

무선 할당 설계와 MC+나선형 설계의 채점자 편향 탐지 성능 비교

권순보 (한국교원대학교 시간강사)*

요약

본 연구는 피험자 수, 채점자 수 및 편향 크기를 달리한 모의실험을 통해 무선 할당 설계와 MC+나선형 설계의 채점자 편향 탐지 성능을 비교하였다. 자료는 GPCM에 기반한 채점자 모형으로 생성하였다. 5개의 피험자 수-채점자 수 조합과 3개의 편향 크기를 교차한 15개 기본 조건에 완전 교차, 무선 할당 및 MC+나선형 설계를 적용하여 총 45개 설계 조건을 구성하고, 각 조건을 100회 반복하였다. 채점자 분류 정확도, 수렴률, 평균제곱오차 및 경험적 표준오차를 산출한 결과, 두 불완전 설계의 성능 차이는 표본 조건에 따라 달라졌다. 피험자 수가 충분하고 채점자 1인당 배정량이 많은 조건에서는 두 설계의 차이가 거의 없었다. 반면 채점자 1인당 배정량이 적고 무선 할당 설계에서 채점자 쌍이 공유하는 답안이 부족한 조건에서는 MC+나선형 설계의 수렴률이 더 높고 평균제곱오차와 경험적 표준오차가 대체로 작았다. 편향 크기에 따른 설계 간 차이는 일부 조건에서 분류 정확도보다 추정 안정성 지표에서 상대적으로 뚜렷하게 나타났다. 채점자 1인당 배정량과 채점자 쌍당 공유 정보가 제한된 조건에서 MC+나선형 설계가 상대적으로 안정적인 추정을 보인 결과는 연결성이 희소한 혼합형 검사의 채점 설계를 결정할 때 고려할 필요가 있음을 시사한다.

주제어 : 채점자 편향, 채점 설계, MC+나선형 설계, 무선 할당 설계, 문항반응이론

* 제1저자 및 교신저자, sbkwon@knue.ac.kr

I. 서론

2022 개정 교육과정은 단편적 지식의 암기를 지양하고 문제 해결 및 사고 과정을 중시하는 평가를 강조하고 있다(교육부, 2022). 이러한 흐름 속에서 고차원적 사고력과 문제해결력을 평가할 수 있는 구성형 문항의 중요성이 더욱 확대되고 있다. 구성형 문항은 선택형 문항으로는 포착하기 어려운 학생의 사고 깊이와 표현 능력을 직접 확인할 수 있어(Ercikan et al., 1998; Wind & Ge, 2024), 학생이 직접 답안을 구성하고 서술하는 과정을 통해 논리적 추론 능력과 같은 역량을 측정할 수 있다는 점에서 주목받고 있다(하민수, 최숙기, 2025).

2022 개정 교육과정이 깊이 있는 학습과 문제 해결 역량 향상을 핵심 방향으로 제시함에 따라(김진희 외, 2025), 서·논술형 평가의 중요성은 한층 높아지고 있으며, 정부 역시 2028학년도 대학입시제도 개편 확정안에서 미래 핵심 역량 신장을 위해 서·논술형 평가를 확대하겠다는 방침을 발표하였다(교육부, 2023). 실제로 2011년 당시 우리나라 국가수준 학업성취도 평가(National Assessment of Educational Achievement: NAEA)의 서답형 비율이 과목별로 13~20%에 그쳐 해외 대규모 평가(50% 내외)의 절반에도 미치지 못하였던 것과 비교하면(심재호, 노은희, 김명화, 2012), 현재 전체 문항 중 약 40%를 단답형·서술형 문항으로 구성하고 있는 수준으로 지난 10여 년간 상당한 정책적 진전이 이루어졌다(신진아 외, 2025).

이 중 서술형 문항은 채점자의 판단이 개입되는 반응 특성상 인간 채점자의 전문적 판단에 의존하는 구조를 유지하고 있다. 이러한 서술형 문항은 선택형 문항과 함께 혼합형 검사(mixed-format assessment)로 구성할 수 있다. 국가수준 학업성취도 평가 역시 약 40%의 단답형·서술형 문항과 60% 정도의 선택형 문항으로 이루어져 혼합형 검사에 해당된다. 혼합형 검사 구조는 채점 문제에 상반된 두 가지 함의를 가진다. 먼저, 구성형 문항의 채점은 인간의 판단에 의존하므로 채점자에 따른 측정 오차가 개입될 수 있다(Guo & Wind, 2023). 다음으로 선택형 문항은 채점자의 개입 없이 즉시 채점되므로, 서술형 채점이 시작되기 이전에 이미 피험자의 능력에 관한 정보를 제공한다. 혼합형 검사에서 선택형 문항 점수는 추가 비용 없이 확보되는 사전 정보이며, 채점자에게 답안을 배정하는 단계에서 활용될 수 있는 자원이다.

채점자 편향은 채점자 간 엄격성(severity)과 관대성(leniency)의 차이, 중심화 경향(central tendency), 비일관성(misfit) 등 다양한 형태로 나타난다(Engelhard, 1994; Myford & Wolfe, 2003). 이는 피험자의 실제 수행 수준과 무관한 체계적 점수 차이를 유발하여 평가 결과의 일관성을 저해할 수 있다(Lunz & Stahl, 1993).

채점자 편향은 학생의 성취도 추정에 영향을 미치고(Wind, 2019), 성취 수준 분류 결과에도 영향을 줄 수 있다. 이러한 문제는 오래전부터 구조적으로 지적되어 왔는데, 현장 교사들을 대상으로 한 면담 연구에서도 채점의 부담과 신뢰도 문제가 서답형 문항 활용을 억제하는 요인으로 일관되게 보고된 바 있다(심재호, 노은희, 김명화, 2012). 한국어 쓰기 평가 연구에서도 가장 엄격한 채점자가 채점할 경우 학습자 점수가 평균 40% 가량 감소할 수 있음이 보고될 만큼(강민석, 2024), 채점자 편향이 점수에 미치는 영향은 결코 사소하지 않다. 채점자 편향이 초래하는 문제는 대규모 평가 맥락에서 더욱 심각하게 부각된다. 실제로 대규모 국가수준 평가 환경에서는 수십 명의 채점자가 수만 건의 답안을 처리해야 하는 상황에 직면하게 되며(신진아 외, 2025), 이러한 환경에서 채점자 간 일관성의 확보는 평가 공정성의 핵심 과제로 대두된다.

채점자 편향을 구조적으로 통제하는 핵심 수단은 채점 설계(rating design)이다. 채점 설계는 피험자 응답을 채점자에게 배정하는 방식을 사전에 계획하는 절차로서, 채점자 간 비교 가능성, 편향 탐지 민감도, 채점 부담, 운영 효율성 사이의 균형을 고려해야 한다(Barnett, 2005; Guo & Wind, 2023). 이론적으로 이상적인 방식은 완전 교차 설계(complete rating design)로, 모든 채점자가 모든 피험자 응답을 채점함으로써 채점자 간 연결성(connectivity)이 완전히 확보되고 편향 탐지 민감도가 가장 높을 것으로 기대되며, 이 설계는 선행 모의실험에서도 우수한 성능을 보였다(Casabianca et al., 2023; Guo & Wind, 2023). 그러나 대규모 평가에서는 시간과 비용의 제약으로 완전 교차 설계의 적용이 어려워, 불완전 설계(incomplete design)가 운영된다(Guo & Wind, 2023).

이러한 현실적 제약에 대응하기 위해 나선형 설계(spiral design)가 대안으로 제안되었다. Wind & Jones(2019)는 나선형, 체계적 연결, 기준 수행 등 세 가지 불완전 설계를 비교하여, 세 설계 모두에서 채점자 편향 탐지가 가능하며 설계 간 민감도 차이는 크지 않으나, 지표와 편향 유형, 편향 채점자 비율에 따라 민감도가 부분적으로 달라질 수 있음을 보였다. 이를 기반으로 Guo & Wind(2023)는 대규모 혼합형 평가 상황에서 완전 교차 설계, MC 연계 설계(Multiple Choice link design), MC+나선형 설계(Multiple Choice+spiral link design)를 비교하는 모의실험 연구를 수행하였다. 연구 결과, 완전 교차 설계가 채점자 분류 정확도와 측정 정밀도 면에서 가장 우수하고, MC+나선형 설계가 비용 대비 성능의 균형을 갖춘 현실적 대안임을 입증하였다. 현재 국가수준 학업성취도 평가에서는 서술형 답안을 두 명의 채점자에게 무작위로 배정하는 무선 할당 설계(random assignment design)가 적용되고 있다(신진아 외, 2025).

기존 선행연구에는 두 가지 중요한 공백이 존재한다. 첫째, Guo & Wind(2023)는 채점자 40명 또는 80명, 피험자 1,000명 또는 2,000명이라는 상대적으로 대규모 단일 조건만을 다루었으며, 다양한 혼합형 평가 특성을 고려한 후속 연구의 필요성을 함께 제기

하였다. 둘째, 신진아 외(2025)는 국가수준 학업성취도 평가 맥락을 반영한 모의실험 연구에서 무선 할당 설계와 MC+나선형 설계의 성능을 비교하였으나, 피험자 10,000명이 라는 단일 대규모 조건에서만 분석이 이루어졌다. 이 연구에서 두 불완전 설계 간 성능 차이는 크지 않았으나, MC+나선형 설계가 일부 지표에서 더 안정적인 결과를 보였으며, 다양한 조건에서의 추가 모의실험 연구가 필요함을 후속 과제로 제안하였다. 또한 이 연구는 채점자 간 연결성 확보 방안으로 전체 답안의 15~20%를 복수 채점하는 표본 기반 교차 채점을 제안하고, 일부 문항을 모든 채점자가 일정 비율 이상 채점하여 집단 내 비교 기준점을 마련할 것을 권고하였다. 그러나 연결성은 정성적 수준에서 다루어졌을 뿐, 채점자 쌍이 공유하는 답안량을 설계 모수로 표현한 지표는 제시되지 않았다.

연결성 확보 방식이라는 이론적 관점에서 보면, 무선 할당 설계와 MC+나선형 설계의 성능 차이는 피험자 수가 감소하거나 채점자 수가 증가하여 채점자 쌍이 공유하는 답안이 줄어들수록 커질 것으로 예상할 수 있다. 각 피험자를 두 명의 채점자에게 배정하는 상황에서 피험자 수(N)와 채점자 수(R)의 조합은 채점자당 배정량과 채점자 쌍당 공유 답안 수에 영향을 미치는데, 공유 답안이 충분할 때에는 두 설계 모두 연결 정보를 확보할 수 있으나 희소해지면 배정 방식의 구조적 차이가 드러나게 된다. 따라서 피험자 수와 채점자 수의 조합에 따라 두 설계의 편향 탐지 성능이 어떻게 달라지는지, 그리고 이러한 차이가 편향 크기에 따라 어떻게 변화하는지를 검토할 필요가 있다.

이러한 소규모 조건은 최근의 정책 환경에서 그 비중이 커지고 있다. 서울특별시교육청과 경기도교육청 등 시·도교육청 단위에서 AI 기반 서·논술형 평가지원시스템의 개발과 시범 운영이 추진되면서(김진희 외, 2025; 하민수, 최숙기, 2025), 서·논술형 평가를 시·도교육청 및 단위학교 수준으로 확대하고 있기 때문이다. 다만 현행 시스템은 교사의 채점을 지원하는 구조로 설계되어 최종 판단은 교사에게 귀속되며, 자동채점 모형 자체도 인간 채점 결과를 학습 자료이자 성능 검증의 준거로 삼는다(김진희 외, 2025; 하민수, 최숙기, 2025). 훈련 자료에 엄격성과 관대성이라는 채점자 효과가 포함되어 있으면 자동채점 점수가 그 효과를 재현할 수 있음이 보고된 바 있어(Wind et al., 2018), 인간 채점 자료의 편향 탐지는 자동채점 도입의 필요조건이다. 더욱이 AI 1차 채점과 교사 확인 채점을 결합한 혼합 운영에서는 교사 1인이 검토하는 답안 수가 줄어 채점자 간 공통 답안이 감소할 수 있으므로, 연결성이 낮은 조건에서의 배정 설계 문제는 완화되기보다 오히려 부각될 수 있다. 이러한 점에서 채점자 수와 피험자 수가 국가 수준 평가보다 작은 소규모 조건에서의 설계 간 성능 비교는 정책적으로 중요한 과제이다.

본 연구는 채점자 수와 피험자 수 조건을 단계적으로 변화시킨 모의실험을 통해, 무선 할당 설계와 MC+나선형 설계의 채점자 편향 탐지 성능을 비교 분석하고자 한다. 본 연구에서의 채점자 편향은 채점자 모형(Rater Model: RM; Muraki, 1993)의 채점자

엄격성 모수로 표현되는 편향, 곧 엄격성과 관대성을 중심으로 정의하며 이를 검증하기 위해 두 가지 연구 문제를 설정한다.

첫째, 채점자 수와 피험자 수 조건이 달라질 때, 무선 할당 설계와 MC+나선형 설계의 채점자 편향 탐지 성능은 어떻게 달라지는가?

둘째, 편향 크기에 따라 무선 할당 설계와 MC+나선형 설계의 채점자 편향 탐지 성능 차이는 어떻게 달라지는가?

II. 이론적 배경

1. 채점자 효과와 측정 모형

가. 채점자 효과가 측정에 미치는 영향

채점자 효과는 점수의 오차뿐 아니라 능력 추정과 분류 결정에도 영향을 미친다. Wind(2019)는 수행평가 맥락의 모의실험을 통해 엄격성, 중심화 경향, 비일관성의 채점자 효과가 피험자의 능력 추정치와 서열 및 평정척도 범주 분류에 영향을 미치며, 능력 추정치와 서열에 대한 영향은 엄격성과 비일관성에서 중심화 경향보다 크게 나타났음을 보고하였다. 구체적으로, 이러한 효과를 보이는 채점자를 포함한 분석과 제외한 분석의 분류 결과를 비교했을 때, 모의실험 조건에 따라 평균 5%에서 71%의 피험자가 서로 다른 평정척도 범주로 분류되었다. Hombo, Donoghue, & Thayer(2001) 역시 채점자 효과를 분석 모형에 반영하지 않으면 능력 추정치에 상당한 편향이 발생하며, 이 편향은 가장 극단적인 채점자에게 채점받은 피험자에서 가장 크게 나타남을 확인하였다.

이처럼 채점자 효과는 피험자 능력 추정의 정확도와 분류 결정의 타당성에 모두 영향을 미친다. 본 연구는 이 중 엄격성과 관대성이라는 채점자 효과에 초점을 둔다. 중심화 경향과 비일관성 역시 평가 결과를 왜곡하는 중요한 유형이나, 본 연구의 분석 모형은 채점자 모형의 엄격성 모수를 중심으로 설계되며, 이 모수를 통해 탐지되는 엄격성과 관대성이 연구의 직접적 분석 대상이다. 엄격성과 관대성은 채점자가 피험자 수행의 질과 무관하게 체계적으로 더 낮거나 높은 점수를 부여하는 경향으로, 그 탐지는 채점자들이 공통 답안을 통해 하나의 척도 위에서 비교될 수 있을 때 가능하다(Engelhard, 1997). 이러한 연결성은 채점 설계와 표본 조건에 의해 결정되므로, 채점 설계 방식의 선택은 편향 탐지의 전제 조건이 된다.

나. 채점자 모형의 구조와 모수 추정

채점자 편향의 정량적 탐지를 위해서는 피험자 능력과 문항 특성, 채점자 특성을 하나의 모형 구조 안에서 분리 추정할 수 있는 분석틀이 요구된다. Muraki(1993)의 채점자 모형은 일반화 부분 점수 모형(Generalized Partial Credit Model: GPCM; Muraki, 1992)을 확장하여 문항 변별도와 난이도를 채점자 엄격성 모수와 함께 추정하는 구조를 가진다. 이 모형은 채점자 간 편향을 단일 수준에서 추정하며, 자료의 위계 구조를 별도로 모형화하지 않아도 되므로 운영 부담이 적고 추정이 안정적이다(Muraki, 1993). 채점자 j 가 피험자 i 의 문항 k 에 대해 범주 c 로 채점할 확률은 다음과 같다.

$$P_{ikjc}(\theta_i) = \frac{\exp\left[\sum_{v=0}^c Da_k(\theta_i - b_{kv} - \lambda_j)\right]}{\sum_{m=0}^{M_k} \exp\left[\sum_{v=0}^m Da_k(\theta_i - b_{kv} - \lambda_j)\right]}$$

여기서 θ_i 는 피험자 i 의 능력 모수, a_k 는 문항 k 의 변별도, b_{kv} 는 v 번째 범주 경계 난이도(step difficulty), λ_j 는 채점자 j 의 엄격성 모수, $D=1.7$ 은 척도 변환 상수, M_k 는 문항 k 의 최대 점수 범주이다. 모형 식별을 위한 기준 범주 제약으로 $b_{k0}=0$ 으로 고정한다. $\lambda_j > 0$ 이면 해당 채점자는 상대적으로 엄격하고 $\lambda_j < 0$ 이면 관대하며, 모형 식별을 위해 $\sum \lambda_j = 0$ 으로 고정한다. 채점자 모형은 채점자 엄격성이 모든 문항에 걸쳐 일정하게 작용한다고 가정하며, GPCM 기반 구조를 공유하므로 GPCM으로 구성형 문항을 분석하는 NAEA 체계와 해석적 정합성이 높다. 한편 Hombo, Donoghue, & Thayer(2001)는 모의실험을 통해 모든 채점자가 모든 피험자를 채점하는 완전 교차 설계에서 피험자 능력 추정치의 편이가 가장 작고 표준오차도 작게 나타남을 확인하였다.

채점자 편향 탐지에서 엄격성 추정치 $\hat{\lambda}_j$ 는 정확도와 정밀도 두 측면에서 평가된다. 정확도는 참값과 추정값 간의 편의(bias)와 평균제곱오차(mean squared error: MSE)로, 정밀도는 채점자별 엄격성 추정치의 반복 간 표준편차인 경험적 표준오차(empirical standard error: EmpSE)로 평가한다. 다만 본 연구의 모형 식별 제약($\sum \lambda = 0$)에 따라 편이는 모든 조건에서 0에 수렴하므로, 정확도는 사실상 MSE를 통해 판단한다. 실제 운영 맥락에서는 채점자를 엄격 또는 관대로 분류하는 의사결정의 정확도인 채점자 분류 정확도(classification accuracy: CA)도 핵심 지표로 활용된다(Guo & Wind, 2023).

단일 문항을 복수의 채점자가 채점한 자료에 rater-as-item 방식을 적용할 때는 채점

자 j 와 문항 k 의 조합을 하나의 가상 문항으로 재구성한다. 위의 식에서 문항 k 의 범주 경계 난이도 b_{kv} 와 채점자 j 의 엄격성 λ_j 는 모두 피험자 능력 θ_i 에서 차감되므로, 가상 문항의 범주 경계는 $b_{kv} + \lambda_j$ 의 결합된 형태로 표현된다. 따라서 λ_j 가 큰 엄격한 채점자가 부여한 점수는 범주 경계가 높은 문항에 대한 응답과 같은 방식으로 작용하여, 동일한 능력을 가진 피험자가 높은 점수를 받을 확률을 낮춘다. 다만 이렇게 추정된 위치 모수에는 문항 효과가 함께 섞여 있으므로, 채점자 효과만을 분리하는 후속 절차가 요구된다. Wu(2017)는 이러한 재구성 방식이 채점자 엄격성 등의 채점자 효과를 문항반응이론(item response theory: IRT)의 문항 모수 형태로 분석하는 데 활용될 수 있음을 실제 자료 분석을 통해 확인하였다.

2. 채점 설계와 편향 탐지

가. 연결성과 채점 설계의 역할

채점자 효과를 모수로 반영하는 측정 모형에서 채점자 엄격성을 공통 척도 위에서 의미 있게 비교하려면, 데이터 내에서 채점자들이 공통 피험자 답안을 통해 직접 혹은 간접적으로 연결되어 있어야 한다. Engelhard(1997)는 이를 연결성이라 명명하며, 데이터의 절대적 크기와 무관하게 연결성이 확보되지 않으면 채점자 모수를 하나의 공통 척도 위에서 비교하기 어렵다고 지적하였다. Casabianca 외(2023)는 GPCM을 확장하여 채점자 효과를 반영한 IRT 모형인 채점자 반응 일반화 부분점수모형(Rater Response Generalized Partial Credit Model: RR-GPCM)을 대규모 평가 운영 맥락에 적용하여, 채점 설계의 연결성이 채점자 모수 회복에 미치는 영향을 모의실험으로 분석하였다. 분석 결과, 완전 교차 조건에서 모수 회복 정확도가 가장 높았으며, 불완전 교차 조건에서는 연결성이 약화될수록 모수 회복 오차가 증가하였다. 특히 채점자 정확도 모수(rater accuracy parameter)가 희소 설계에 더 민감하게 반응하였으나, 엄격성 모수는 설계 유형과 무관하게 상대적으로 강건하였다.

그러나 Casabianca 외(2023)가 보고한 엄격성 모수의 강건성은 다른 채점자 모수와 비교한 상대적인 결과로, 불완전 채점 설계에서도 엄격성 추정의 정밀도와 개별 채점자 탐지 성능이 동일하게 유지된다는 의미는 아니다. 예를 들어, Guo & Wind(2023)의 대규모 혼합형 평가 모의실험에서는 엄격성 추정치의 평균 편향이 모든 설계에서 $-0.001 \sim 0.001$ 범위로 매우 작았지만, 엄격하거나 관대한 채점자를 분류하는 정확도는 완전 교차 설계에서 $0.993 \sim 1.000$, MC+나선형 설계에서 $0.561 \sim 0.972$, MC 연계 설계에서 $0.480 \sim 0.672$ 로 설계에 따라 뚜렷하게 달랐다. 또한 채점자 1인당 배정 피험자 수

가 감소할수록 엄격성 추정의 정밀도도 낮아졌다. 이는 설계별 평균 편향이 거의 나타나지 않더라도 개별 채점자를 엄격 또는 관대로 식별하는 성능은 채점자 간 연결 구조와 채점자 1인당 배정 피험자 수에 따라 달라질 수 있음을 보여준다.

엄격성과 관대성은 피험자 점수에 일관된 상승 또는 하락을 초래하므로 채점자 훈련과 모니터링에서 우선적으로 확인된다. 평균적인 모수 회복에서 상대적으로 안정적인 엄격성은 불완전 채점 설계가 채점자 편향을 탐지하는 데 필요한 기본적인 연결성을 확보하고 있는지를 평가하기 위한 기준 효과로 활용될 수 있다. 다만 이러한 강건성은 평균 수준에서 관찰된 것으로, 개별 채점자 탐지 성능의 안정성까지 보장하지는 않는다.

불완전 설계에서 연결성을 확보하는 방법으로는 연결 집합(linkage set)이 있다. 연결 집합은 채점자 풀의 전부 또는 거의 전부가 공통으로 채점하는 답안 묶음으로, 자료의 일부를 완전 교차 구조로 만들어 채점자들을 서로 연결한다(Casabianca et al., 2023). Wind & Jones(2018)는 희소 채점 네트워크(sparse rating network) 환경에서 연결 집합의 크기와 모형-자료 적합도가 채점자 모수 추정의 안정성에 미치는 영향을 모의실험으로 분석하였다. 분석 결과, 피험자 및 문항 국면의 추정치는 연결 집합의 변화에 상대적으로 강건했으나, 채점자 국면 추정치는 연결 집합 크기가 줄어들수록 안정성이 더 민감하게 저하되는 패턴을 보였다. 채점자 모수 추정이 연결성 수준에 취약하므로, 설계 방식의 선택이 편향 탐지 민감도를 좌우한다.

나. 채점자 수 및 피험자 수 조건과 연결성

연결성의 수준은 채점 설계의 구조만이 아니라 채점자 수와 피험자 수 조건에 의해서도 결정된다. 피험자 수가 충분히 클 때는 무선 할당 설계에서도 각 채점자가 다수의 답안을 배정받으므로, 채점자 쌍들이 우연히 공통 답안을 공유할 확률이 높아진다. 이 경우 설계의 구조적 차이가 연결성에 미치는 영향은 상대적으로 작아진다. 반면 채점자 수가 많거나 피험자 수가 적은 조건에서는 채점자 1인당 배정 피험자 수가 감소하고, 채점자 쌍 간 공통 답안 확보 가능성이 낮아져 설계 간 연결성 차이가 커진다. Guo & Wind(2023)의 모의실험 결과 채점자 수가 증가할수록 MC 연계 설계와 MC+나선행 설계 모두에서 채점자 엄격성 추정치의 평균절대오차(mean absolute error)와 평균표준오차(mean standard error)가 증가하였으며, 채점자 분류 정확도도 하락하였다. 채점자 한 명당 담당 피험자 수가 줄어들수록 추정에 활용 가능한 정보량이 감소하여 엄격성 추정의 정밀도와 분류 정확도가 모두 저하된다.

본 연구와 근접한 조건을 다룬 선행연구에서도 지표에 따른 차이가 보고되었다. Wind & Jones(2019)는 채점자 수를 30명 또는 60명으로 설정하고, 피험자 수와 채점

자 수의 비율을 50:1로 고정하였다. 이들은 설계 유형뿐 아니라 편향 유형과 편향 보유 채점자 비율에 따라 각 지표의 민감도가 달라진다고 보고하였다. 따라서 설계 간 비교에서는 분류 정확도와 추정 안정성을 함께 살펴볼 필요가 있다.

다. 무선 할당 설계와 MC+나선형 설계

현재 국가수준 학업성취도 평가에서는 서술형 답안을 두 명의 채점자에게 무작위로 배정하는 무선 할당 설계를 적용하고 있다(신진아 외, 2025). 운영이 단순하고 채점 부담이 고르게 분산된다는 점이 이 설계의 장점이다. 그러나 배정 과정에서 피험자 능력이나 문항 특성에 관한 정보는 사용되지 않는다. 그 결과 채점자 간 연결성은 배정의 무선성에 좌우된다. 피험자 수가 충분하면 이러한 의존이 문제가 되지 않으나, 표본이 작아지면 일부 채점자 쌍은 공통 답안을 거의 확보하지 못한다. 이때 해당 채점자들의 엄격성을 공통 척도 위에서 비교하기 어렵다(Casabianca et al., 2023).

MC+나선형 설계는 Guo & Wind(2023)가 MC 연계 설계와 함께 비교한 불완전 설계이다. MC(Multiple Choice)는 선택형 문항을 뜻하며, 두 설계는 선택형 문항을 채점자 연결의 기제로 삼는다는 공통점이 있다. MC 연계 설계가 선택형 문항만으로 연결을 확보하는 데 비해, MC+나선형 설계는 나선형 배정을 더해 채점자 간 연결을 추가로 형성한다. 나선형 배정에서는 피험자를 일정한 기준으로 정렬한 다음 채점자 목록을 차례로 순회하며 답안을 배정하고, 목록의 끝에 이르면 처음으로 돌아가 순서를 이어 간다. MC+나선형 설계는 정렬 기준으로 선택형 문항 총점을 사용한다. 채점 이전에 확보된 능력 정보를 배정 단계에서 활용하는 것으로, 각 채점자는 이 절차를 통해 능력 분포 전반의 답안을 고르게 배정받는다. 채점자 간 간접 연결도 능력 수준 전반에 걸쳐 형성되므로, 연결이 우연에 의해 끊길 위험이 줄어든다.

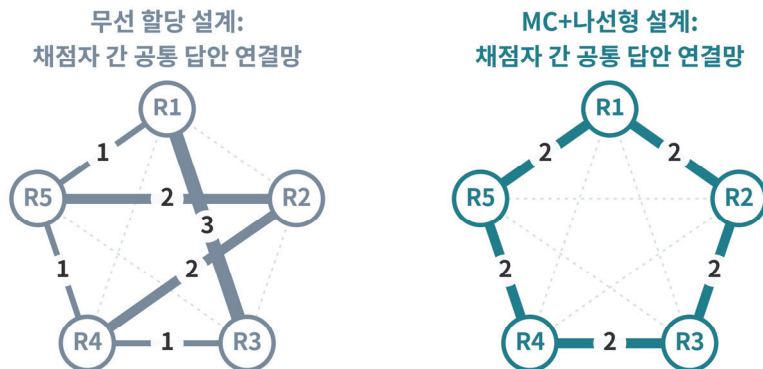
[그림 1]은 두 채점 설계가 각 피험자의 답안을 어떤 채점자에게 배정하는지 비교한 것이다. 무선 할당 설계에서는 어떤 채점자 쌍이 연결될지가 무작위로 정해진다. MC+나선형 설계는 선택형 문항 총점으로 피험자를 능력순 정렬한 후 인접한 채점자 쌍에 순환 배정한다.

무선 할당 설계: 배정행렬						MC+나선형 설계: 배정행렬					
	R1	R2	R3	R4	R5		R1	R2	R3	R4	R5
	n=4	n=4	n=4	n=4	n=4		n=4	n=4	n=4	n=4	n=4
P1 (4)	0	1	0	1	0	P1 (4)	1	1	0	0	0
P2 (6)	1	0	1	0	0	P2 (6)	0	1	1	0	0
P3 (7)	0	0	1	1	0	P3 (7)	0	0	1	1	0
P4 (8)	0	1	0	0	1	P4 (8)	0	0	0	1	1
P5 (9)	1	0	1	0	0	P5 (9)	1	0	0	0	1
P6 (11)	0	0	0	1	1	P6 (11)	1	1	0	0	0
P7 (12)	1	0	0	0	1	P7 (12)	0	1	1	0	0
P8 (13)	0	1	0	1	0	P8 (13)	0	0	1	1	0
P9 (15)	1	0	1	0	0	P9 (15)	0	0	0	1	1
P10 (17)	0	1	0	0	1	P10 (17)	1	0	0	0	1

R1~R5 = 개별 채점자, P1~P10 = 개별 피험자, P1~P10 괄호 안 = 선택형 총점
행렬의 1 = 배정, 0 = 미배정 · n = 채점자별 배정 피험자 수

[그림 1] 무선 할당 설계와 MC+나선형 설계의 채점자-피험자 배정행렬

[그림 2]는 [그림 1]의 배정행렬에서 형성되는 채점자 간 연결 구조이다. 두 채점자를 잇는 선의 숫자와 굵기는 공통으로 채점한 피험자 수를, 점선은 공통 답안이 없어 직접 연결되지 않은 쌍을 나타낸다. 무선 할당 설계에서는 10개 채점자 쌍 중 여섯 쌍만 연결되고 쌍별 공통 피험자 수가 1~3명으로 불균등하다. MC+나선형 설계에서는 인접한 다섯 쌍이 모두 2명씩 균등하게 연결된다.



선의 숫자: 공통 채점 피험자 수 · 점선: 공통 답안 없음

[그림 2] 무선 할당 설계와 MC+나선형 설계의 채점자 간 공통 답안 연결망

III. 연구 방법

1. 모의실험 설계

본 연구는 피험자 수, 채점자 수, 편향 크기, 채점 설계 방식의 네 요인을 교차한 몬테 카를로 모의실험을 통해, 소규모와 중규모 조건에서 설계, 표본, 편향 크기에 따른 성능 차이를 분석하였다. 요인별 수준과 조건 수의 산출 과정은 <표 1>과 같다. 총 45개 조건에 각각 100회 반복하여 모의실험을 수행하였다.

<표 1> 모의실험 요인과 분석 조건 수의 구성

구분	모의실험 요인	조건 수준	수준 수	조건 수 산출
1단계	피험자 수 × 채점자 수	(500, 5), (500, 10), (2,000, 5), (2,000, 10), (2,000, 30)	5	5
2단계	편향 크기	0.2, 0.5, 0.8	3	15
3단계	채점 설계	완전 교차, 무선 할당, MC+나선형	3	45

각 피험자의 구성형 문항 답안은 두 명의 채점자에게 배정되어 이중 채점된다. 이는 대규모 평가 운영에서 일반적으로 채택되는 방식이며(Casabianca et al., 2023), 채점자 간 연결성이 공통 답안을 통해서만 형성되는 조건을 반영한다. 따라서 두 불완전 설계에서 총 평정 수는 $2N$ 이고, 채점자 1인당 배정 피험자 수는 $2N/R$ 이 된다. 무선 할당 설계에서는 각 피험자에 대해 R 명 중 2명을 비복원 무선 추출하여 배정하였고, MC+나선형 설계에서는 선택형 문항 총점으로 피험자를 정렬한 뒤 나선형 순서에 따라 배정하여 두 설계의 1인당 채점 부담이 동일하도록 통제하였다. 완전 교차 설계에서는 모든 채점자가 모든 피험자를 채점하므로 1인당 배정 피험자 수는 N 이다.

피험자 능력 θ 는 $N(0, 1)$ 에서 생성하였으며, $N = 500$ (단위학교, 시·도교육청 소규모 평가)과 $N = 2,000$ (시·도 수준 표집 평가)의 두 수준을 설정하였다. 채점자 수는 $R = 5$ (시·도교육청, 단위학교 수준), $R = 10$ (중간 규모), $R = 30$ (대규모 근사값)의 세 수준으로 설정하였다.

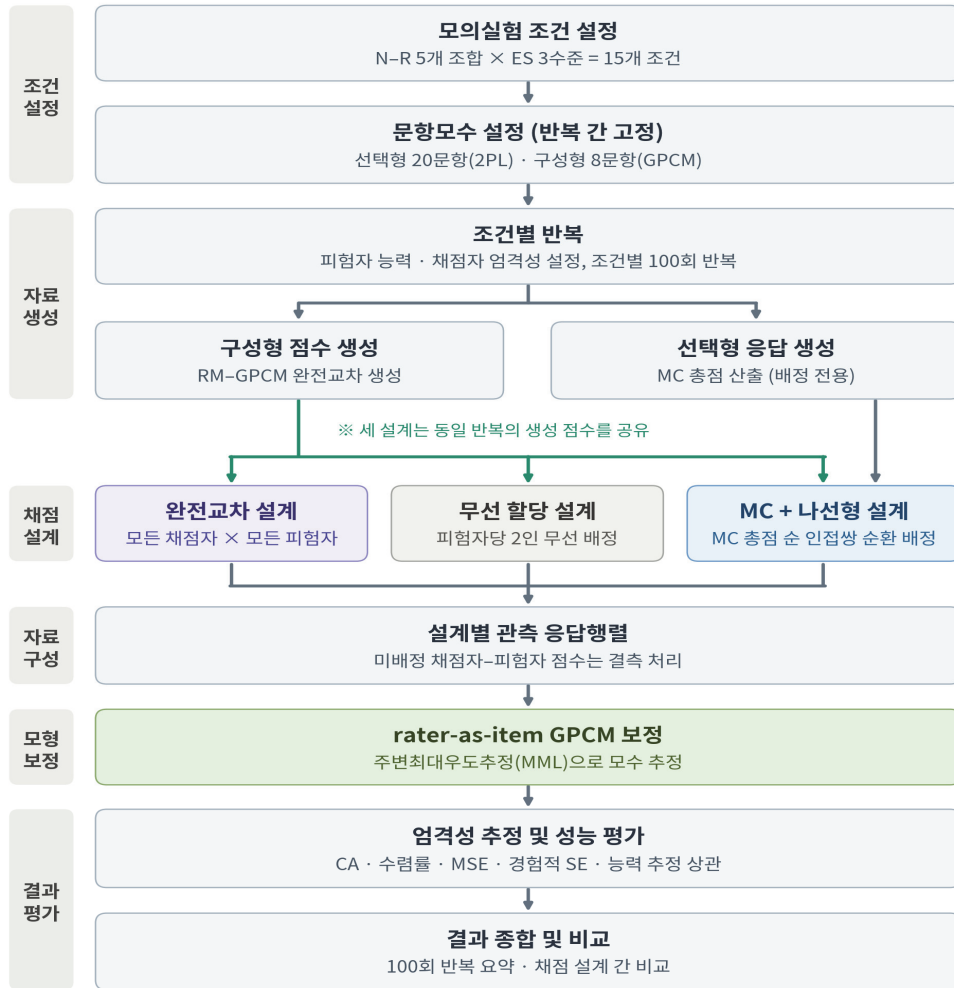
조건 간 연결성 수준을 비교 가능한 형태로 나타내기 위해 두 가지 파생 지표를 사용한다. 첫째는 $2N/R$ 로 정의되는 채점자 1인당 배정 피험자 수로, 개별 채점자의 엄격성 모수 추정에 활용되는 정보량을 나타낸다. 둘째는 $2N/\{R(R-1)\}$ 로 정의되는 채점자 쌍당 기대 공통 피험자 수로, 무선 할당 설계에서 임의의 두 채점자가 공유하는 답안 수의 기댓값이다. 피험자 수와 채점자 수는 그 절대적 크기가 아니라 이 두 지표를 통해 편향

탐지 성능에 작용한다. 특히 채점자 쌍당 기대 공통 피험자 수는 N 과 R 의 조합에 따라 달라지므로, 피험자 수가 많더라도 채점자 수가 함께 증가하면 낮아질 수 있다. 예를 들어 $N = 2,000$, $R = 30$ 조건의 쌍당 기대 공통 피험자 수는 4.6명으로, 본 연구의 조건 중 가장 적다. 따라서 조건 간 결과는 피험자 수의 크기만으로 해석하지 않고, 해당 N , R 조합과 두 파생 지표를 함께 고려하여 해석하였다.

이 두 지표를 고려하여 $N = 500$, $R = 30$ 조건은 두 가지 이유로 제외하였다. 첫째, 교과별 채점 인력 구조이다. 신진아 외(2025)가 2024년 중학교 국가수준 학업성취도 평가의 실제 조건으로 설정한 채점자 수(국어 36명, 수학 10명)를 기준으로, 피험자 10,000명과 답안당 2인 채점을 적용하면 채점자 1인당 피험자 수는 국어 약 556명, 수학 2,000명이다. 본 연구의 $N = 2,000$, $R = 5$ 조건(1인당 800명)은 이 범위 안에 포함된다. 반면 $N = 500$, $R = 30$ 조건은 1인당 33명에 그치는데, 채점자가 해당 교과 전문 인력임을 고려하면 피험자 500명 규모에서 동일 교과 채점자 30명을 확보하는 상황은 상정하기 어렵다. 둘째, 해당 조건의 채점자 쌍당 기대 공통 피험자 수가 1.15명에 불과하여, 약 3분의 1의 채점자 쌍이 공통 피험자를 확보하지 못할 것으로 예상된다. 이에 따라 유효 $N \times R$ 조합은 5개이다.

편향 채점자 비율은 원칙적으로 전체의 20%로 설정하되, 엄격 채점자와 관대 채점자를 같은 수로 두기 위해 $R = 5$ 조건에서는 2명(40%)을 편향 채점자로 지정하였다. $R = 10$ 은 엄격·관대 각 1명(총 2명, 20%), $R = 30$ 은 각 3명(총 6명, 20%)이며, 나머지 채점자의 엄격성 모수는 0으로 고정하였다. $R = 5$ 조건의 결과 해석 시 실제 편향 채점자 비율(40%)을 별도로 고려하였다. 편향 크기는 0.2, 0.5, 0.8의 세 수준으로 설정하고, 엄격성 모수는 엄격 채점자에게 $+d$, 관대 채점자에게 $-d$ 를 부여하였다.

전체 절차는 [그림 3]과 같다. 조건별로 100회 반복하여, 매 반복마다 동일한 자료에 완전 교차 설계와 무선 할당 설계, MC+나선형 설계를 각각 적용하였다. rater-as-item 방식으로 재구성한 자료에 GPCM을 적합하여 채점자 엄격성을 추정하고, 채점자 분류 정확도, 수렴률, MSE, 경험적 표준오차 및 피험자 능력 추정치와 자료 생성에 사용된 능력 모수 간 상관을 비교하였다.



[그림 3] 모의실험의 자료 생성, 채점 배정, 모형 추정 및 성능 평가 절차

2. 데이터 생성

본 연구의 혼합형 검사는 선택형 20문항과 구성형 8문항으로 구성하였다. 신진아 외 (2025)는 2024년 중학교 국가수준 학업성취도 평가의 검사 구성을 반영하여 국어 선택형 14문항, 구성형 8문항과 수학 선택형 18문항, 구성형 6문항을 설정하였다. 본 연구에서는 이러한 구성을 참고하되, MC+나선형 설계에서 피험자 정렬에 사용하는 선택형 총점의 점수 범위를 다소 넓히기 위해 선택형 문항을 20개로 설정하였다. 구성형 문항의 비율은 28.6%로, Guo & Wind(2023)가 설정한 20~40% 범위에 포함된다.

선택형 문항 반응은 2PL 모형에 따라 생성하고, 그 총점을 MC+나선형 설계에서 피

험자를 정렬하는 대리 지표로 사용하였다. 피험자를 선택형 총점순으로 정렬한 후 정해진 채점자 순서에 따라 나선형으로 배정하였으며, 선택형 문항 반응은 구성형 문항에 적용한 채점자 모형의 추정에는 투입하지 않았다. 문항 난이도는 표준정규 능력척도에서 피험자 집단에 대응하도록 $N(0, 1)$ 에서 생성하였다. 문항 변별도는 20개의 이분문항에 2PL 모형을 적용한 Hombo, Donoghue, & Thayer(2001)의 모의실험 조건을 참고하여 $U(0.8, 1.8)$ 에서 생성하였다. 반응자료 생성에 척도상수 $D=1.7$ 을 적용한 결과, 선택형 문항의 실질적인 로지스틱 기울기는 1.36~3.06으로 Hombo, Donoghue, & Thayer(2001)가 사용한 0.85~3.40의 범위에 포함된다.

구성형 문항은 0~4점의 5개 순서범주로 설정하였다. 특정 검사형의 추정치에 의존하지 않도록 문항모수는 확률분포에서 생성하였다. 문항 변별도는 Huang(2023)이 다국면 GPCM 모의실험에서 사용한 $U(0.5, 1.5)$ 와 Wu 외(2025)가 제시한 0.5~2.0을 참고하여 $U(0.8, 1.5)$ 로 설정하였다. 변별도를 균등분포에서 생성하고 선택형 문항보다 낮은 상한을 적용한 것은 8개의 구성형 문항을 포함한 혼합형 검사를 모의실험한 김성훈(2013)의 접근도 참고한 것이다. 다만 구체적인 모수 범위와 점수 범주의 수는 본 연구에서 설정하였다. 각 문항의 네 범주 경계 난이도는 $N(0, 1)$ 에서 추출한 후 오름차순으로 정렬하였다. 한 번 생성한 문항모수는 모든 실험조건과 반복에서 동일하게 사용하였다.

구성형 문항의 점수는 Muraki(1993)의 채점자 모형에 따라 생성하였다. 피험자 능력, 문항 변별도, 범주 경계 난이도 및 채점자 엄격성을 모형에 대입하여 각 점수 범주의 반응확률을 산출하고, 이를 최저 점수 범주부터 누적한 뒤 0과 1 사이에서 추출한 균일 난수가 속하는 누적확률 구간의 점수를 부여하였다. 설계별 점수 생성에 따른 추가적인 무선 차이를 줄이기 위해 완전 교차 설계에서 모든 피험자, 문항, 채점자 조합의 점수를 생성한 후, 두 불완전 설계에서는 각 설계에 따라 배정된 조합의 점수만을 추출하여 사용하였다(신진아 외, 2025).

3. 분석 모형과 추정 절차

본 연구의 분석 모형은 Muraki(1993)의 채점자 모형이다. 이 모형은 GPCM을 확장하여 문항 변별도와 범주 경계 난이도를 채점자 엄격성 모수와 함께 추정한다. 이를 추정하기 위해 다음과 같은 자료 처리와 추정 절차를 적용하였다.

분석 소프트웨어는 R 프로그램의 TAM 패키지(Robitzsch, Kiefer, & Wu, 2022)를 사용하였으며, 채점자 모형의 추정은 Wu(2017) 및 Robitzsch & Steinfeld(2018)의 rater-as-item 재구성 방식을 적용하였다. 채점자 j 가 문항 k 를 채점한 조합은 하나의 가상 문항 (j, k) 으로 처리되며, 피험자 i 가 채점자 j 에게 배정되지 않은 경우 해당 열은 결

측(NA)으로 처리된다. TAM 패키지가 추정하는 xsi 모수로부터 채점자 엄격성을 다음 절차로 분리한다.

가상 문항의 위치 모수에는 문항 효과와 채점자 효과가 함께 포함되므로, 각 가상 문항 (j, k)의 추정 변별도 $\hat{a}_{jk}$ 로 xsi_{jk} 를 나누어 위치 모수 δ_{jk} 를 산출하고, 여기서 동일 문항 k 에 걸친 채점자 평균(문항 효과 $\bar{\delta}_k$)을 제거하여 최종 채점자 엄격성 $\hat{\lambda}_j$ 를 추출한다. 즉, $\delta_{jk} = xsi_{jk} / \hat{a}_{jk}$ 이고, $\hat{\lambda}_j = mean_k(\delta_{jk} - \bar{\delta}_k)$ 이다.

채점자 모수는 TAM 패키지의 tam.mml.2pl 함수를 이용하여 GPCM에 기반한 주변 최대우도추정법(Marginal Maximum Likelihood Estimation: MML)으로 추정하였으며, 별도의 사전분포를 부과하지 않았다. 채점자별 차이는 엄격성 모수에만 반영한다는 모형 가정에 따라, 동일한 원 구성형 문항에서 파생된 채점자별 가상 문항의 변별도는 공통으로 추정하였다. 잠재 능력 분포의 적분에는 -6에서 6까지 등간격으로 배치한 구적점을 사용하였으며, 채점자 수가 5명인 조건에는 41개, 10명과 30명인 조건에는 81개를 적용하였다. 모수 변화에 대한 수렴 허용 기준은 10^{-4} , 최대 반복 횟수는 2,000회로 설정하였으며, 가속 알고리즘은 적용하지 않았다. 반복 횟수가 사전 설정된 최대 반복 횟수(2,000회)에 도달하기 전에 추정이 종료된 경우를 수렴으로 판정하였다.

4. 분석 지표

본 연구는 채점 설계의 편향 탐지 성능을 두 측면에서 평가하였다. 하나는 채점자 분류 정확도이고, 다른 하나는 채점자 엄격성 추정의 안정성이다. 채점자 분류 정확도는 채점자에 대한 모니터링과 개입이 분류 결정의 형태로 이루어진다는 점에서 편향 탐지의 실무적 결과를 반영한다. 추정 안정성은 수렴률과 MSE 분포, SE로 평가하였으며, 분류 결정의 기반이 되는 추정이 조건에 따라 얼마나 견고한지를 보여준다. 편의는 중심화 제약에 따라 모든 조건에서 0에 수렴하므로 설계 간 비교에서 제외하였다.

채점자 분류 정확도 계산은 모의실험에서 설정된 참값 λ_j 와 추정값 $\hat{\lambda}_j$ 간의 부호 방향 일치 여부를 기준으로 채점자를 분류하였다. 구체적으로, 엄격 편향 채점자($\lambda_j > 0$)는 $\hat{\lambda}_j > 0$ 일 때, 관대 편향 채점자($\lambda_j < 0$)는 $\hat{\lambda}_j < 0$ 일 때 정분류로 처리하였으며, 비편향 채점자($\lambda_j = 0$)는 $|\hat{\lambda}_j| < \text{편향 크기}/2$ 일 때 정분류로 처리하였다. 분류 기준으로 Wolfe & McVay(2012)가 제안한 로짓 절대값 기준($|\hat{\lambda}_j| > 0.5$)을 적용하는 것이 일반적이나 (Guo & Wind, 2023), 해당 기준은 채점자 풀에 대한 상대적 위치를 기반으로 정의되며 추정 척도의 구성 방식과 채점자 구조에 따라 해석이 달라질 수 있다. 이에 본 연구는 편

향 크기 조건과 생성된 참값 구조에 직접 연동된 분류 기준을 적용하였다. 이로 인해 본 연구의 채점자 분류 정확도는 선행연구의 절대 기준 기반 분류 정확도와 동일한 개념이 아니라, 각 반복에서 생성된 참값에 근거하여 편향 채점자를 식별하는 분류 정확도로 해석된다. 지표별 산출 대상 반복은 다음과 같다. 수렴률과 범주 경계모수 결손 반복률은 전체 100회 반복을 기준으로 산출하였다. 분류 정확도, MSE, 표준오차 및 피험자 능력 추정 상관은 해당 설계가 수렴한 반복만을 대상으로 산출하였으며, 설계 간 대응 차이는 두 설계가 모두 수렴한 반복만을 대상으로 계산하였다.

IV. 연구 결과

1. 수렴률과 범주 경계모수 결손

완전 교차 설계는 15개 조건에서 모두 100% 수렴하였다. <표 2>에 나타나듯 두 불완전 설계의 수렴률은 $N = 500$, $R = 5$ 와 $N = 2,000$, $R = 5$, 10 조건에서 거의 100%로 나타났다. 이 조건들의 채점자 쌍당 기대 공통 피험자 수는 각각 50.0명, 200.0명, 44.4명이었다. 반면, 수렴률은 쌍당 기대 공통 피험자 수가 각각 11.1명과 4.6명인 $N = 500$, $R = 10$ 과 $N = 2,000$, $R = 30$ 조건에서 크게 낮아졌다.

<표 2> 모의실험 조건 및 채점 설계별 수렴률과 범주 경계모수 결손 반복률

피험자수	채점자수	채점자당 피험자수	쌍당 기대 피험자수	편향 크기	수렴률(%)			범주 경계모수 결손 반복률(%)		
					완전 교차	MC+ 나선형	무선 할당	완전 교차	MC+ 나선형	무선 할당
500	5	200	50.0	0.2	100	100	100	0	2	4
				0.5	100	100	100	1	13	14
				0.8	100	98	100	9	43	50
	10	100	11.1	0.2	100	100	99	0	42	57
				0.5	100	98	96	1	68	72
				0.8	100	87	86	5	87	87
2,000	5	800	200.0	0.2	100	100	100	0	0	0
				0.5	100	100	100	0	0	0
				0.8	100	100	100	0	2	1
	10	400	44.4	0.2	100	100	100	0	0	0
				0.5	100	100	100	0	2	3
				0.8	100	100	100	0	18	14
	30	133	4.6	0.2	100	100	98	0	49	66
				0.5	100	98	93	0	81	84
				0.8	100	89	72	0	100	98

특히 $N = 2,000$, $R = 30$, 편향 크기 = 0.8에서 무선 할당 설계의 수렴률은 72%로, MC+나선형 설계의 89%보다 17%p 낮았다. 같은 조건에서 범주 경계모수 결손 반복률은 무선 할당 설계 98%, MC+나선형 설계 100%였으며, $N = 500$, $R = 10$, 편향 크기 = 0.8 조건에서는 두 설계 모두 87%로 나타났다. 범주 경계모수 결손 반복은 하나 이상의 가상 문항에서 산출된 범주 경계모수의 수가 5개 점수 범주에 대해 기대되는 4개에 미달한 반복을 의미하며, 이 반복은 완전 교차 설계에서도 $N = 500$ 의 일부 조건에서 1~9% 발생하였으나, 두 불완전 설계에 비해 전반적으로 낮았다. 이 지표는 채점자 간 연결성을 직접 측정하는 값이 아니라, 채점자와 문항별 범주 정보를 확인하기 위한 지표로 볼 수 있다.

2. 피험자 수와 채점자 수 조건에 따른 설계 간 성능

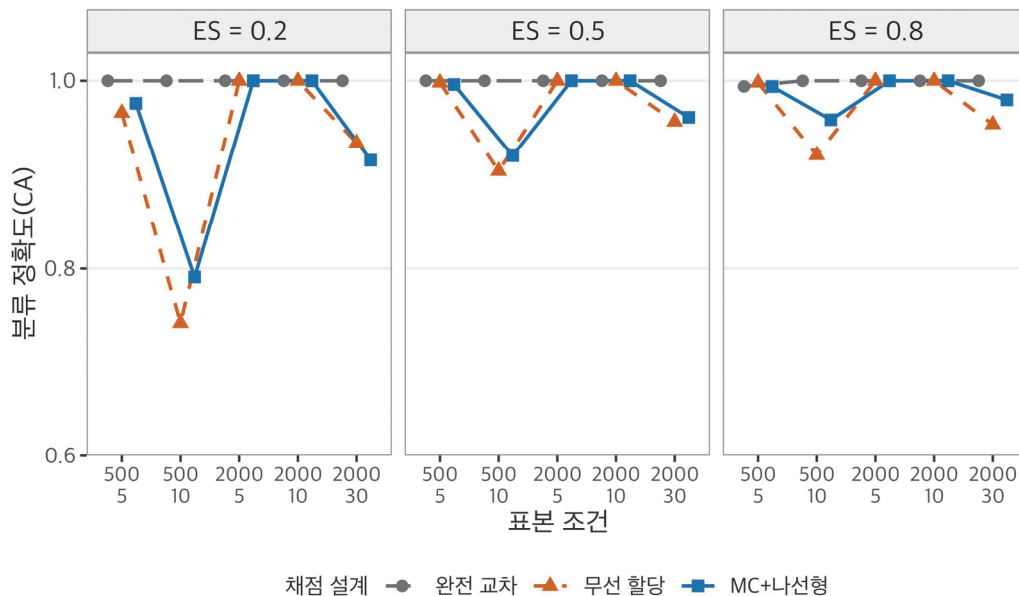
가. 분류 정확도

<표 3>은 채점 설계별 분류 정확도의 평균과 반복 간 표준편차를 조건별로 제시한다. 완전 교차 설계의 분류 정확도는 0.994~1.000으로, 충분한 연결성이 확보된 조건에서는 편향 채점자가 거의 완전하게 식별되었다. 두 불완전 설계도 $N = 2,000$, $R = 5$ 와 $N = 2,000$, $R = 10$ 에서는 모든 편향 크기에서 1.000을 기록하였다. 반면 $N = 500$, $R = 10$, 편향 크기 = 0.2에서는 MC+나선형 설계와 무선 할당 설계의 분류 정확도가 각각 0.791과 0.741로 가장 낮았으며, 반복 간 표준편차도 각각 0.232와 0.289로 크게 나타났다.

<표 3> 채점 설계별 채점자 분류 정확도

피험자수	채점자수	채점자당 피험자수	쌍당 기대 피험자수	편향 크기	완전 교차	MC+나선형 M (SD)	무선 할당 M (SD)
500	5	200	50.0	0.2	1.000	0.976 (0.107)	0.966 (0.130)
				0.5	1.000	0.996 (0.028)	0.998 (0.020)
				0.8	0.994	0.994 (0.045)	0.998 (0.020)
	10	100	11.1	0.2	1.000	0.791 (0.232)	0.741 (0.289)
				0.5	1.000	0.920 (0.101)	0.904 (0.151)
				0.8	1.000	0.959 (0.087)	0.921 (0.170)
2,000	5	800	200.0	0.2	1.000	1.000 (0.000)	1.000 (0.000)
				0.5	1.000	1.000 (0.000)	1.000 (0.000)
				0.8	1.000	1.000 (0.000)	1.000 (0.000)
	10	400	44.4	0.2	1.000	1.000 (0.000)	1.000 (0.000)
				0.5	1.000	1.000 (0.000)	1.000 (0.000)
				0.8	1.000	1.000 (0.000)	1.000 (0.000)
	30	133	4.6	0.2	1.000	0.916 (0.135)	0.934 (0.112)
				0.5	1.000	0.961 (0.082)	0.957 (0.109)
				0.8	1.000	0.980 (0.024)	0.954 (0.126)

[그림 4]는 표본 조건과 편향 크기에 따른 채점 설계별 평균 분류 정확도의 변화를 보여준다. 두 불완전 설계를 보면 $N = 2,000$, $R = 5$ 와 $R = 10$ 에서는 완전 교차 설계와 동일한 성능을 보였지만, 표본이 작거나 채점자 수가 많아 연결성이 상대적으로 낮아지는 조건에서는 정확도가 하락하였다. 이러한 하락은 $N = 500$, $R = 10$ 에서 가장 뚜렷했으며, 편향 크기가 증가할수록 완전 교차 설계와의 격차는 대체로 감소하였다. $N = 2,000$ 이더라도 채점자 수가 30명으로 증가하면 두 불완전 설계의 정확도가 다시 낮아졌다는 점에서, 편향 탐지 성능은 피험자 수뿐 아니라 채점자 수와 연결 구조의 영향을 함께 받는 것으로 나타났다. 두 불완전 설계 사이에는 일관된 우열이 나타나지 않았다. 15개 조건 중 MC+나선행 설계가 6개 조건에서, 무선 할당 설계가 3개 조건에서 더 높았으며, 나머지 6개 조건에서는 두 설계가 1.000으로 동일하였다. MC+나선행 설계의 우위가 0.02를 초과한 경우는 $N = 500$, $R = 10$ 의 편향 크기 = 0.2와 0.8, 그리고 $N = 2,000$, $R = 30$, 편향 크기 = 0.8의 세 조건에 한정되었다. 따라서 MC+나선행 설계는 일부 연결성이 낮은 조건에서 정확도 저하를 완화하였다.



[그림 4] 채점 설계, 표본 조건 및 편향 크기에 따른 분류 정확도

나. 평균제곱오차와 추정 정밀도

〈표 4〉는 수렴한 반복을 대상으로 산출한 채점 설계별 MSE 분포를 제시한다. MSE는 완전 교차 설계에서 가장 작았고, 두 불완전 설계의 차이는 어려운 조건의 분포 꼬리

에서 나타났다. $N = 500$, $R = 10$, 편향 크기 = 0.8 조건에서 MSE가 0.1을 초과한 반복은 무선 할당 설계가 수렴한 86회 중 31회(36.0%), MC+나선형 설계가 수렴한 87회 중 19회(21.8%)였다. $N = 2,000$, $R = 30$, 편향 크기 = 0.8 조건에서 MSE가 0.1을 초과한 반복은 무선 할당 설계가 수렴한 72회 중 29회(40.3%), MC+나선형 설계가 수렴한 89회 중 9회(10.1%)였다. 다만 이 결과는 수렴한 반복을 대상으로 한 조건부 비교이므로, <표 2>의 수렴률과 함께 해석할 필요가 있다.

<표 4> 채점 설계별 MSE 분포의 중앙값, 상위 분위수 및 극단값 발생 빈도

피험자 수	채점자 수	채점자당 피험자 수	쌍당 기대 피험자 수	편향 크기	설계	중앙값	P90	P95	최댓값	n (MSE> 0.1)
500	5	200	50.0	0.2	완전 교차	0.0002	0.0004	0.0005	0.0010	0
					MC+나선형	0.0006	0.0018	0.0032	0.1270	1
					무선 할당	0.0007	0.0021	0.0607	0.0930	0
				0.5	완전 교차	0.0002	0.0007	0.0009	0.0020	0
					MC+나선형	0.0009	0.0296	0.0566	0.1170	2
					무선 할당	0.0008	0.0824	0.1029	0.1910	6
				0.8	완전 교차	0.0004	0.0016	0.1469	0.6520	7
					MC+나선형	0.0021	0.1143	0.1823	0.6980	13
					무선 할당	0.0035	0.1331	0.1560	0.4460	18
	10	100	11.1	0.2	완전 교차	0.0002	0.0004	0.0005	0.0010	0
					MC+나선형	0.0180	0.0680	0.0801	0.1340	2
					무선 할당	0.0294	0.0951	0.1348	0.4980	9
				0.5	완전 교차	0.0002	0.0005	0.0006	0.0010	0
					MC+나선형	0.0347	0.0849	0.1257	0.2480	9
					무선 할당	0.0338	0.1347	0.1672	0.5320	19
				0.8	완전 교차	0.0004	0.0128	0.1829	0.2740	10
					MC+나선형	0.0395	0.2149	0.2560	0.4300	19
					무선 할당	0.0649	0.4443	0.6942	2.9970	31
2,000	5	800	200.0	0.2	완전 교차	0.0001	0.0001	0.0001	0.0002	0
					MC+나선형	0.0001	0.0003	0.0004	0.0010	0
					무선 할당	0.0001	0.0003	0.0004	0.0010	0
				0.5	완전 교차	0.0001	0.0003	0.0003	0.0010	0
					MC+나선형	0.0002	0.0004	0.0006	0.0010	0
					무선 할당	0.0002	0.0004	0.0005	0.0010	0
				0.8	완전 교차	0.0001	0.0005	0.0006	0.0010	0
					MC+나선형	0.0003	0.0008	0.0010	0.2060	1
					무선 할당	0.0002	0.0006	0.0009	0.1350	2
	10	400	44.4	0.2	완전 교차	0.0001	0.0001	0.0001	0.0001	0
					MC+나선형	0.0005	0.0010	0.0012	0.0020	0
					무선 할당	0.0004	0.0008	0.0008	0.0010	0
				0.5	완전 교차	0.0001	0.0001	0.0001	0.0002	0
					MC+나선형	0.0005	0.0010	0.0012	0.0370	0
					무선 할당	0.0004	0.0008	0.0027	0.0580	0

피험자 수	채점 자수	채점 자당 피험자 수	쌍당 기대 피험자 수	편향 크기	설계	중앙값	P90	P95	최댓값	n (MSE> 0.1)
				0.8	완전 교차	0.0001	0.0002	0.0002	0.0004	0
					MC+나선형	0.0006	0.0016	0.0388	0.0590	0
					무선 할당	0.0005	0.0304	0.0447	0.0710	0
	30	133	4.6	0.2	완전 교차	0.0001	0.0001	0.0002	0.0003	0
					MC+나선형	0.0095	0.0480	0.0623	0.3890	2
					무선 할당	0.0136	0.0338	0.0419	0.4900	2
				0.5	완전 교차	0.0003	0.0006	0.0006	0.0010	0
					MC+나선형	0.0178	0.0797	0.1184	3.5070	9
					무선 할당	0.0223	0.0808	0.1502	1.0820	8
				0.8	완전 교차	0.0006	0.0013	0.0016	0.0030	0
					MC+나선형	0.0298	0.1023	0.1800	0.6950	9
					무선 할당	0.0590	0.3023	0.7051	3.1520	29

다. 설계 간 쌍별 성능 차이

동일 반복 내에서 두 설계를 직접 비교하기 위해, 두 설계가 모두 수렴한 반복만을 대상으로 쌍별 성능 차이를 산출하였다. <표 5>의 ΔCA 는 MC+나선형 설계 값에서 무선 할당 설계 값을 뺀 것으로, 양수이면 MC+나선형 설계가 우위에 있음을 뜻한다. ΔMSE 는 두 설계의 MSE 차이로, 음수이면 MC+나선형 설계의 오차가 더 작음을 의미한다. 설계 간 우열은 조건에 강하게 의존하였다. ΔCA 는 $N = 2,000$, $R = 5$ 와 $N = 2,000$, $R = 10$ 에서 0으로 수렴하였고, $N = 500$, $R = 10$, 편향 크기 = 0.2에서 0.048로 가장 커 MC+나선형 설계가 우위를 보였다. 반면 $N = 2,000$, $R = 30$, 편향 크기 = 0.2에서는 ΔCA 가 -0.020으로 무선 할당 설계가 근소하게 앞섰다. ΔMSE 의 중앙값은 $N = 500$, $R = 10$ 과 $N = 2,000$, $R = 30$ 조건에서 음수로 MC+나선형 설계의 오차가 더 작았다. 설계 간 차이는 수렴 및 추정 안정성이 가장 낮았던 $N = 2,000$, $R = 30$, 편향 크기 = 0.8 조건에서 가장 크게 나타났으며, 중앙값은 -0.031이었다.

<표 5> 동일 반복 내 두 불완전 설계의 성능 차이(ΔCA , ΔMSE)

피험자 수	채점 자수	채점 자당 피험자 수	쌍당 기대 피험자 수	편향 크기	동시 수렴	채점자 분류 정확도		ΔCA		ΔMSE	
						MC+ 나선형	무선 할당	평균	SD	평균	중앙값
500	5	200	50.0	0.2	100/100	0.976	0.966	0.010	0.154	-0.0018	0.0000
				0.5	100/100	0.996	0.998	-0.002	0.035	-0.0078	0.0000
				0.8	98/100	0.994	0.998	-0.004	0.050	0.0022	-0.0005
	10	100	11.1	0.2	99/100	0.790	0.741	0.048	0.370	-0.0150	-0.0003
				0.5	94/100	0.920	0.903	0.017	0.184	-0.0160	-0.0019
				0.8	75/100	0.955	0.931	0.024	0.175	-0.1001	-0.0203

시험 자수	채점 자수	채점 자당 시험 자수	쌍당 기대 시험 자수	편향 크기	동시 수렴	채점자 분류 정확도		ΔCA		ΔMSE	
						MC+ 나선형	무선 할당	평균	SD	평균	중앙값
2000	5	800	2000	0.2	100/100	1.000	1.000	0.000	0.000	0.0000	0.0000
				0.5	100/100	1.000	1.000	0.000	0.000	0.0000	0.0000
				0.8	100/100	1.000	1.000	0.000	0.000	0.0003	0.0001
	10	400	44.4	0.2	100/100	1.000	1.000	0.000	0.000	0.0001	0.0001
				0.5	100/100	1.000	1.000	0.000	0.000	-0.0017	0.0001
				0.8	100/100	1.000	1.000	0.000	0.000	-0.0016	0.0001
	30	133	4.6	0.2	98/100	0.914	0.934	-0.020	0.174	0.0006	-0.0001
				0.5	91/100	0.970	0.957	0.013	0.115	-0.0206	-0.0046
				0.8	66/100	0.981	0.956	0.025	0.131	-0.1312	-0.0306

동시 수렴 = 두 설계가 모두 수렴한 반복 수 / 전체 반복 수. 표의 통계량은 동시 수렴 반복 기준으로 산출.

3. 편향 크기에 따른 설계 간 성능 차이

가. 편향 크기 수준별 분류 정확도

〈표 5〉의 ΔCA 를 편향 크기 수준별로 보면, 편향 크기가 작을수록(편향 크기 = 0.2) 소규모 조건에서 MC+나선형 설계의 분류 정확도 우위가 가장 컸다. 미세 편향의 탐지는 채점자 간 연결성의 질에 민감할 수 있으며, 이 조건에서 나타난 MC+나선형 설계의 상대적 우위는 능력 수준에 따른 체계적 답안 배정과 관련된 것으로 해석할 수 있다. 최소 탐지 가능 편향의 관점에서, 〈표 3〉에 따르면 $N = 2,000$, $R = 5$ 와 $R = 10$ 에서는 편향 크기 = 0.2에서도 분류 정확도가 1.000에 이른 반면, $N = 500$, $R = 10$ 에서는 편향 크기 = 0.2의 분류 정확도가 0.74~0.79에 그쳐 미세 편향 탐지가 상대적으로 어려운 것으로 나타났다.

나. 조건 상호작용과 최소 탐지 가능 편향 크기

〈표 5〉의 ΔCA 를 조건별로 종합하면, 설계 간 격차는 피험자 수가 작고 채점자 수가 많으며 편향이 작은 조건($N = 500$, $R = 10$, 편향 크기 = 0.2)에서 0.048로 가장 컸고, 연결성 밀도가 충분한 $N = 2,000$, $R = 5$ 와 $N = 2,000$, $R = 10$ 조건에서는 모든 편향 크기 수준에서 0으로 수렴하였다. 설계 간 격차가 피험자 수가 적고 채점자 수가 많은 조건에 집중된 것은 채점자 1인당 배정 피험자 수와 채점자 쌍당 기대 공통 피험자 수가 함께 작용한 결과로 볼 수 있다. 무선 할당 설계에서 연결성이 우연히 약화될 가능성이 이들 조건에서 성능 차이로 나타났다.

4. 경험적 표준오차와 피험자 능력 추정 상관

〈표 6〉은 세 채점 설계의 표준오차와 피험자 능력 추정 상관을 정리한 것이다. 채점자 엄격성 추정의 편의는 앞서 밝힌 대로 모든 조건에서 0에 수렴하여 설계 간 차이가 없으므로 별도로 제시하지 않았다. 정밀도 지표인 평균 EmpSE는 채점자별 $\hat{\lambda}_j$ 의 반복 간 표준편차를 채점자에 걸쳐 평균한 값이다. 설계 간 격차는 $N = 500$, $R = 10$ 과 $N = 2,000$, $R = 30$ 의 편향 크기 0.8 조건에서 가장 컸다. 〈표 6〉에 나타나듯 $N = 2,000$, $R = 30$, 편향 크기 = 0.8 조건에서 SE는 MC+나선행 설계가 0.178인 데 비해 무선 할당 설계는 0.284로 더 컸다. 반면 추정된 피험자 능력과 참 능력의 평균 상관은 불완전 설계에서 약 0.98, 완전 교차 설계에서 0.99 이상으로 설계 간 차이가 미미하였다. 즉 채점 설계의 선택은 피험자 능력 추정보다 채점자 편향 탐지의 정밀도에 더 크게 작용한다.

〈표 6〉 채점 설계별 경험적 표준오차와 피험자 능력 추정치의 상관

피험자수	채점자수	채점자당 피험자수	쌍당 기대 피험자수	편향 크기	경험적 표준오차 평균(EmpSE)			상관		
					완전 교차	MC+ 나선행	무선 할당	완전 교차	MC+ 나선행	무선 할당
500	5	200	50.0	0.2	0.0147	0.0588	0.0684	0.992	0.980	0.980
				0.5	0.0171	0.0789	0.0983	0.992	0.980	0.980
				0.8	0.1171	0.1541	0.1487	0.992	0.980	0.979
	10	100	11.1	0.2	0.0152	0.1601	0.1976	0.996	0.980	0.979
				0.5	0.0164	0.1908	0.2069	0.996	0.979	0.980
				0.8	0.0824	0.2055	0.3141	0.996	0.979	0.979
2,000	5	800	200.0	0.2	0.0074	0.0134	0.0129	0.992	0.980	0.980
				0.5	0.0086	0.0143	0.0139	0.992	0.980	0.980
				0.8	0.0105	0.0471	0.0441	0.992	0.980	0.980
	10	400	44.4	0.2	0.0072	0.0239	0.0212	0.996	0.980	0.980
				0.5	0.0080	0.0282	0.0373	0.996	0.980	0.980
				0.8	0.0090	0.0462	0.0491	0.996	0.980	0.980
	30	133	4.6	0.2	0.0080	0.1286	0.1280	0.999	0.980	0.980
				0.5	0.0094	0.2051	0.1867	0.999	0.979	0.980
				0.8	0.0109	0.1778	0.2835	0.999	0.979	0.980

V. 결론 및 논의

본 연구는 표본 조건과 편향 크기를 달리한 모의실험으로 두 불완전 채점 설계의 편향 탐지 성능을 비교하였다. 두 연구 문제에 대한 답은 다음과 같다.

첫째, 두 설계의 성능 차이는 피험자 수와 채점자 수 조건에 의존하였다. 피험자 수가 충분하고 채점자당 배정 답안이 많은 조건($N = 2,000$, $R = 5, 10$)에서는 두 설계가 분류 정확도와 수렴률에서 사실상 동등하였다. 그러나 채점자 간 연결 정보가 상대적으로 희소한 $N = 500$, $R = 10$ 과 $N = 2,000$, $R = 30$ 의 편향 크기 0.8 조건에서는 MC+나선형 설계가 더 높은 수렴률과 낮은 MSE 및 EmpSE를 보였다.

이러한 차이는 연결성 확보 방식과 관련된다. 무선 할당 설계는 공통 답안 형성을 무선성에 의존하므로, 피험자 수가 제한적이거나 채점자 수가 많아질수록 채점자 쌍당 기대 공통 피험자 수가 감소한다. 이 경우 채점자 엄격성 모수 λ_j 의 추정이 불안정해질 수 있으며, 극단적인 조건에서는 수렴률이 낮아질 수 있다. 실제로 $N = 2,000$, $R = 30$, 편향 크기 = 0.8 조건에서 무선 할당 설계의 수렴률은 72%로 MC+나선형 설계의 89%보다 낮았는데, 이는 채점자 쌍당 공유 정보가 적은 조건에서 채점자 모수 추정이 상대적으로 불안정할 수 있음을 보여준다. 반면 MC+나선형 설계는 피험자를 선택형 점수에 따라 정렬하여 배정하므로, 채점자 쌍이 점수 분포 전반에서 공통 답안을 공유한다. 채점자 모수 추정이 연결 정보의 감소에 상대적으로 민감하다는 점은 선행 연구(Wind & Jones, 2018; Casabianca et al., 2023)와 부합하며, 연결 정보가 제한된 일부 조건에서 나타난 설계 간 차이는 Guo & Wind(2023)가 보고한 결과와 같은 방향이다.

이러한 조건 효과는 두 파생 지표로 정리된다. 첫 번째 지표인 채점자 1인당 배정 피험자 수를 기준으로 정렬하면, 모든 편향 크기 수준에서 분류 정확도는 지속적으로 높아지고 MSE와 표준오차는 대체로 낮아졌다. 완전 교차 설계에서는 채점자마다 모든 피험자를 채점하므로, $N=2,000$ 에서 R 이 5, 10, 30으로 달라져도 EmpSE는 0.0072~0.0109의 비교적 작은 범위에 머물렀다. 이는 완전 교차 설계에서 채점자 1인당 배정 피험자 수가 채점자 수와 무관하게 N 으로 고정되기 때문이다. <표 6>의 무선 할당 설계에서 $N = 2,000$, 편향 크기 = 0.5 조건을 보면, 채점자 1인당 배정 피험자 수가 800, 400, 133 명으로 줄어들어 따라 표준오차는 0.0139, 0.0373, 0.1867로 커졌다. 채점자 수가 많아지면서 1인당 배정량과 쌍당 공유 정보가 함께 감소하였고, 추정 정밀도도 낮아지는 경향이 나타났다.

두 번째 지표인 채점자 쌍당 기대 공통 피험자 수를 고려하면, 수렴률은 채점자 1인당 배정 피험자 수가 증가하더라도 일관되게 높아지지 않는 않았다. <표 2>에서 $N =$

2,000, $R = 30$ 조건은 채점자 1인당 배정 피험자 수가 133명으로 $N = 500$, $R = 10$ 조건의 100명보다 많았다. 그럼에도 무선 할당 설계의 편향 크기 = 0.8에서 $N = 2,000$, $R = 30$ 조건의 수렴률은 72%로 $N = 500$, $R = 10$ 조건의 86%보다 낮았다. 두 조건의 쌍당 기대 공통 피험자 수는 4.6명과 11.1명으로 배정량의 순서와 반대였다. 이는 채점자 1인당 배정량뿐 아니라 채점자 쌍당 공유 정보도 모형 수렴과 관련될 수 있음을 보여준다.

둘째, 편향 크기에 따른 두 설계의 차이는 추정 안정성에서 상대적으로 뚜렷하게 나타났다. 분류 정확도는 대부분의 조건에서 두 설계가 사실상 동등하였으며, <표 3>과 <표 5>에서 보듯 두 설계의 차이는 반복 간 표준편차 안에 머물렀다. 예컨대 <표 5>에서 $N = 500$, $R = 10$, 편향 크기 = 0.2 조건의 분류 정확도 평균 차이는 0.048이었으나, 같은 조건의 반복 간 표준편차는 0.370으로 그 차이를 크게 상회하였다. <표 3>에서 보듯 일부 조건($N = 2,000$, $R = 30$, 편향 크기 = 0.2)에서는 무선 할당 설계의 분류 정확도가 오히려 더 높게 나타나, 이 지표만으로는 두 설계의 우열을 일관되게 규정하기 어려웠다. 이는 채점자 위치 추정치가 불완전 채점 설계의 종류와 무관하게 엄격성과 관련성을 탐지하는 데 유용하다고 보고한 Wind & Jones(2019)의 결과와 부합한다. <표 2>의 $N = 2,000$, $R = 30$ 조건에서 무선 할당 설계의 수렴률은 편향 크기 0.2의 98%에서 0.8의 72%로 떨어진 반면, MC+나선행 설계는 100%에서 89%로 감소하는 데 그쳤다. <표 4>의 MSE 극단값 발생 빈도와 <표 6>의 무선 할당 설계 표준오차에서도 같은 방향의 증가가 나타났다. MC+나선행 설계도 일부 조건에서 추정 불안정성이 나타났다. 연결 정보가 가장 희소한 $N=2,000$, $R=30$ 의 편향 크기 0.8 조건에서는 무선 할당 설계보다 수렴률이 높고 MSE와 EmpSE가 낮아 불안정성의 증가를 상대적으로 완화하였다. 이는 구성형 문항 응답의 무선 배정을 포함한 MC 연계 설계가 채점자 추정 정밀도에서 가장 취약한 설계로 보고된 Guo & Wind(2023)의 결과와도 같은 방향이다. 채점자 쌍당 공유 정보가 많은 조건에서 두 설계의 분류 정확도가 비슷했다고 해서, 연결 정보가 희소한 조건에서도 그러하리라고 보기는 어렵다.

이러한 결과는 단일 대규모 조건($N = 10,000$)에서 두 불완전 설계 간 성능 차이가 크지 않다고 보고한 선행 연구(신진아 외, 2025)와 모순되지 않는다. 오히려 본 연구는 그 차이가 작게 나타난 까닭이 연결성 밀도가 높은 대규모 조건이라는 전제에 있었음을 보이고, 피험자 수가 줄거나 채점자 수가 늘어 연결성 밀도가 낮아질수록 두 설계의 추정 안정성 차이가 드러남을 확인하였다. 설계 간 차이가 작다는 선행 연구의 관찰은 표본 조건을 함께 고려하여 해석할 필요가 있다.

주목할 점은 설계 선택의 효과가 추정 대상에 따라 비대칭적으로 나타났다는 것이다. <표 6>에서처럼 피험자 능력 추정에서는 추정값과 참값의 상관이 불완전 설계에서 약

0.98, 완전 교차 설계에서 0.99 이상으로 설계 간 차이가 거의 없었던 반면, 채점자 편향 탐지의 정확도는 설계에 민감하게 반응하였다. 이는 피험자 능력 θ_i 가 여러 문항과 복수 채점자의 평정 정보를 통합해 추정되기 때문으로 해석할 수 있다. 반면 한 채점자가 전반적으로 낮은 점수를 부여하는 경우, 공동 채점 정보가 없으면 그 점수 수준이 채점자의 엄격성 때문인지 그에게 배정된 피험자의 능력이 낮았기 때문인지 구분하기 어렵다(Wu, 2017). 채점자들이 일부 답안을 공동으로 채점하거나 이러한 공동 채점 관계를 통해 간접적으로 연결될 때, 비교 가능한 피험자 수행을 기준으로 채점자 효과를 피험자 능력 차이와 분리하고 채점자 간 엄격성 λ_j 를 공통 척도에서 비교할 수 있다(Engelhard, 1997). 따라서 채점자 엄격성 추정은 피험자 능력 추정보다 배정 구조가 형성하는 공동 채점 관계에 더 민감하다.

피험자 점수만 산출하는 것이 목적이라면 설계에 따른 차이가 적지만, 채점자 모니터링과 개입을 목적에 포함하면 설계의 선택이 결과에 영향을 준다. 따라서 시·도교육청 및 단위학교 수준의 평가와 같이 운영 여건에 따라 채점자 1인당 배정량과 채점자 쌍당 공유 정보가 제한될 수 있는 상황에서 채점자 편향을 안정적으로 탐지할 필요가 있다면, MC+나선형 설계를 우선적으로 고려할 수 있다. 반면 피험자 수가 충분한 대규모 표집 평가에서는 현행 무선 할당 설계로도 편향 탐지에 큰 손실이 없다. 이는 평가 규모와 운영 목적에 따라 채점 설계를 차등 적용하는 전략의 근거를 제공한다.

본 연구는 몇 가지 한계를 가진다. 첫째, 채점자 모형의 엄격성과 관대성에 초점을 두었으며 중심화 경향이나 비일관성 등 다른 편향 유형은 다루지 않았다. 둘째, 문항 모수를 반복 간 고정하고 100회 반복으로 추정하였다. 셋째, $R = 5$ 조건에서는 편향 채점자 수를 정수로 배정하는 과정에서 실제 비율이 명목 20%가 아닌 40%로 설정되었으므로, 해당 조건의 결과 해석에는 주의가 필요하다. 넷째, 본 연구의 채점자 분류 정확도는 절대적 로짓 기준(Wolfe & McVay, 2012)이 아니라 각 반복에서 생성된 참값에 근거한 상대적 분류로 정의되었다. 이 방식은 채점자의 참 상태에 따라 정분류와 오분류의 기준이 달라지는 구조이므로, 추정치에 단일 기준을 적용해 분류하는 표준적 분류 정확도보다 다소 관대하게 산출되는 경향이 있다. 따라서 본 연구의 분류 정확도는 선행 연구의 절대 기준 분류 정확도와 직접 비교되지 않으며, 그 절대 수준보다는 조건 간 상대적 변화로 해석되어야 한다. 다섯째, 본 연구는 문항 모수를 특정 검사의 실제 추정치가 아니라 통계적 분포에서 생성하였다. 이는 일반화 가능성을 위한 선택이나, 그 결과 국가수준 학업성취도 평가나 시·도교육청, 단위학교 수준 평가에서 실제로 사용되는 문항의 특성, 예컨대 국가수준 평가에서 교과에 따라 달라지는 배점 분포를 그대로 반영하지는 않는다. 따라서 실제 사·논술형 자료에 적용할 경우 결과가 달라질 수 있다.

특히 수렴에 실패한 반복을 채점자 분류 정확도 계산에서 제외한 처리는 결과 해석에서 유의할 부분이다. 수렴에 실패한 반복은 추정이 가장 어려웠던 사례일 가능성이 높으므로, 이를 제외한 분류 정확도는 수렴률이 더 낮았던 설계, 즉 무선 할당 설계에 다소 유리하게 작용했을 수 있다. 그렇다면 본 연구에서 관찰된 MC+나선형 설계의 우위는 보수적으로 추정된 것으로 볼 수 있다. 다만 이 해석은 탈락한 반복의 분류 정확도가 더 낮았으리라는 가정에 의존하며, 두 설계의 수렴률 차이가 큰 조건에 한정되므로, 수렴 실패를 포함한 분석으로 확인할 필요가 있다. 향후 연구에서는 다양한 편향 유형의 동시 발생, 피험자 수와 채점자 수의 더 넓은 조합, 실제 서·논술형 자료에의 적용을 통해 본 결과의 일반화 가능성을 검토할 필요가 있다.

참고문헌

- 강민석(2024). 채점자 간 협의가 채점자의 채점 특성에 미치는 영향: 예비 한국어 교원을 대상으로. **이중언어학**, 98, 1-26.
- 교육부(2022). **초·중등학교 교육과정 총론(교육부 고시 제2022-33호 [별책 1])**. 교육부.
- 교육부(2023). **미래 사회를 대비하는 2028 대학입시제도 개편 확정안**. 교육부.
- 김성훈(2013). 혼합형 검사에서 IRT 문항모수와 기저 능력분포의 동시 추정이 문항모수 추정의 정확도에 미치는 영향. **교육평가연구**, 26(3), 647-668.
- 김진희, 김위정, 김호영, 김준엽, 박종임, 최숙기(2025). 「**하이러닝AI 서·논술형 평가시스템**」 시범운영 실태 분석 및 현장 도입 지원 방안(정책연구 2025-18). 경기도교육연구원.
- 신진아, 박인용, 김미림, 곽민호, 박민호, 강혜진, 구슬기, 정호현(2025). **컴퓨터 기반 학업 성취도 평가의 채점 체제 개선 방안(RRE 2025-3)**. 한국교육과정평가원.
- 심재호, 노은희, 김명화(2012). 국내외 학업성취도 평가의 서답형 문항 특성과 채점 체제 고찰. **열린교육연구**, 20(4), 159-185.
- 하민수, 최숙기(2025). **중등 학교 서·논술형 평가 지원을 위한 AI 기반 평가 지원(채점-피드백) 도구 개발 타당성 검토 연구**. 서울특별시교육청.
- Barnett, S. G. (2005). Relative performance of scoring designs for the assessment of constructed responses. Doctoral dissertation, University of British Columbia.
- Casabianca, J. M., Donoghue, J. R., Shin, H. J., Chao, S. F., & Choi, I. (2023). Using linkage sets to improve connectedness in rater response model estimation. *Journal of Educational Measurement*, 60(3), 428-454.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.
- Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.
- Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, 35(2), 137-154.

- Guo, W., & Wind, S. A. (2023). The effects of rating designs on rater classification accuracy and rater measurement precision in large-scale mixed-format assessments. *Applied Psychological Measurement, 47*(2), 91-105.
- Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (Research Report RR-01-05). Educational Testing Service.
- Huang, H.-Y. (2023). Modeling rating order effects under item response theory models for rater-mediated assessments. *Applied Psychological Measurement, 47*(4), 312-327.
- Lunz, M. E., & Stahl, J. A. (1993). The effect of rater severity on person ability measure: A Rasch model analysis. *American Journal of Occupational Therapy, 47*(4), 311-317.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.
- Muraki, E. (1993, April). *Variations of polytomous item response models: Raters' effect model, DIF model, and trend model*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.
- Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules* (R package version 4.1-4) [Computer software]. <https://CRAN.R-project.org/package=TAM>
- Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101-139.
- Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement, 43*(2), 159-171.
- Wind, S. A., & Ge, Y. (2024). Detecting rater bias in mixed-format assessments. *Measurement: Interdisciplinary Research and Perspectives, 22*(1), 20-30.
- Wind, S. A., & Jones, E. (2018). The stabilizing influences of linking set size and model-data fit in sparse rater-mediated assessment networks. *Educational and Psychological Measurement, 78*(4), 679-707.

- Wind, S. A., & Jones, E. (2019). The effects of incomplete rating designs in combination with rater effects. *Journal of Educational Measurement*, 56(1), 76–100.
- Wind, S. A., Wolfe, E. W., Engelhard, G., Jr., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, 18(1), 27–49.
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31–37.
- Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, 59(4), 453–470.
- Wu, T., Kim, S. Y., Westine, C., & Boyer, M. (2025). IRT observed–score equating for rater–mediated assessments using a hierarchical rater model. *Journal of Educational Measurement*, 62(1), 145–171.

논문접수 : 2026.7.2. / 수정본접수 : 2026.7.30. / 게재승인 : 2026.8.11.

ABSTRACT

A Comparison of Random Assignment and MC + Spiral Link Designs for Detecting Rater Bias

Soonbo Kwon

Part-time lecturer, Korea National University of Education

This simulation study compared random assignment and MC + spiral link designs for detecting rater bias across different numbers of examinees and raters and three bias magnitudes. Responses were generated using a GPCM-based rater model. Five examinee-by-rater combinations crossed with three bias magnitudes yielded 15 base conditions. Applying complete-crossed, random-assignment, and MC + spiral link designs produced 45 design conditions, each replicated 100 times. Outcomes included rater classification accuracy, convergence rate, mean squared error, and empirical standard error. The performance of the two incomplete designs varied by sample condition. When examinees and assignments per rater were sufficient, the designs performed similarly. When assignments per rater were limited and rater pairs shared few examinees under random assignment, the MC + spiral link design showed higher convergence and generally smaller errors. In some conditions, differences between the designs were more apparent in estimation stability than in classification accuracy. These findings support considering the MC + spiral link design for mixed-format assessments with limited rater-pair linking information.

Key Words: rater bias, rating design, MC + spiral link design, random assignment design, IRT