2026-1 DSL EDA Project

같은 점수, 다른 과정:
과정 기반 문항·학습자 단위 분석

EDA 교육팀

14기 김서윤 윤동영
15기 박성하 배소윤 이세원 이지원

목차

왜 '정답률'만으로는 충분하지 않은가?

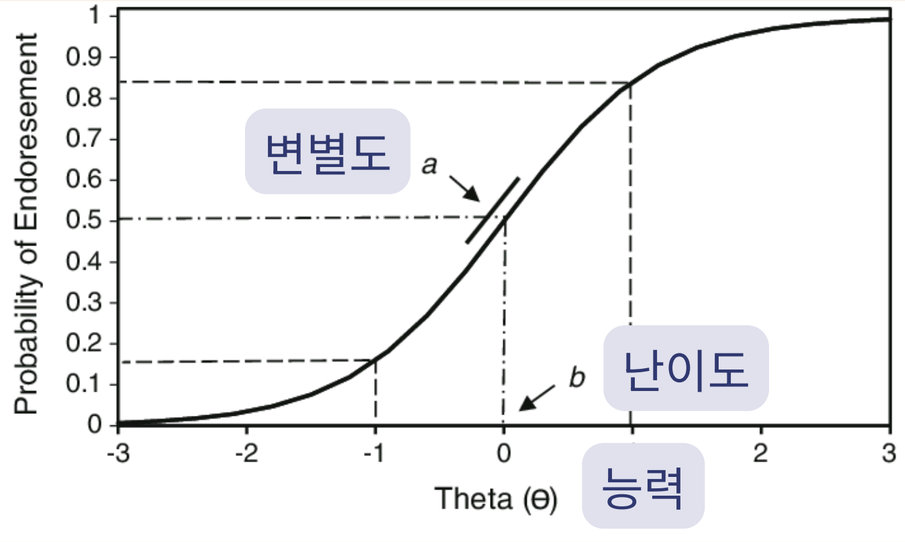
IRT(Item Response Theory)란?

피험자의 전체 검사 점수가 아니라 각 문항에 대한 응답(정/오답)을 분석하여
피험자의 능력(θ)과 문항 특성(난이도, 변별도 등)을 개별적으로 추정하는 통계학적 검사이론

전통적인 문항 분석(CTT)는 총점을 중심으로 학습 성취를 평가
한단계 더 발전한 문항반응이론(IRT)에서는 학습자 능력과 문항 난이도를 분리하여 측정

기본 가정

1) 지역독립성(local independence)이 확보됨
2) 문항특성곡선(ICC; Item Characteristic Curve)이
   특정한 형태를 가짐



왜 '정답률'만으로는 충분하지 않은가?

같은 점수, 다른 과정
그러나 동일한 점수, 동일한 능력(θ)을 가진 학습자라도
풀이 시간, 시도 횟수, 반응 패턴은 서로 다를 수 있음
→ 즉, 결과 변수만으로는 **학습 과정의 차이**를 설명하기 어려움

학습 로그 데이터에는 Response Time, Attempt 등 많은 과정 정보를 담은 지표가 존재

우리 연구는 이를 바탕으로 분석 시작!

왜 '정답률'만으로는 충분하지 않은가?

연구문제 1:
문항난이도로 설명되지 않는 과정난이도를 학습 로그 기반 '혼란도'로 정량화할 수 있는가?
→ 새롭게 정의한 과정 기반 지표 '혼란도'와 IRT 기반 지표 '난이도'로 문항 유형을 클러스터링,
유형별 패턴 분석 후 시스템 개선 방안 분석

연구문제 2:
학습 유형을 학습자의 능력과 더불어 학습과정 변수로 군집화할 수 있는가?
→ 학습자의 능력(θ)과 학습 로그 데이터를 가공하여 만든 User Feature 기반으로 학습 유형 클러스터링,
학습 유형별 특징 분석 후 학습 방법 피드백 방안 분석

=> 최종적으로는 연구문제 1과 2를 연결해 과정 기반 학습의 중요성을 담은 인사이트 도출

데이터 소개 (기본 정보)

EdNet(TOEIC 학습 플랫폼 로그)

- “EdNet is the dataset of all student-system interactions collected over 2 years by Santa, a multi-platform AI tutoring service with more than 780K users in Korea available through Android, iOS and web.” (Riiid, 2020)
- 정답 여부, 학습 로그를 함께 제공
- 과정 기반 학습 행동 분석이 가능하다는 점에서 본 연구에 적합

데이터 소개 (사용 로그 구성)

EdNet(TOEIC 학습 플랫폼 로그)

Dataset

- KT1 (Submission log): 문항 제출 결과(응답, 정오답 판단, 풀이시간)
- KT2 (Process log): 풀이 과정 행동(respond/submit, bundle enter/quit 등)
- KT4 (Full behavior log): 행동 로그(예: pay 포함, 추가 행동 이벤트)
- Contents: 문항 메타 + 정답키(correct_answer, part, tags, bundle_id)

분석 설계 포인트

- Q1은 문항 단위(item-level), Q2는 학생 단위(user-level) 분석
- 두 분석 모두 동일 표본(공통 코호트)에서 출발

데이터 전처리 ①: 학습자(코호트) 선발

목적: Q1/Q2에 필요한 로그(KT1, KT2)가 모두 존재하는 “유료 사용자” 코호트 확정

- **유료 사용자 정의 (KT4)**
  - KT4 로그에서 action_type = pay가 1회 이상이면 **유료 사용자**로 분류 (23,789명)
- **분석 가능 코호트 확정 (교집합)**
  - KT1(정답률·풀이시간) + KT2(응답변경) 지표 생성을 위해 **KT4_pay ∩ KT1 ∩ KT2**만 유지
- **최종 분석 대상 확정**
  - 297,915명 중 최종 코호트 23,477명

=> 23,477명 분석에 활용

데이터 전처리 ②: 문항 선별

목적: 선발된 학습자 코호트(23,477명)의 KT1 제출 로그를 기준으로 IRT 분석에 사용할 “문항 집합”을 정리함

IRT 입력 문항 구성
전체 13,169개 → 11,574개 문항 사용

기준:

- KT1(선발된 코호트)의 제출 기록에서 등장한 모든 문항을 수집
- question_id 기준으로 중복을 제거하여 문항 구성(각 문항은 1회만 포함)

=> 11,574개 문항 분석에 활용

연구문제 1

문항난이도로 설명되지 않는 **과정난이도**를
학습 로그 기반 ‘혼란도’로 정량화할 수 있는가?

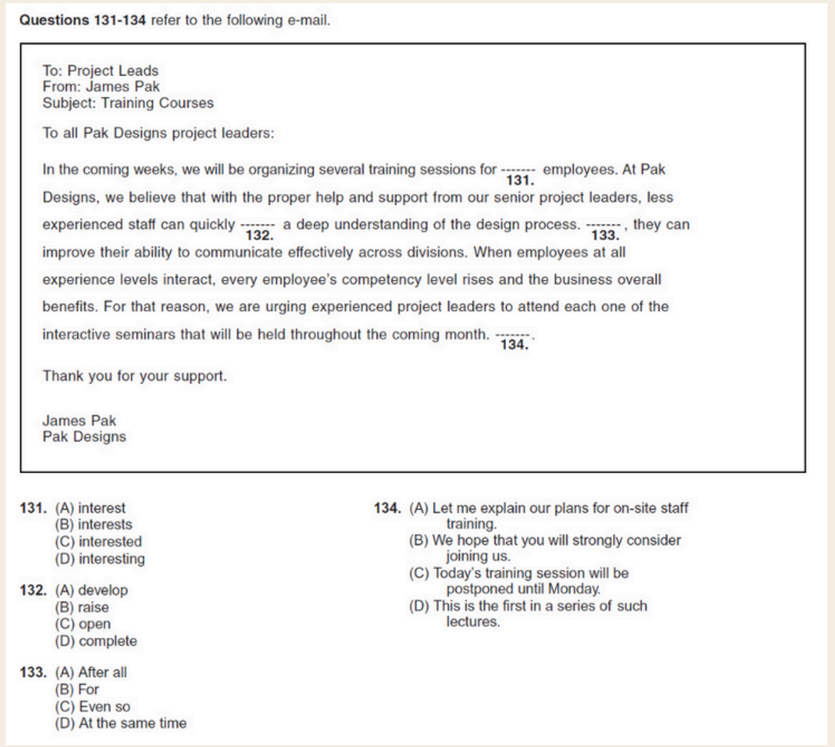
RQ1. 결과난이도- IRT 문항난이도 모수 추정방식

- **활용 모형:** IRT Rasch Testlet Model (Wang & Wilson, 2005)
  - 단위검사는 IRT 가정인 지역독립성 가정을 위배하기 때문

$$\log\left(\frac{p_{ni1}}{p_{ni0}}\right) = \theta_n - b_i + \gamma_{nd(i)}$$

  - **단위검사(testlet)란?**

    A **testlet** is a bundle of items that **share a common stimulus** (e.g., a reading comprehension passage or a figure) (Wainer&Kiely, 1987, as cited in Wang & Wilson, 2005)



Questions 131-134 refer to the following e-mail.

To: Project Leads
From: James Pak
Subject: Training Courses

To all Pak Designs project leaders:

In the coming weeks, we will be organizing several training sessions for ------- **131.** employees. At Pak Designs, we believe that with the proper help and support from our senior project leaders, less experienced staff can quickly ------- **132.** a deep understanding of the design process. ------- **133.**, they can improve their ability to communicate effectively across divisions. When employees at all experience levels interact, every employee's competency level rises and the business overall benefits. For that reason, we are urging experienced project leaders to attend each one of the interactive seminars that will be held throughout the coming month. ------- **134.**.

Thank you for your support.

James Pak
Pak Designs

131. (A) interest
(B) interests
(C) interested
(D) interesting

132. (A) develop
(B) raise
(C) open
(D) complete

133. (A) After all
(B) For
(C) Even so
(D) At the same time

134. (A) Let me explain our plans for on-site staff training.
(B) We hope that you will strongly consider joining us.
(C) Today's training session will be postponed until Monday.
(D) This is the first in a series of such lectures.

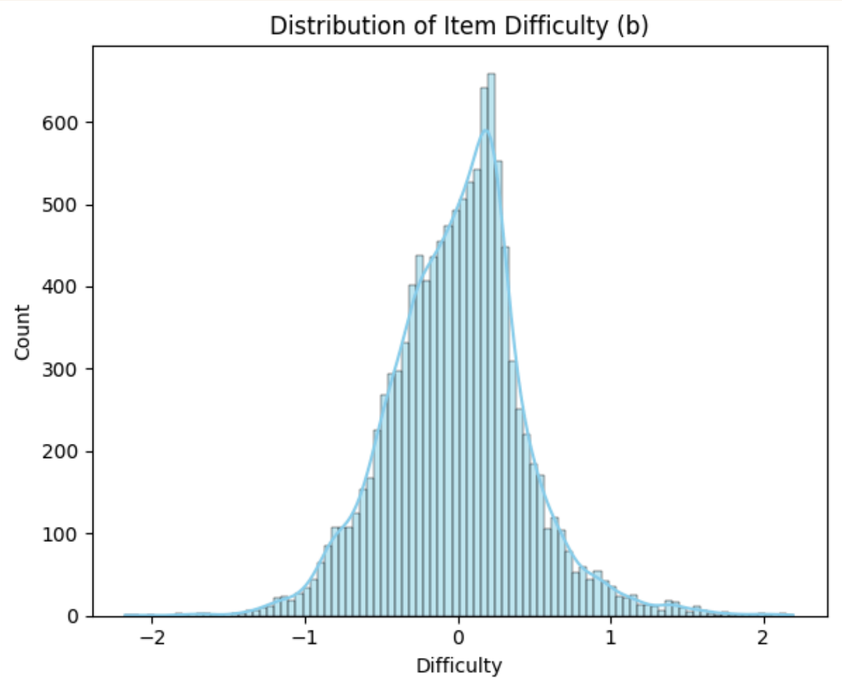
RQ1. 결과난이도- IRT 문항난이도 모수 추정방식

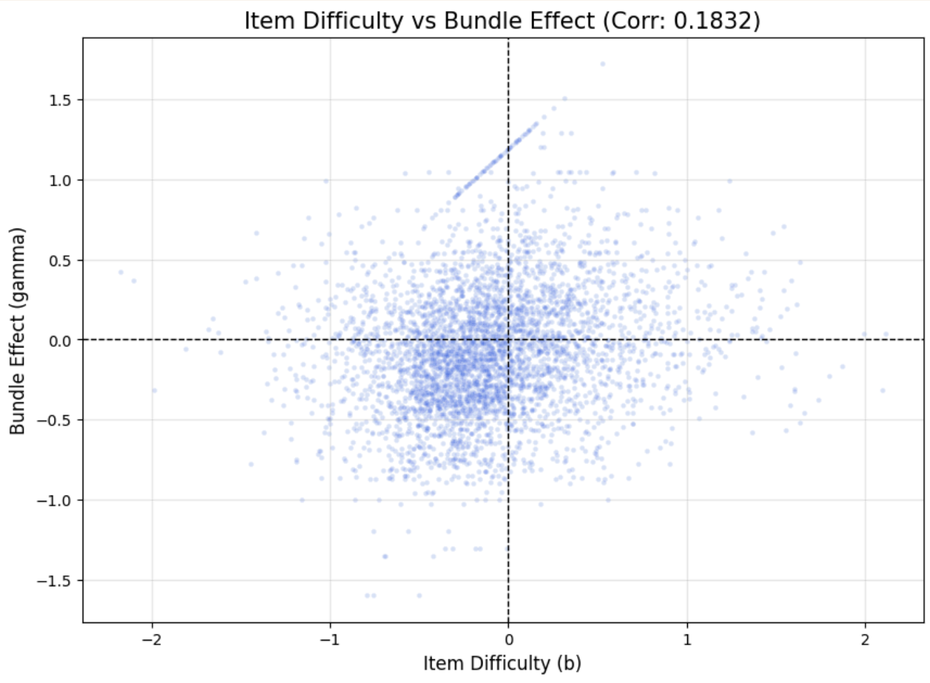
- 파라미터 업데이트: CJMLE (constrained joint maximum likelihood estimator)
  - 대규모 데이터에 적합

In particular, we establish a notion of statistical consistency for a constrained JML estimator, under an asymptotic setting that both the numbers of items and people grow to infinity and that many responses may be missing. A parallel computing algorithm is proposed for this estimator that can scale to very large datasets. Via simulation studies, **we show that when the dimensionality is high, the proposed estimator yields similar or even better results than those from the MML estimator, but can be obtained computationally much more efficiently.** (Chen et al., 2019, Abstract)

- 해당 문항에 대한 학습자의 ‘첫 시도'를 기준으로 추정

RQ1. 결과난이도- IRT 문항난이도 모수 추정방ㅠ식





- -3 < 난이도 < +3
  (-5 < 난이도 < +5로 제약을 주었음에도 일반적 이론과 일치)
- 정규분포 형태

- 문항난이도와 단위검사 효과 상관 = 0.18
  → 해당 모형이 단위검사 효과와 문항 난이도를 잘 구분하고 있음 확인

RQ1. 혼란도 지표 산출방법

- 정답률만으로는 **맞췄지만 헤맨 문제 / 틀렸지만 그냥 찍은 문제** 구분이 어려움
- Response time(RT)은 속도-정확도/노력을 구분하는 핵심 신호로 자주 사용됨

Process Data in Computer-Based Assessment

Challenges and Opportunities in Opening the Black Box

Marlit Annalena Lindner[1,2] and Samuel Greiff[3]

[1] IWM – Leibniz-Institut für Wissensmedien, University of Tübingen, Germany
[2] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

- **Inputs (attempt-level)**
  - RT: $RT = \log\left(1 + \text{elapsed\_time}_{ms}\right)$
  - Change: $Ch = \mathbf{1}[\text{response\_change} > 0]$

- **Weight Selection (α,β)**
  - 고정 가중치(α,β)는 데이터/플랫폼별 스케일 차이 때문에 부적절 → EdNet 내부에서 추정
  - RT/Change가 proxy를 얼마나 설명하는지로 계수 추정 (간단 logistic regression 사용)

RQ1. 혼란도 지표 산출방법

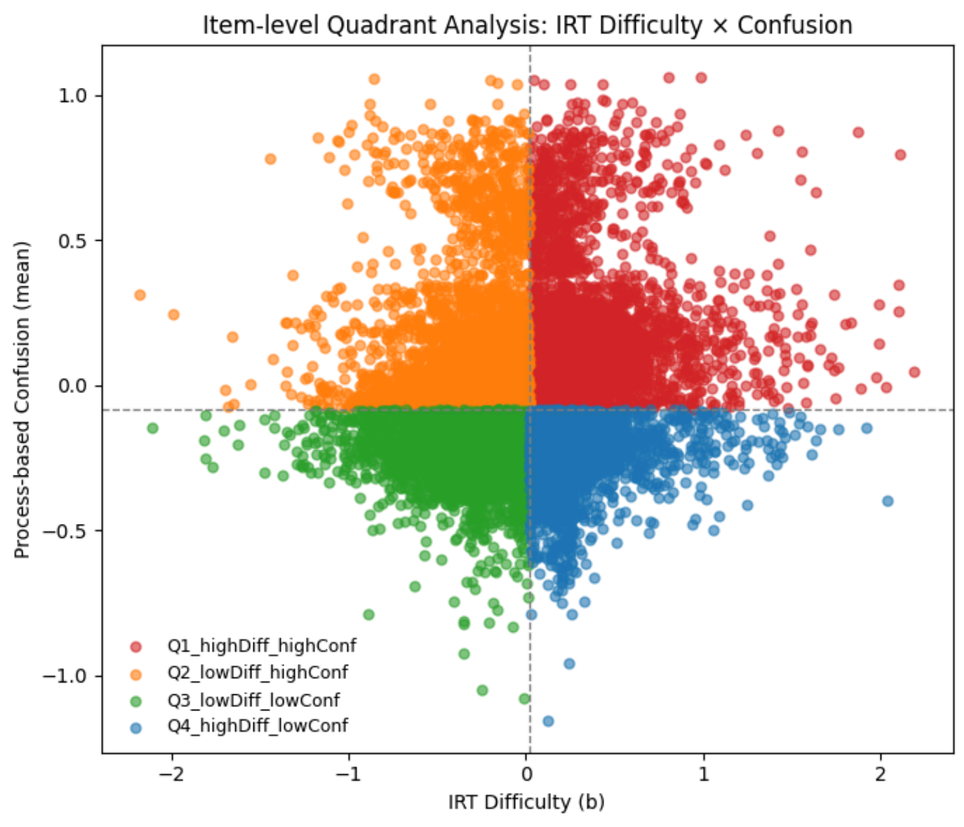
- **confused_attempt(0/1)**: 상위 p%로 라벨링 (분석에서는 10% 사용)

- **Confusion_attempt**
  - $\text{Confusion} = \alpha' x_{RT} + \beta' x_{Ch}$

- **question-level**
  - $q_{\text{confusion\_mean}} = \text{mean}(\text{Confusion})$
  - $q_{\text{confusion\_median}} = \text{median}(\text{Confusion})$
  - $q_{\text{confusion\_rate}} = \text{mean}(\text{confused})$

RQ1. 결과지표 & 과정지표로 문항 유형 구분

x축: IRT 문항난이도 & **y축**: 혼란도 평균 값으로 설정



문항 유형	비율
1사분면 (Q1) 고난이도-고혼란도	0.2730
2사분면 (Q2) 저난이도-고혼란도	0.2260
3사분면 (Q3) 저난이도-저혼란도	0.2740
4사분면 (Q4) 고난이도-저혼란도	0.2270

RQ1. 문항 유형별 행동로그 패턴 분석

1. **선지선택 불확실성**

- 선지 선택 분포의 shannon entropy를 선지 수로 정규화한 값
- 정오답 선지 모두 포함해서 분석
- 값이 낮을수록 응답이 특정 선지에 집중된 것

2. **응답시간 분산**

- 초단위로 변환
- 값이 클수록 학습자 간 응답시간 패턴의 이질성 큰 것

RQ1. 문항 유형별 행동로그 패턴 분석

문항 유형	선지선택 불확실성 평균	응답시간 분산 평균
1사분면 (Q1) 고난이도-고혼란도	0.742160	41.041864
2사분면 (Q2) 저난이도-고혼란도	0.490423	37.059290
3사분면 (Q3) 저난이도-저혼란도	0.472258	26.360258
4사분면 (Q4) 고난이도-저혼란도	0.644829	27.794383

[선지선택 불확실성]

- 저난이도(Q2&Q3)에서 특정 선지에 집중 → **답이 명확했을 경우로 생각 가능**
- 특히 Q1 (고난이도- 고혼란도)에서 높은 수치 보임 → **여러 선지가 매력적**
- 저혼란도 유형 (Q3 & Q4)끼리 비교해보면, 이 중 고난이도 유형(Q4)가 더 선지 선택이 퍼져있음.

[응답시간 분산]

- 고혼란 유형 (Q1 & Q2) > 저혼란 유형(Q3 & Q4)
- **고난이도-고혼란도 유형(Q1)에서 가장 큰 값**

RQ1. 문항 유형별 행동로그 패턴 분석

3. Tag의 개수

- tag는 스킬을 나타내며 정수로 표현되어 있음
- 정수가 정확히 어떤 스킬을 나타내는지 알 수는 없지만, 문항 특성 파악 가능

문항 유형	tag 개수
1사분면 (Q1) 고난이도-고혼란도	1.675949
2사분면 (Q2) 저난이도-고혼란도	2.195719
3사분면 (Q3) 저난이도-저혼란도	2.866604
4사분면 (Q4) 고난이도-저혼란도	2.240198

- 고난이도- 고혼란 문항 (Q1)의 tag 수가 가장 적음
- **고혼란도 유형의 tag 수가 적은 편**

→ tag가 정확히 어떤 스킬을 요구하는지는 알 수 없지만, **어려운 하나의 능력을 요구하는 문항**이 혼란도와 난이도가 모두 높은 문항이라고 추론해볼 수 있음

RQ1. 문항 유형별 행동로그 패턴 분석

4. 파트별 문항유형 구성 비율

파트 / 문항 유형	1사분면 고난이도-**고혼란도**	2사분면 저난이도-**고혼란도**	3사분면 저난이도-**저혼란도**	4사분면 고난이도-저혼란도
LC 1	0.07717	**0.70097**	0.21865	0.00322
LC 2	0.03613	0.01290	**0.53226**	0.41871
LC 3	0.04040	0.08448	**0.64004**	0.23508
LC 4	0.03861	0.08020	**0.58020**	0.30099
RC 5	**0.39202**	0.14676	0.18245	0.27876
RC 6	0.37000	**0.61583**	0.01000	0.00417
RC 7	**0.51746**	0.47059	0.00000	0.01195

- **RC 파트(5,6,7) 문항들이 고혼란도 비율 높음**
- 파트 6,7은 대부분의 문항이 고혼란도 유형(Q1&2)

  → 상대적으로 긴 지문 + 단위검사 특성

RQ1. 시스템 및 시험의 품질 개선 시 참고사항

- **Q2(저난이도-고혼란도): 학습 가능한 실패**
  - 풀 수는 있지만 학생이 고민해서 풀어야 함 → 문항의 학습적 가치가 존재
  - 비계(scaffolding) 작동 → 해설/피드백의 도움으로 학습 효과가 크게 날 수 있는 구간
  - ZPD(근접발달영역; zone of proximal development)에 속하는 문항
  - 주의: 문항 문구 및 조건이 애매해서 불필요한 혼란 발생 가능성 존재
    → Q2를 무조건 좋은 문항유형 단정 X

- **Q4(고난이도-저혼란도): 학습이 발생하지 않는 문항**
  - 고난도 문항에서 혼란이 발생하지 않는다는 것 → 고민하지 않음
  - 문항 추측 (찍기) 유발 문항
  - 평가/학습 모두에 대한 정보제공 능력 약함
  - 해결책: 보기 강화(실패 원인 분화) / 단계형 분해(어디서 막히는지 로그 확보) / 브릿지 문항 추가 (문항 풀이 시도를 유도) 등

연구문제 2

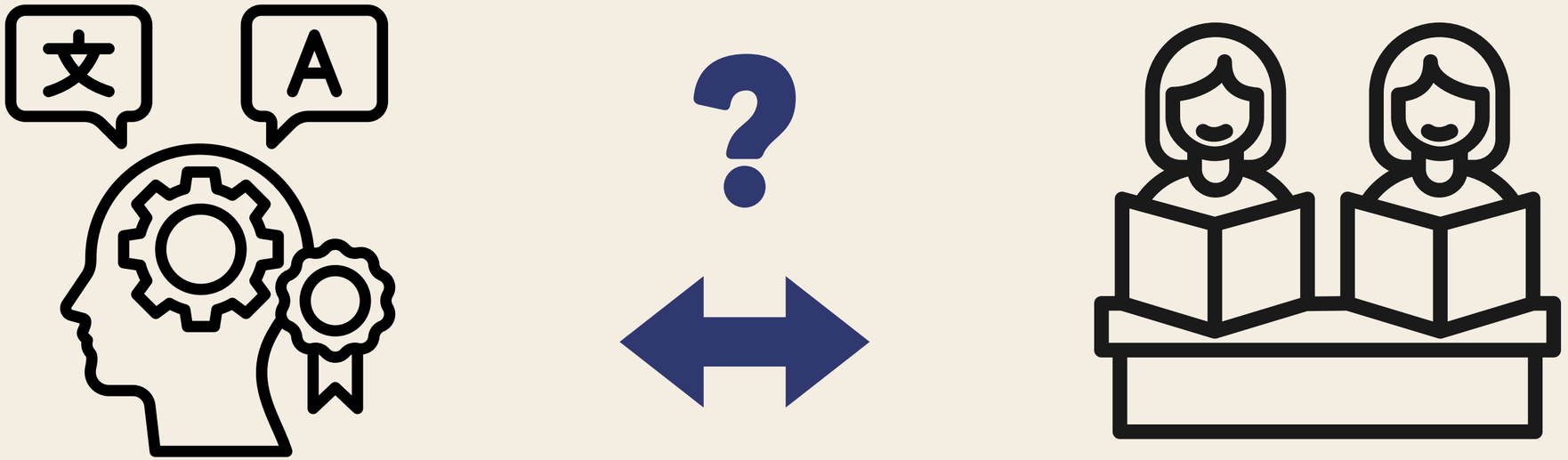
학습 유형을 학습자의 **능력**과 더불어

학습과정 변수로 군집화할 수 있는가?

RQ2. 변수 요약

변수명	설명
accuracy	정답률
n_problems	푼 문항 수
time_on_task_total_ms	번들 투입시간 합
avg_response_time_ms	문항당 평균 풀이시간
accuracy_per_time	투입시간 대비 정답률
accuracy_per_problem	푼 문제 대비 정답률
abandon_rate	시작 대비 미완료 비율
consecutive_days	최장 연속학습일수
active_days	학습일수

RQ2. 변수 해석 관점



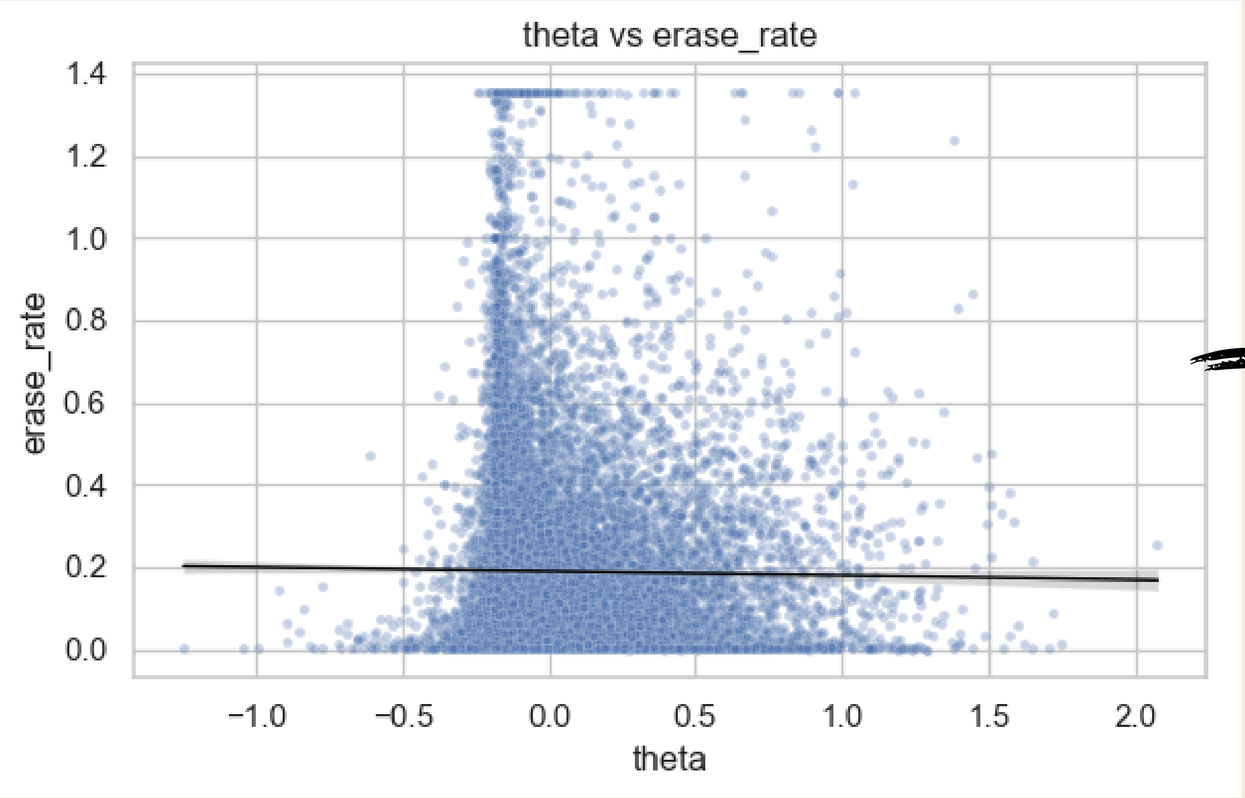
능력을 단순히 결과물로 받아들이는 것이 아닌

WHY?라는 관점에서 학습 과정을 통해 학습자를 파악하고자 함

RQ2. 학습자 특성 - 능력 간 Pearson 상관분석

FEATURE	PEARSON	PEARSON_P
accuracy	0.593152	0
n_problems	0.647012	0
time_on_task_total_ms	0.550002	0
active_days	0.530994	0
consecutive_days	0.497193	0
accuracy_per_time	0.276144	0
adaptive_offer	0.170166	0
explanation_adoption_rate	0.084195	0
explanation_after_wrong_rate	0.027385	0.0000271
source_entropy	0.09456	0
explanation_time_per_problem_ms	-0.006887	0.3066
answer_change_rate	0.023033	0.000842
avg_response_time_ms	0.021907	0.000789
erase_rate	-0.012061	0.1247
media_play_rate	-0.0332	0
accuracy_per_problem	-0.209332	0
undo_rate	-0.149635	0
abandon_rate	NaN	

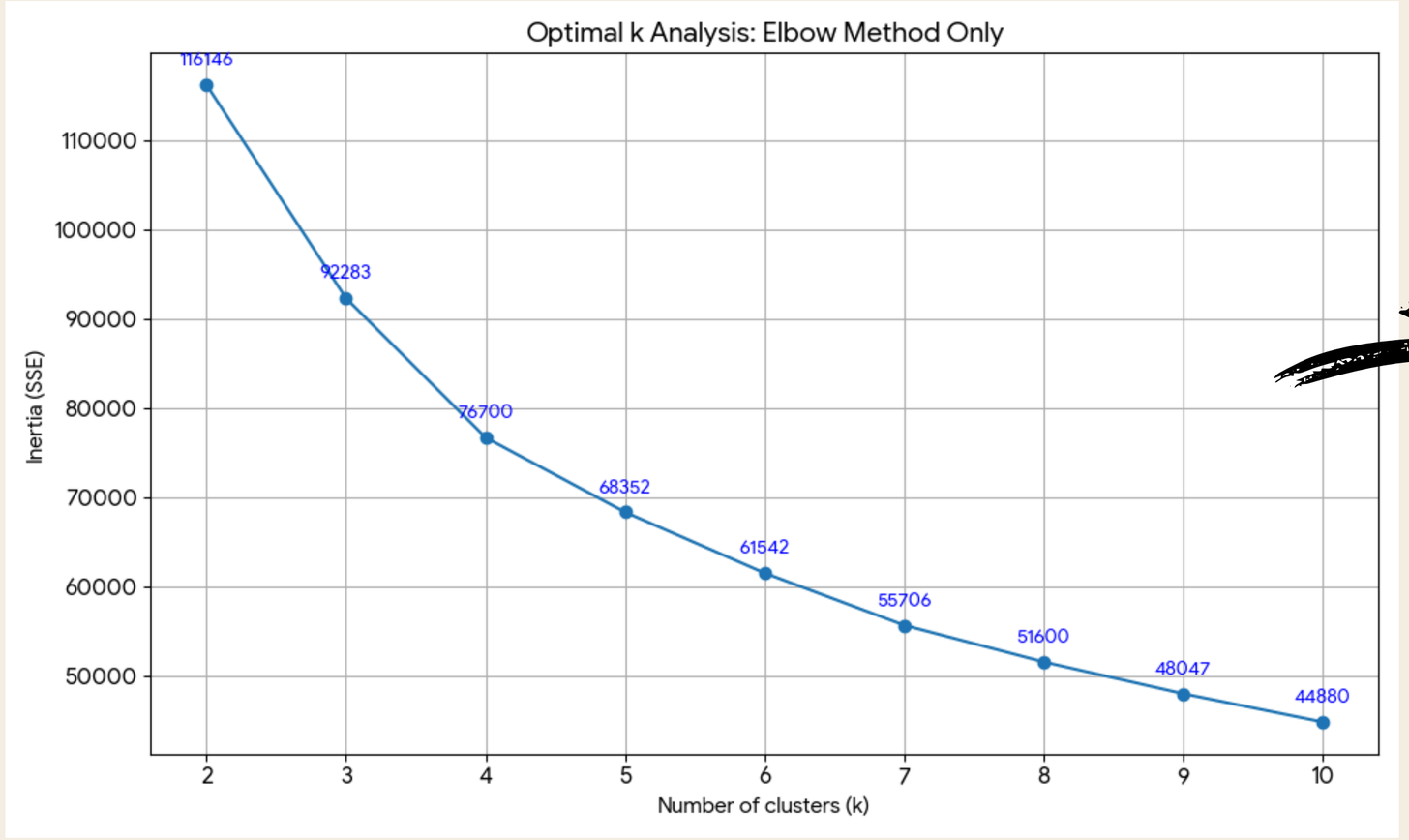
RQ2. 비선형 관계: Spearman 상관 분석



비선형성

- Spearman 상관계수

RQ2. 클러스터링: 최적 K수 결정 (Elbow Method)



N = 4

RQ2. 클러스터링: 군집 수 결정 기준

SSE : 군집 내 데이터 포인트와 centroid 간의 거리 제곱합
=> inertia / WCSS (Within Clusters Sum of Squares)

- 값이 작을수록 군집내 응집도가 높음!

Elbow point

- SSE의 감소비율이 급격하게 작아지는 지점이 최적의 군집수!

RQ2. 클러스터별 학습자 특성 요약

cluster	n_students	accuracy	avg_response_time_ms	n_problems	time_on_task_total_ms	consecutive_days	active_days	erase_rate	undo_rate	answer_change_rate	explanation_adoption_rate	explanation_after_wrong_rate	source_entropy	adaptive_offer	media_play_rate	accuracy_per_time	accuracy_per_problem	theta	theta_z
0	8670	0.6143	10.2868	6.3771	14.0626	2.3211	6.6141	0.0889	0.0072	0.0624	0.2499	0.0756	0.6124	0.0493	0.3015	0.0438	0.0978	0.009	0.0333
1	4103	0.5049	10.1955	3.6413	11.8554	1.0441	1.2445	0.2273	0.0768	0.1434	0.1653	0.0302	0.1148	0.0004	0.4648	0.0429	0.1507	-0.1697	-0.6281
2	6922	0.6419	10.1984	5.2387	14.3404	2.8264	7.8755	0.2917	0.0222	0.1528	0.7789	0.2973	0.5656	0.0395	1.0165	0.0449	0.1259	-0.0766	-0.2836
3	3782	0.6683	10.2893	7.5494	16.5086	9.6362	38.6433	0.2052	0.0119	0.2066	0.6508	0.2666	0.5773	0.0716	0.7602	0.0405	0.0891	0.3037	1.1242

- **Cluster 0** : 자신의 촉을 믿는 (답 변경 적음) 학습자
- **Cluster 1** : 학습 이탈/부진 위험 학습자
- **Cluster 2** : 전략적 (선택지 제거/해설 채택) 학습자
- **Cluster 3** : 우수 학습자

RQ2. 클러스터별 학습자 특성 요약

cluster	accuracy	avg_response_time_ms	n_problems	time_on_task_total_ms	consecutive_days	active_days	erase_rate	undo_rate	answer_change_rate	explanation_adoption_rate	explanation_after_wrong_rate	source_entropy	adaptive_offer	media_play_rate	accuracy_per_time	accuracy_per_problem	theta	theta_z
0	0.6143	10.2868	6.3771	14.0626	2.3211	6.6141	0.0889	0.0072	0.0624	0.2499	0.0756	0.6124	0.0493	0.3015	0.0438	0.0978	0.009	0.0333
1	0.5049	10.1955	3.6413	11.8554	1.0441	1.2445	0.2273	0.0768	0.1434	0.1653	0.0302	0.1148	0.0004	0.4648	0.0429	0.1507	-0.1697	-0.6281
2	0.6419	10.1984	5.2387	14.3404	2.8264	7.8755	0.2917	0.0222	0.1528	0.7789	0.2973	0.5656	0.0395	1.0165	0.0449	0.1259	-0.0766	-0.2836
3	0.6683	10.2893	7.5494	16.5086	9.6362	38.6433	0.2052	0.0119	0.2066	0.6508	0.2666	0.5773	0.0716	0.7602	0.0405	0.0891	0.3037	1.1242

- **Cluster 0** : 자신의 촉을 믿는 (답 변경 적음) 학습자
- **Cluster 1** : 학습 이탈/부진 위험 학습자
- **Cluster 2** : 전략적 (선택지 제거/해설 채택) 학습자
- **Cluster 3** : 우수 학습자

RQ2. 클러스터별 학습자 특성 요약

cluster	accuracy	n_problems	consecutive_days	active_days	undo_rate	answer_change_rate	explanation_adoption_rate	explanation_after_wrong_rate	adaptive_offer	media_play_rate	accuracy_per_problem	theta
0	0.6143	6.3771	2.3211	6.6141	0.0072	0.0624	0.2499	0.0756	0.0493	0.3015	0.0978	0.009
2	0.6419	5.2387	2.8264	7.8755	0.0768	0.1528	0.7789	0.2973	0.0395	1.0165	0.1259	-0.0766

Cluster 0 : 자신의 촉을 믿는 (답 변경 적음) 학습자

vs.

Cluster 2 : 전략적 (선택지 제거/해설 채택) 학습자

연결고리

군집 0과 군집 2를 분석하여

앞선 연구문제 둘 사이의 연결점 파악

RQ1

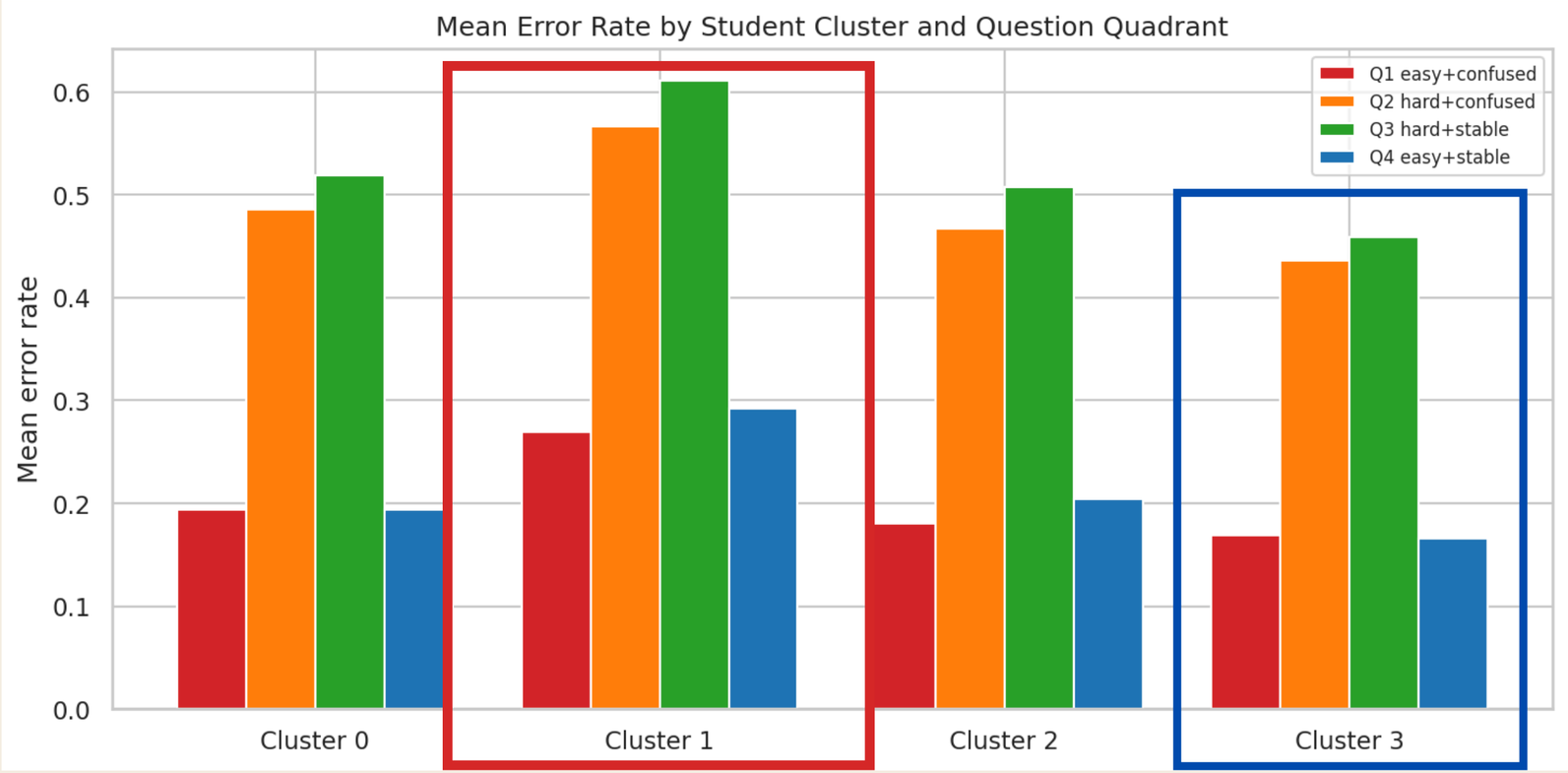
문항
단위

연결고리?

RQ2

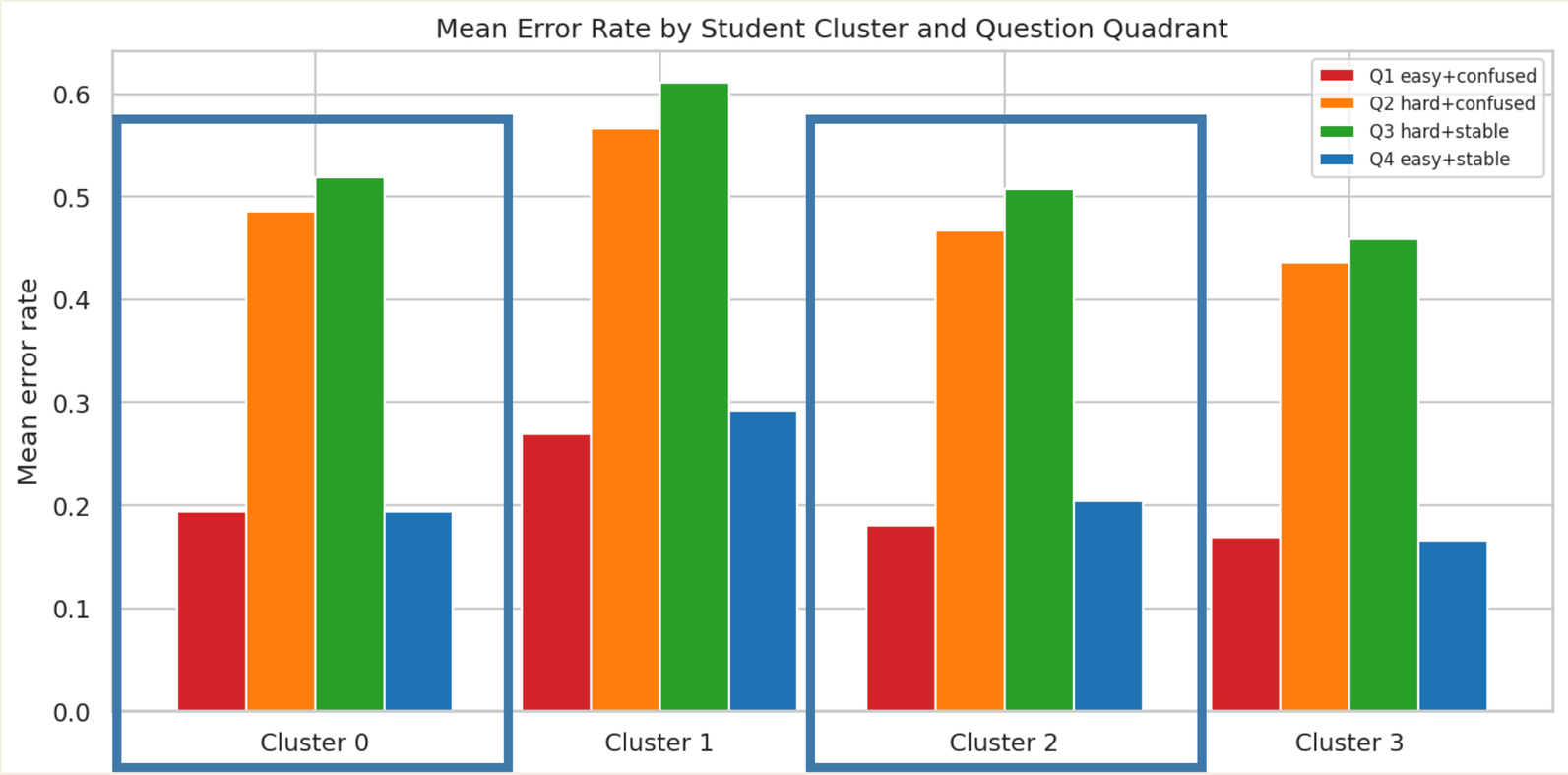
학습자
단위

문항별, 클러스터별 오답률



문항 전반적으로 오답률이 가장 높고 낮은 군집은 시각적으로 **구분 가능**

문항별, 클러스터별 오답률



나머지 군집(특히 군집 0, 2)에 대해서는 문항 오답률만으로는 **구분 불가능**

학생 능력치, 정답률 비교

cluster	학생 수	정답률	문항당 평균 풀이시간	...	풀이량 대비 성과	theta
0	8670	0.6143	10.2868	...	0.0978	0.009
1	4103	0.5049	10.1955	...	0.1507	-0.1697
2	6922	0.6419	10.1984	...	0.1259	-0.0766
3	3782	0.6683	10.2893	...	0.0891	0.3037

마찬가지로 cluster 0과 2는 theta값(능력치)과 정답률(accuracy)이 크게 차이 나지 않는다

→ cluster의 feature로 판단!

(앞서 Q2에서 k를 4로 잡은 이유와도 연결 가능)

Cluster 0-2 비교

feature	cluster_0_mean	cluster_2_mean	mean_difference	test_statistic	significant	cohens_d	effect_size
문항당 평균 풀이시간	10.2868	10.1984	0.0884	36034112	TRUE	0.3044	small
푼 문항 수	6.3771	5.2387	1.1385	47373832	TRUE	1.1651	large
번들 투입시간 합	14.0626	14.3404	-0.2778	24839811	TRUE	-0.2527	small
…	…	…	…	…	…	…	…
최장 연속 학습일	2.3211	2.8264	-0.5052	24074801	TRUE	-0.3163	small
투입시간 대비 성과	0.0438	0.0449	-0.001	26495127	TRUE	-0.129	negligible
풀이량 대비 성과	0.0978	0.1259	-0.0282	11583576	TRUE	-1.1772	large

MannWhitney-U 검정 :

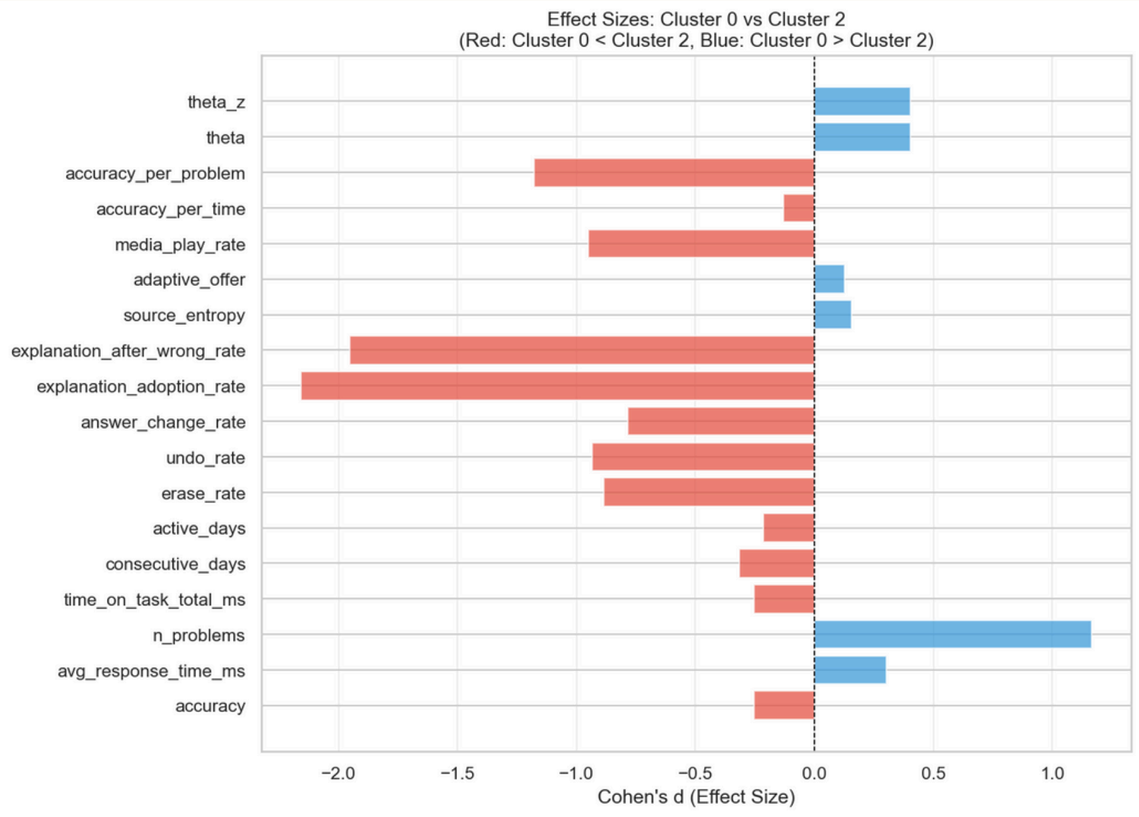
비모수적 추정 방법, 한쪽으로 치우친 데이터에 적용하기 좋음

중앙값과 순위(rank)를 비교중심으로 사용

MannWhitney-U값과 함께 p-value를 보고 통계적으로 유의한지 판단

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

Cluster 0-2 비교 (feature 영향력)



Cohen's d

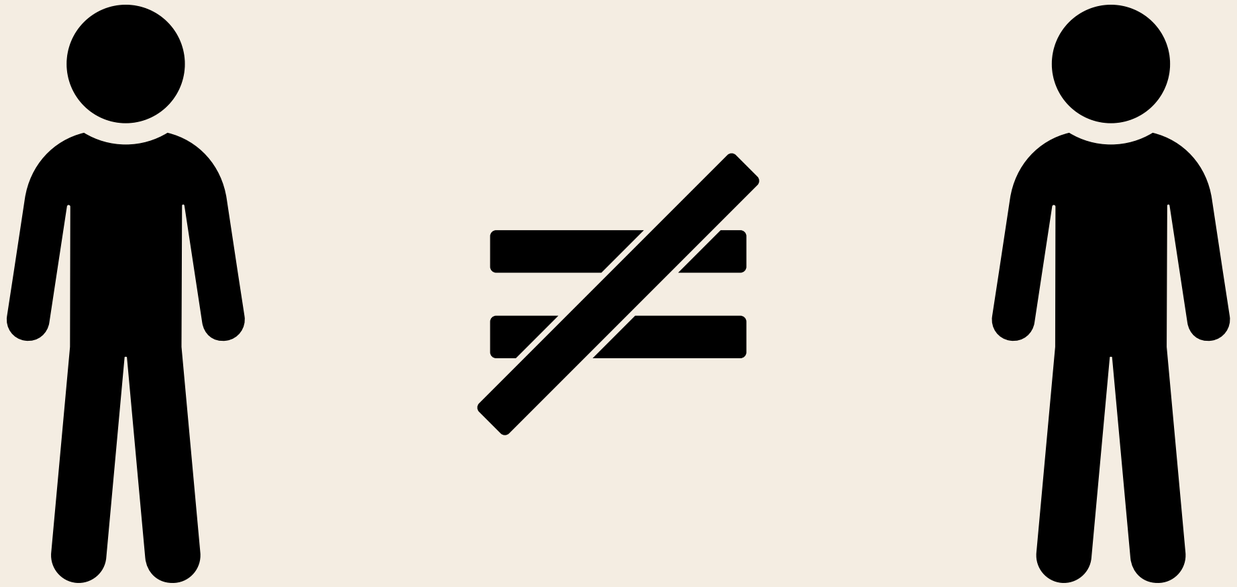
두 집단 사이의 평균 차이를 표준편차로 나눈 값
→ 차이가 얼마나 큰지 표준화된 수치로 보여줌

- 0.2 (Small): 효과가 작음
- 0.5 (Medium): 효과가 중간
- 0.8 (Large): 효과가 큼

→ 0.2보다 작은 경우 ‘negligible’(무시가능)

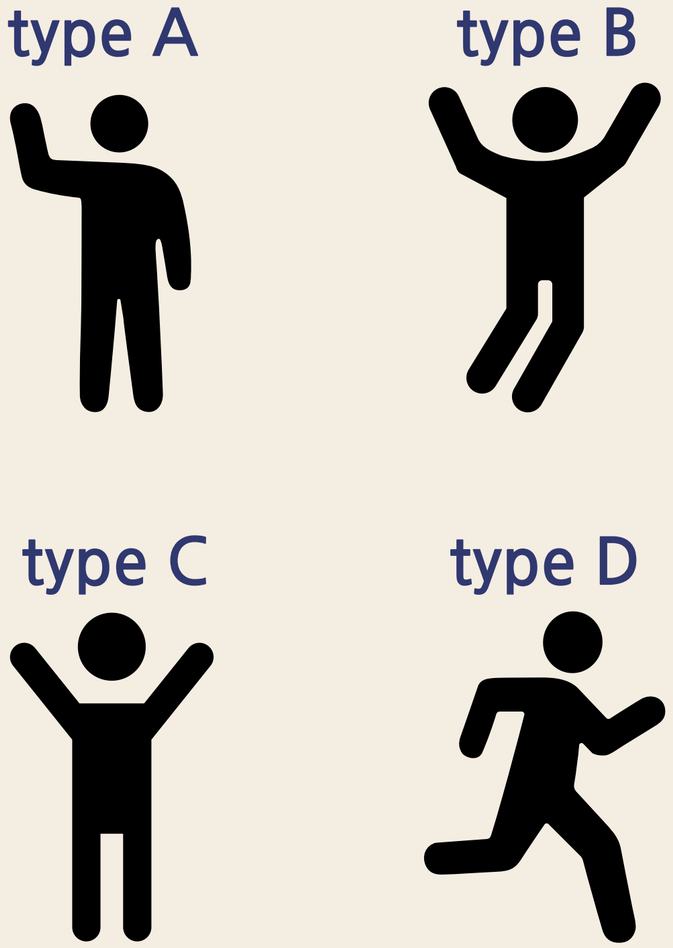
$$d = \frac{M_1 - M_2}{s_p} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

CLUSTER 0 CLUSTER 2



IRT기반의 능력치만으로는 구별이 쉽지않은 cluster 0과 2은
학습자 행동 feature 비교분석을 통해 명확히 구분 가능
결과가 아닌, **과정을 확인하고 판단할 수 있는** 좋은 비교집단

결론



결과 값에 따른 ‘줄세우기’가 아닌,
유형별 군집화를 통해 **‘과정 중심 교육’**을 기대해볼 수 있다

분석 한계 및 보완 방향

- **Confusion 지표 불안정성**
  - 지표 정의 방식에 따라 해석이 달라질 수 있음
  - 정량적 검증 필요
- **response_change 희소성**
  - 로그 자체가 적어 행동 신호가 약함
- **문항 메타정보 미결합**
  - 문항 내용, 정답, 태그와의 통합 분석 미비
- **학습자 유형 분석 미완성**
  - 피처 설계 단계로 분석 모델은 아직 없음
- **Inference 체계 부족**
  - 해석 프레임워크 부재, 가설 기반 분석 필요

참고문헌

- Chen, Y., Li, X., & Zhang, S. (2017, December 19). Joint maximum likelihood estimation for high-dimensional exploratory item response analysis. arXiv.org. https://arxiv.org/abs/1712.06748
- Education Testing Service(ETS). (2025). TOEIC Listening & Reading Test Examinee Handbook. https://www.ets.org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf
- Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment: Challenges and opportunities in opening the black box [Editorial]. European Journal of Psychological Assessment, 39(4), 241-251. https://doi.org/10.1027/1015-5759/a000790
- Riiid. (2020). EdNet: A large-scale hierarchical dataset in education. GitHub. https://github.com/riiid/ednet
- Wang, W., & Wilson, M. (2005). The Rasch Testlet model. Applied Psychological Measurement, 29(2), 126-149. https://doi.org/10.1177/0146621604271053