설문 결과가 이상했던 적이 있다. - 이 보도는 문제가 있다. 정보가 부족하기 때문이다. 여론조사는 조심해야 한다. - 개선하려면 정확하게 보도해야 한다.

초록

전이(Transfer) 평가를 위한 세트형 서·논술형 과제 설계 체계 연구: 구인·과제·증거의 정렬(Alignment) 메커니즘을 중심으로

박고운

본 연구는 깊이 있는 학습을 요구하는 서·논술형 평가가 현장에서 단답화·형식화되는 현상을 '정렬 붕괴'로 진단하고, 전이(transfer) 수행을 관찰 가능한 증거로 포착하기 위한 세트형 과제 설계 체계를 제안하는 데 목적을 둔다. 증거중심설계(ECD)와 개념기반 탐구학습(CBIL)을 결합하여 구인–과제–증거의 정렬을 복원하는 3층위 설계 모델을 구축하고, 2022 개정 '독서와 작문' 맥락에서 여론조사 자료 비평 세트를 설계 사례로 제시한다. 전이를 '조건부 근거 판단'으로 운영화하고, 지문 기능 분업(개념 토대 → 맥락 적용 → 공적 규범)과 문항 간 책임 이전을 통해 사고 경로를 누적하는 과제 구조를 제시하였다. 루브릭은 관점/기준 이원화 원리와 CBIL 경로 기반 질적 차등으로 설계하며, 역방향 검증을 통해 구인 충실성과 구인 무관 분산(CIV) 통제를 점검한다. 본 연구는 서·논술형 평가를 '정렬 가능한 설계 체계'로 재정식화하며, 전이 수행을 증거화하는 과정적 평가 구조의 이론적·실천적 토대를 제공한다. 후속 연구는 학습자 수행 데이터와 채점자 간 신뢰도 검증을 통한 경험적 타당화가 필요하다.

핵심어 서·논술형 평가, 전이(transfer), 증거중심설계(ECD), 개념기반 탐구학습(CBIL), 정렬(alignment), 학문 문식성, 독서와 작문

ABSTRACT

Designing Set-Based Constructed-Response Tasks for Transfer Assessment: An Alignment Framework of Construct, Task, and Evidence

Park Goun

This study diagnoses the recurring simplification and formalization of constructed-response assessments as "alignment collapse" and proposes a set-based task design framework to capture transfer performance as observable evidence. By integrating Evidence-Centered Design (ECD) and Concept-Based Inquiry Learning (CBIL), this study develops a three-layer design model that restores construct-task-evidence alignment and presents a design case of opinion poll critique task sets within the 2022 revised Korean Language Arts curriculum's "Reading and Writing" course. The study operationalizes transfer as "conditional evidence-based reasoning" and proposes a task structure that accumulates cognitive pathways through text function division (conceptual foundation → contextual application → public norms) and the gradual transfer of responsibility across items. The rubric employs dual principles of perspective/criteria bifurcation and CBIL pathway-based qualitative differentiation, with backward validation procedures to ensure construct fidelity and control for construct-irrelevant variance (CIV). This study reconceptualizes constructed-response assessment as an "alignment-ready design system" rather than a mere item format, providing the theoretical and practical foundations for process-oriented assessment structures that overcome single-item limitations. Future research should pursue empirical validation through student performance data and inter-rater reliability assessment.

KEYWORDS Constructed-response assessment, transfer, evidence-centered design (ECD), concept-based inquiry learning (CBIL), alignment, disciplinary literacy, reading and writing