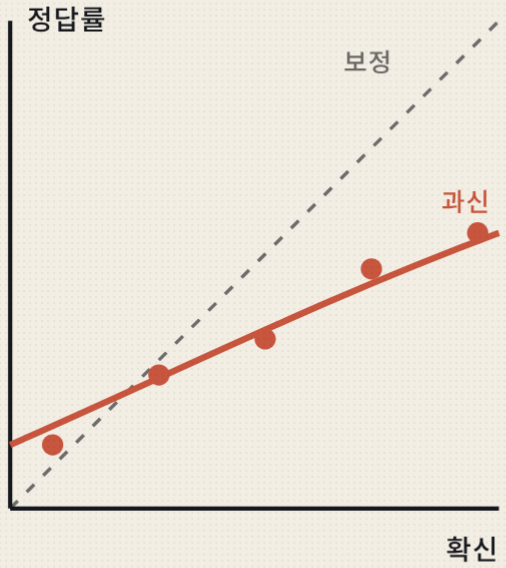
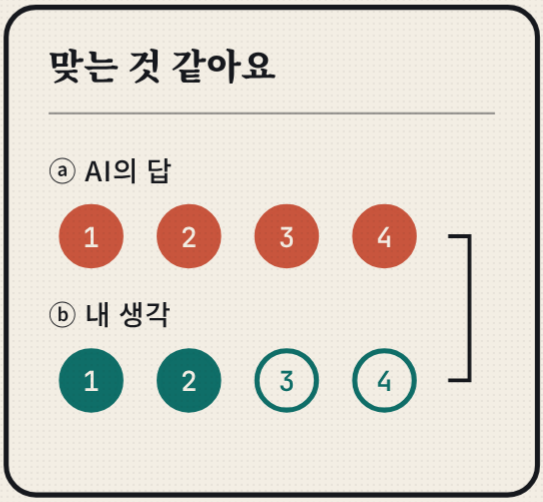
09

제9회

3부 · 여섯 요건을 하나씩




확신은 하나가 아니라 둘이다




ⓐ – ⓑ 판단이 어디로 기울었나

*AI Insight Column*

# 확신도 — 메타인지적 보정

'맞는 것 같아서요' · AI와 함께 배우는 교실에서 확신은 하나가 아니라 둘이다

보정과 변별 | 두 칸 확신도 | 과신 | 확신하고 틀린 순간

# 9. 확신도 — 메타인지적 보정

연재 〈에이전틱 사고와 온톨로지〉 제 9 회 — 3 부: 여섯 요건을 하나씩

## ‘맞는 것 같아서요’

학생에게 객관식 문제를 제시하고 답을 고른 후에 ‘왜 이 답을 골랐느냐’고 물으면 가장 자주 돌아오는 대답이 있습니다. 바로 ‘맞는 것 같아서요’입니다. 맞는 것 같다? 얼마의 확신이 있는 것일까요?

**저는 오랫동안 이 말을 성의 없는 대답으로 여겼습니다. 하지만 지금은 다르게 봅니다. 이 말은 학생이 자기 머릿속에 있는 정보를 있는 그대로 알려 주는 문장입니다. 학생에게 있는 것은 ‘맞는 것 같다’는 막연한 느낌 하나뿐이고 그 느낌을 얼마나 믿어도 되는지에 대한 정보는 아예 없는 것입니다. 사실 앞서 제시한 확신의 정도를 학생들이 직접 판단하기에는 어려워 보입니다. 그렇기 때문에 막연한 느낌만을 이야기하는 것이죠.**

장면 하나를 떠올려 보겠습니다. 두 학생이 분수가 들어 있는 일차방정식 문제를 나란히 풀었는데, 한 명은 맞고 한 명은 틀렸습니다. 이유를 물으면 두 학생 모두 ‘맞는 것 같아서요’라고 답하니, 이 말만으로는 둘을 구별할 수 없습니다. 하지만 한 걸음 더 들어가 물으면 차이가 드러납니다. 맞힌 학생은 ‘양변에 같은 수를 곱했으니 해가 바뀌지 않는다’고 이유를 댔지만, 틀린 학생은 ‘이런 문제는 늘 이렇게 풀었다’는 말밖에 하지 못했습니다. 겉으로 보인 확신은 같았지만 그 확신이 발 딛고 선 근거는 전혀 달랐던 것입니다.

느낌만 있고 그 느낌을 얼마나 믿을 수 있는지 모르는 상태, 이번 회는 바로 그 빈자리에 대한 이야기입니다. AI 가 교실에 들어온 뒤로 이 빈자리가 왜 더 넓어졌는지, 그리고 교사가 그 자리를 무엇으로 채울 수 있는지 함께 살펴보겠습니다.

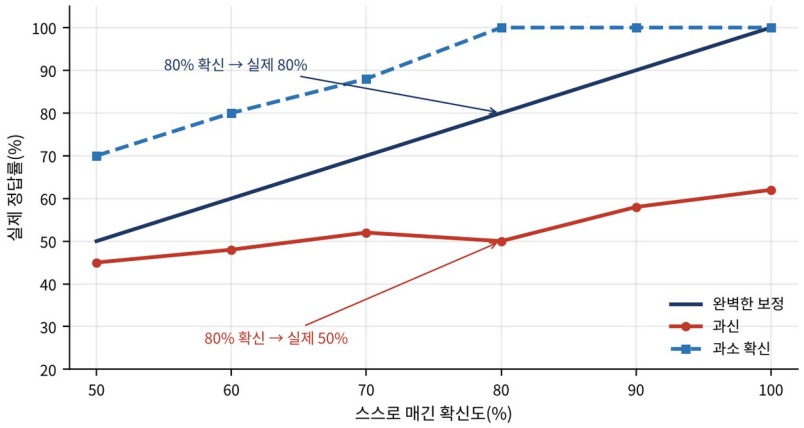
## 요건: 자기 판단이 얼마나 확실한지 스스로 알고 있을 것

다섯 번째 요건은 ‘확신도’입니다. 무언가를 안다고 할 때 자신이 그것을 얼마나 확실하게 아는지 스스로 깨닫고 밖으로 표현할 수 있는 상태를 뜻합니다. 목표가 분명하고 지식의 구조가 잡혀 있고 제약 조건과 기억이 잘 정리되어 있더라도, 지금 내린 판단이 그중 어디에 기대고 있는지 모르면 어디부터 확인해야 할지 정할 수 없습니다. 즉, 확신도는 나머지 요건 가운데 무엇을 먼저 점검할지 우선순위를 정해 주는 요건입니다.

**이 요건은 세 단계로 나누어 볼 수 있습니다. 첫째는 판단마다 확신의 크기가 다르다는 사실을 아는 것이고 둘째는 자신이 느끼는 확신의 크기가 실제 정답 여부와 맞아떨어지는 것이며 셋째는 그 확신을**

**말이나 숫자로 분명하게 꺼내 놓을 수 있는 것입니다. 머릿속으로만 어렴풋이 느끼는 확신은 교사도 알아볼 수 없고, 학생 스스로도 다음 행동을 결정하는 근거로 삼기 어렵습니다.**

제목에 붙인 '보정'이라는 말은 이 가운데 두 번째를 가리킵니다. 저울의 영점을 맞추듯이 내가 느끼는 확신의 눈금과 실제 정답률의 눈금에 맞게 맞추는 일입니다. 80 퍼센트쯤 확신한다고 답한 문제들 가운데 실제로 열 문제 중에 여덟 문제를 맞혔다면 보정이 잘된 것이고 다섯만 맞혔다면 '과신' 이 되고 열 문제를 모두 맞혔다면 '과소 확신'이 되는 거죠(〈그림 1〉).




**〈그림 1〉 보정 곡선으로 본 과신과 과소 확신(개념도)**

**보정과 자주 헷갈리는 개념으로 '변별'이 있습니다. 보정이 확신의 크기와 실제 정답률의 일치를 묻는다면 변별은 맞힌 문제와 틀린 문제에 서로 다른 크기의 확신을 주었는지를 묻습니다. 그러니까 보정은 전체적인 것에 대한 자신의 생각이나 확신의 정도를 조정하는 것이 되며 변별은 그 과정에서 하나의 문제에 대해 내가 얼마나 확신을 주었는가를 이야기합니다.**

**Schraw(2009)가 메타인지적 감시 지표들을 정리하며 밝혔듯이 둘은 서로 다른 측면을 잽니다. 모든 문제에 확신을 조금씩 높게 매기는 학생은 보정은 부정확해도 변별은 뛰어날 수 있고, 평균 확신도는 정답률과 비슷해도 어느 문제가 위험한지 전혀 가려내지 못하는 학생도 있습니다.**

**〈표 1〉 보정과 변별의 구분**

| 구분 | 보정(calibration) | 변별(resolution) |
|---|---|---|
| 핵심 질문 | 내가 느끼는 확신의 크기가 실제 정답률과 일치하는가? | 맞힌 문제와 틀린 문제에 서로 다른 확신을 주었는가? |
| 어긋날 때의 모습 | 전반적으로 지나치게 과신하거나 전반적으로 위축된다. | 쉬운 문제와 어려운 문제에 똑같은 확신을 보인다. |
| 교실에서 확인하는 법 | 단원이 끝날 때 확신도 점수 구간별 실제 정답률을 모아 본다. | 틀린 문제에 매긴 확신도와 맞힌 문제에 매긴 확신도를 나란히 비교한다. |

| 구분 | 보정(calibration) | 변별(resolution) |
|---|---|---|
| 수업에 주는 의미 | 공부를 언제 멈출지 결정하는 데 영향을 준다. | 어떤 문제를 다시 복습할지 고르는 데 영향을 준다. |

이 연재에서 말하는 확신도 요건은 보정과 변별을 모두 포함하지만, 교실에서는 변별이 먼저입니다. 학생이 전반적으로 조금 과신하는 경향이 있다면 교사가 감안해서 들으면 되지만, 어떤 문제가 불안한지 전혀 구별하지 못하는 상태는 감안해 줄 방법이 없습니다. 이번 회에서 제안하는 수업 장치도 결국은 이 변별력을 살려 내기 위한 것입니다.

## 심리학에서는 오래전부터 다뤄 온 주제입니다

Flavell(1979)은 '메타인지'라는 개념을 정립하면서 인지에 대한 지식과 인지에 대한 조절을 하나로 묶었습니다. 우리가 이미 알고 있듯이 메타인지는 내가 무엇을 알고 무엇을 모르는지 아는 능력이 바로 여기에 해당합니다. Nelson 과 Narens(1990)는 이를 더 체계적으로 정리하여, 생각의 층위를 '대상 수준'과 '메타 수준'으로 나누고 메타 수준이 대상 수준을 살피는 '감시'와 행동을 조절하는 '통제'를 맡는다고 보았습니다. 감시만 있고 통제가 뒤따르지 않으면, 모른다는 사실을 알면서도 그냥 넘어가게 됩니다.

감시가 실제로 통제를 움직이는지는 생각보다 증명하기 까다로운 문제였습니다. **학습자가 잘 모른다고 느낀 내용을 더 오래 공부한다는 사실은 오래전부터 알려져 있었습니다. 하지만 그것은 상관관계일 뿐이어서 모른다는 느낌 때문에 더 공부한 것인지 실제로 어려운 내용이라 더 공부한 것인지는 가를 수 없었습니다.** 모른다는 느낌은 그저 함께 따라온 것일 수도 있었기 때문입니다. Metcalfe 와 Finn(2008)은 이 둘을 떼어 놓는 방식으로 답을 찾았습니다. 실제 기억 수준은 똑같이 맞춰 두고 학습자가 느끼는 이해도만 다르게 만든 다음, 무엇을 다시 공부하겠다고 고르는지 살펴본 것입니다.

세 차례의 실험에서 결과는 한결같았습니다. 첫 번째 실험에서는 단어 쌍을 두 번에 나누어 공부하게 했습니다. 어떤 쌍은 처음에 세 번, 두 번 째에 한 번 보여 주었고, 다른 쌍은 반대로 처음에 한 번, 두 번 째에 세 번 보여 주었습니다. 모두 합한 반복 횟수가 같으니 최종 기억 성적은 비슷했지만, 두 번째 학습 직후에 매긴 이해도 판단은 서로 달랐습니다. 그리고 학습자들이 다시 공부할 항목을 고를 때는 실제 성적이 아니라 이 착각 섞인 판단을 따라갔습니다. 두 번째 실험에서는 시간이 조금 지난 뒤에 이해도를 판단하게 했습니다. 그러자 판단의 치우침이 사라졌고, 다시 공부할 항목을 고를 때의 치우침도 함께 사라졌습니다.

세 번째 실험에서는 Koriat 와 Bjork(2005)의 방법을 빌려, 앞 단어에서 뒤 단어로 떠올리는 쌍과 뒤 단어에서 앞 단어로 떠올리는 쌍을 비교했습니다. 뒤에서 앞으로 떠올리는 쌍은 실제 성적이 훨씬 나빴는데도 학습자들은 두 종류에 비슷한 확신을 보였고, 다시 공부할 항목도 그 잘못된 판단대로 골랐습니다. 결국 학습자는 실제로 얼마나 아는지가 아니라 얼마나 안다고 느끼는지에 따라 무엇을 다시 공부할지 정한다는 뜻입니다. 확신은 머릿속의 기분에 그치지 않고 다음 행동을 정하는 '입력값' 입니다. 이 입력값이 잘못되면 공부가 필요한 곳은 놔두고 엉뚱한 곳을 공부하게 됩니다. 두 번째 실험은 희망적인 단서도 하나 줍니다. 판단하는 시점만 바꾸어도 입력값이 바로잡혔고 행동도 따라 바뀌었다는 점입니다. 이는 뒤에서 다룰 시험 직후 '사후 판단'의 쓸모와도 맞닿아 있습니다.

**그렇다면 이 감시는 얼마나 정확할까요? Koriat(1997)는 학습자가 자신이 얼마나 아는지 판단할 때 기억을 직접 들여다보는 것이 아니라 문제가 쉬워 보이는지(내재적 단서), 몇 번 공부했는지(외재적 단서), 답이 얼마나 쉽게 떠오르는지(기억적 단서) 같은 간접 단서에 기댄다고 보았습니다. 핵심은 막힘없이 매끄럽게 처리되는 느낌, 곧 '유창성'이 중요한 단서로 쓰인다는 점입니다. Alter 와 Oppenheimer(2009)가 종합했듯이, 글이 읽기 편하거나 익숙하거나 여러 번 본 내용일수록 사람들은 그것이 더 옳고 자신도 더 잘 안다고 느끼며, 이 둔갑은 대부분 의식하지 못하는 사이에 일어납니다.**

게다가 사람의 확신은 대체로 부풀려져 있습니다. 우리가 학교에서 모의고사를 보고 아이들에게 물어보면 대부분 가채점 점수가 더 높은 경우들이 많이 있잖아요. 이런 것이 확신이 부풀려지는 사례에 해당이 됩니다. Fischhoff, Slovic, Lichtenstein(1977)의 연구에서 참가자들은 '확실히 맞다'고 장담한 문제에서도 적지 않게 틀렸습니다. 이 과신은 실제 학습 손실로 이어져서, Dunlosky 와 Rawson(2012)의 연구에서는 자기 답에 후한 점수를 준 대학생일수록 공부를 일찍 멈추었고 나중 시험에서도 덜 기억해 냈습니다. '이제 다 알았다'는 느낌이 착각이었을 때의 손해는 고스란히 학생에게 돌아갑니다.

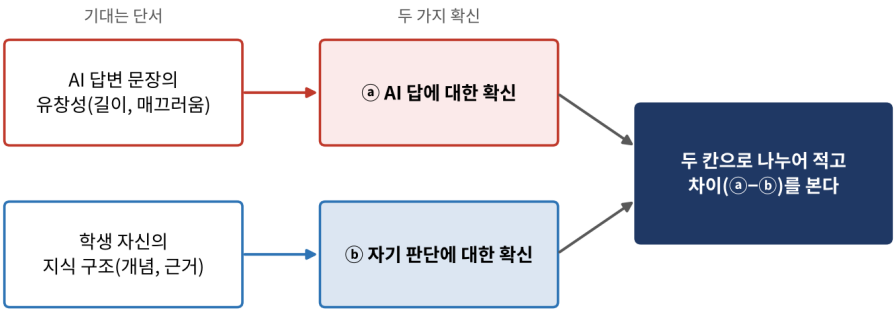
**실제 대학생들을 대상으로 한 연구도 있습니다. Hacker 와 동료들(2000)이 대학 강의에서 한 학기 동안 매 시험 전후로 자기 점수를 예측하게 하자 성적이 높은 학생들은 예측이 점점 정확해졌지만 성적이 낮은 학생들의 시험 전 예측은 그만큼 나아지지 않았습니다. 다만 성적이 낮은 학생들도 시험 직후의 '사후 판단'만큼은 비교적 정확했습니다. 문제와 직접 부딪치고 나면 막연했던 확신이 현실로 끌려 내려오는 것입니다.**

## 여기서부터가 이 연재의 새로운 제안입니다

지금까지 살펴본 연구들은 모두 '내가 이것을 얼마나 아는가'라는 한 가지 확신만을 다룹니다. 하지만 AI 와 함께 문제를 푸는 교실에서는 확신이 둘로 나뉩니다. 아마 지금까지 연구되어오고 이 칼럼에서도 이야기한 확신의 상식을 깨버리는 일이 AI 가 교실로 들어오면서 벌어지게 된 것입니다.

하나는 'AI 가 준 이 답이 맞을 것 같은가'라는 AI 답에 대한 확신이고 다른 하나는 '이 문제에 대한 내 생각이 맞을 것 같은가'라는 자기 판단에 대한 확신입니다. 이 글에서는 이제부터 앞의 확신을 ⓐ, 뒤의 확신을 ⓑ라고 부르겠습니다.

**학생들은 이 두 확신을 잘 구분하지 못하고 하나로 뭉뚱그려 '맞는 것 같아요'라고 말합니다. 그렇기 때문에 이후의 학습이 더 잘 일어나지 않는 것이라고도 생각이 됩니다. 그러나 두 확신은 뿌리가 완전히 다릅니다. ⓐ는 화면에 적힌 문장의 매끄러움과 유창함에 기대고 있습니다. 이 확신에 기대는 순간 학생들은 이전에 이야기한 '인지적 마찰(cognitive friction)이나 바람직한 어려움(desirable difficulties)(Bjork, 1994)를 겪지 못하게 됩니다. ⓑ는 학생 자신이 가진 지식 구조에 기댑니다 ( 〈그림 2〉 ). 이는 ⓐ보다는 매끄럽게 읽히지 않고 심지어 부자연스럽기까지도 하지만 매끄럽게 읽히는 유창성이 확신을 높이는 단서라면 학생들이 매일 마주하는 AI 의 답변은 인류가 지금껏 만들어 낸 가장 유창한 문장입니다. 확신을 부풀리는 조건이 교실 안에 상시로 켜져 있는 셈입니다.**




**〈그림 2〉 두 확신이 기대는 서로 다른 단서**

이것은 단순한 짐작이 아닙니다. Steyvers 와 동료들(2025)의 연구에서 사람들은 대규모 언어 모델(LLM)의 정확성을 모델 내부의 실제 확신도보다 높게 평가했고, 설명이 길어질수록 더 강하게 확신했습니다. **설명이 길다고 답이 더 정확해진 것은 아니었는데도 말입니다.** 연구자들은 이 간극을 '보정 격차'라고 불렀고, 설명의 어조를 모델의 실제 확신도에 맞춰 조절하자 사람들의 판단도 실제 정확성에 가까워졌습니다. 즉, ⓐ의 휘둘림은 어떠한 것에 대해서 휘둘리는 성향 – 쉽게 이야기해서 귀가 얇은 사람 – 과 관련된 것이 아닌 판단의 단서에 관련된 문제입니다. 따라서 단서를 바꾸면 판단도 바뀌므로 교사가 개입할 자리가 있습니다. 아직은 희망이 있다는 것이지요!

자동화 연구에서도 사람이 장치의 판단을 확인 없이 따르는 현상을 오래 다루어 왔는데, Parasuraman 과 Manzey(2010)는 이를 게으름이 아니라 주의 배분의 문제로 정리했습니다. 연재 1 회에서 인용한 Lee 와 동료들(2025)의 조사는 두 확신의 구분을 더 직접적으로 뒷받침합니다. 과제에 대한 '자기 확신'이 높으면 더 꼼꼼히 따져 보고, 'AI 에 대한 확신'이 높으면 덜 따져 봅니다. 두 확신이 하나였다면 이런 상반된 결과는 나올 수 없습니다.

그래서 저는 학생에게 확신도를 하나로 묻지 말고 두 칸으로 나누어 묻자고 제안합니다. 그리고 교사가 눈여겨봐야 할 것은 각각의 점수 자체가 아니라 '두 점수의 차이'입니다.

**〈표 2〉 두 확신의 조합과 교실에서의 의미**

| 구분 | AI 답에 대한 확신(ⓐ) 높음 | AI 답에 대한 확신(ⓐ) 낮음 |
|---|---|---|
| 자기 판단에 대한 확신(ⓑ) 높음 | 가장 바람직한 자리. 다만 두 답이 서로 다를 때 무엇을 근거로 고를지 물어야 한다. 바로 여기서 진짜 학습이 일어난다. | 건강한 의심이 작동하는 자리. 왜 AI 의 답이 미덥지 않은지 말해 보게 하면 학생이 알고 있는 제약 조건이 드러난다. |
| 자기 판단에 대한 확신(ⓑ) 낮음 | 가장 위험한 자리. 검증 없이 AI 의 답을 그대로 받아들인다. 실제 교실에서는 이 칸에 몰려 있는 학급이 많다. | 학습이 멈춰 있는 상태. 무엇을 모르는지부터 교사와 함께 정리해야 한다. 적절한 비계(도움 발판)가 필요하다. |

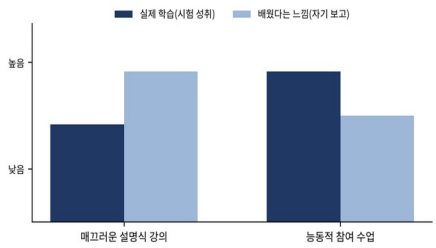
※ 이 표는 학생에게 이름표를 붙여 분류하려는 것이 아니라, 교사가 건넬 '다음 한마디'를 정하기 위한 기준입니다. 같은 학생이라도 문제마다 속하는 칸이 달라집니다.

실제 수업에서는 왼쪽 아래 칸, 곧 내 확신은 낮고 AI 확신은 높은 경우가 가장 많습니다. 어떻게 보면 AI 가 우리 교실로 들어오면서 이런 현상들이 더더욱 두드러지게 나타난다고 볼 수 있는거 같습니다. 2 회에서 제가 했던 실수가 바로 이 칸의 학생들에게 'AI 의 답을 비판적으로 검토해 보라'고 말한 것이었습니다. **자기 판단에 대한 확신이 없으면 AI 의 답을 검토할 근거 자체가 없습니다. 자기가 자신의 판단에 대한 굳건한 확신이 있어야 가능한 검토와 판단도 할 수 있는 것이죠. 그래서 저는 AI 의 답을 보기 전에 반드시 자기 답과 그 확신도를 먼저 적게 합니다. 틀려도 좋으니 내 답이 있어야 AI 의 답과 맞대어 비교할 기준이 생기고, 그렇지 않으면 학생은 AI 의 답을 검토하는 대신 그저 읽게 됩니다.**

나머지 칸도 건넬 말이 다릅니다. 오른쪽 위 칸(내 확신 높음, AI 확신 낮음)은 가장 값진 자리로 왜 그렇게 생각하는지 말해 보게 하면 학생의 지식 구조가 드러납니다. 왼쪽 위 칸(둘 다 높음)은 최종 답만 맞춰 보고 넘어가기 쉬우므로 두 풀이가 어디서부터 달라지는지 찾아보게 합니다. 오른쪽 아래 칸(둘 다 낮음)은 지금 무엇을 모르는지 한 문장으로 적어 보는 것부터 시작합니다.

## 더 많이 배운 학생이 오히려 덜 배웠다고 느낍니다

**Deslauriers와 동료들(2019)은 하버드대학교의 대형 물리 강의에서 같은 내용을 같은 교수가 두 가지 방식으로 가르쳤습니다. 한 집단은 학생이 직접 문제를 풀며 참여하는 능동적 수업을, 다른 집단은 교수가 매끄럽게 준비한 설명식 강의를 들었습니다. 결과는 정반대로 갈렸습니다. 능동적 수업의 학생들은 실제 시험에서 더 많이 배운 것으로 나타났지만 스스로는 덜 배웠다고 느꼈고 설명식 강의의 학생들은 덜 배우고도 더 많이 배웠다고 느꼈습니다(〈그림 3〉). 자 그러니까 이 부분이 주목할 만한 부분입니다. 대부분 능동적인 수업의 효과가 학습자가 더 많이 배우는 환경이라고 생각을 합니다. 실제로도 그렇습니다. 하지만 학습자의 확신의 정도가 반대로 되어 있다는 것이 바로 주목할 만한 부분입니다 연구자들이 지목한 원인은 역시 유창성입니다. 막힘없이 흘러가는 강의는 '잘 이해했다'는 착각을, 자주 막히는 씨름은 '잘 못하고 있다'는 느낌을 만듭니다.**



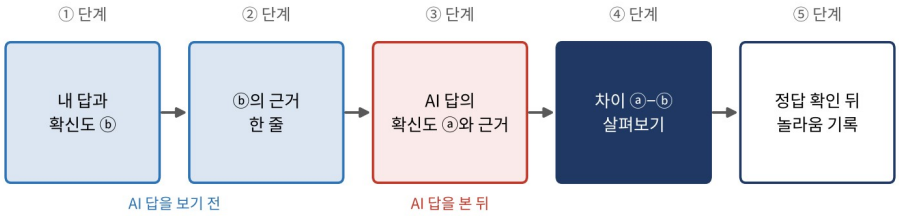

**〈그림 3〉 실제 학습과 배웠다는 느낌의 엇갈림(Deslauriers et al., 2019 개념도)**

Bjork, Dunlosky, Kornell(2013)이 정리한 '학습에 관한 착각'이 강의실 규모로 재현된 셈입니다. 반복해서 읽기는 쉽지만 오래 남지 않고, 책을 덮고 떠올려 보기는 힘들지만 오래 남습니다. 공부하는 동안의 수월함과 실제 학습 효과는 흔히 반대로 움직입니다.

**이 연구가 주는 시사점은 두 가지입니다. 첫째, 확신도라는 숫자를 곧이곧대로 믿을 수 없습니다. ⓑ가 낮은 것은 문제와 치열하게 씨름하고 있어서일 수도 있기 때문에 저는 점수와 함께 그 점수를 준 이유 한 줄을 걷습니다. 둘째, ⓐ가 높은 이유가 문장의 유창성이라면 학생을 탓할 일이 아닙니다. 확신이 저절로 부풀도록 설계된 환경에서 의지로 확신을 조절하라고 요구하는 것은 공평하지 않으므로, 저는 기록지로 두 확신을 물리적으로 갈라놓는 장치를 택했습니다. 이는 교사에게도 마찬가지여서, '잘 이해되었어요'라는 학생의 말 역시 한 번 보정해서 들어야 합니다.**

## 두 칸 확신도를 수업에 적용하는 방법

말로 설명하면 단순해 보이지만 실제 수업에 넣으려면 구체적인 순서와 양식이 필요합니다. 이를 다섯 단계로 정리하면 〈그림 4〉와 같습니다(실제로 사용해본 것은 아닙니다). 그저 제안입니다.




〈그림 4〉 두 칸 확신도 수업의 다섯 단계

**첫째, 내 답과 내 확신도(ⓑ)를 먼저 적습니다. 순서가 바뀌어 AI 의 답을 먼저 보면 ⓑ는 그 답에 영향을 받아 흐려집니다. 한 번 본 AI 의 답은 안 본 상태로 되돌릴 수 없기 때문입니다.**

**둘째, ⓑ 옆에 그렇게 매긴 근거를 한 줄 덧붙입니다. 점수만 적게 하면 학생들은 금방 비슷한 숫자를 기계적으로 쓰기 시작합니다. '나는 무엇 때문에 이만큼 확신한다'는 문장 틀을 함께 주면, 그 확신이 개념 구조에 기대는지, 단순한 기억이나 막연한 느낌에 기대는지가 드러납니다.**

**셋째, AI 의 답을 보고 AI 에 대한 확신도(ⓐ)와 그 근거를 적습니다. 가장 자주 나오는 이유가 '설명이 자세해서', '말이 그럴듯해서'인데, 이것이 바로 앞서 살펴본 유창성 단서입니다. 학생들이 적은 이유를 함께 읽어 보면 어려운 용어를 따로 가르치지 않아도 자신이 왜 흔들렸는지 스스로 알아차리는 경우가 많습니다.**

**넷째, 두 점수의 차이(ⓐ−ⓑ)를 살펴봅니다. 이 값이 클수록 판단의 무게중심이 AI 쪽으로 쏠려 있고, 0 에 가깝거나 음수이면 자기 생각에 중심을 둔 것입니다. 특히 두 답이 다른데도 점수가 AI 쪽으로 크게 기울어 있다면 갈등 없이 AI 의 답을 받아들이려는 상태이므로, 그때가 교사가 개입할 시점입니다.**

**다섯째, 정답을 확인한 뒤 실제 결과가 처음 예상과 같았는지 한 줄로 적습니다. 여기서 가장 눈여겨봐야 할 것은 '확신도가 높았는데 틀린 문제'이며, 그 이유는 다음 절에서 이어서 말씀드리겠습니다.**

**〈표 3〉 두 칸 확신도 기록지 예시**

| 단계 | 내 답과 확신도(ⓑ) | AI 답과 확신도(ⓐ) | 차이(ⓐ−ⓑ) | 정답 확인 뒤 한 줄 |
|---|---|---|---|---|
| 예시 1 | x=3 (2 점) 이항할 때 부호가 헷갈렸다. | x=−3 (4 점) 풀이가 단계별로 자세하다. | +2 (AI 쪽으로 기움) | 정답은 x=3. 설명이 자세하다고 해서 무조건 맞는 것은 아니었다. |
| 예시 2 | 12 (4 점) 넓이 공식을 두 번 확인했다. | 12 (4 점) 내 답과 결과가 같다. | 0 (균형) | 정답. 그런데 AI 는 나와 다른 공식을 썼다. 왜 두 방법이 모두 성립하는지 찾아봐야겠다. |

| 단계 | 내 답과 확신도(ⓑ) | AI 답과 확신도(ⓐ) | 차이(ⓐ−ⓑ) | 정답 확인 뒤 한 줄 |
|---|---|---|---|---|
| 예시 3 | 모르겠음 (1 점) 무엇을 구해야 하는지부터 막막하다. | 5 (2 점) 풀이의 중간 단계가 이해되지 않는다. | +1 (둘 다 낮음) | 구해야 하는 것이 넓이가 아니라 둘레였다. 문제를 정확히 다시 읽는 것부터 시작하자. |

※ 확신도는 1 점부터 4 점까지 매기게 합니다. 어중간한 가운데 점수(3 점)를 없애야 학생이 어느 한쪽으로든 분명한 판단을 내립니다. 표의 예시는 설명을 돕기 위해 구성한 것입니다.

**처음에는 5 점 척도를 생각했는데 가운데 3 점에 몰릴 가능성이 있어 4 점 척도로 바꾸었습니다. ( 실제로 사용해보지는 않았습니다.) 학년이 올라가면 퍼센트로 적게 해도 좋습니다. 단원이 끝나면 확신도 점수별 실제 정답률을 막대그래프로 그려 보는 '보정 곡선' 활동을 합니다. 4 점이라고 적은 문제의 정답률이 2 점보다 오히려 낮은 학생이 생각보다 많은데, 내 확신이 실력과 거꾸로 움직였다는 사실을 눈으로 확인하는 경험은 교사의 백 마디 말보다 오래 남습니다.**

다만 Hacker 와 동료들(2000)의 연구에서 보듯 보정이 가장 절실한 하위권 학생일수록 이런 피드백으로 쉽게 바뀌지 않습니다. 그래서 저는 문제를 풀기 전의 확신도와 풀고 난 직후의 '사후 판단'을 나란히 놓고 비교해 보게 합니다. 직접 부딪친 직후에는 하위권 학생들도 자기 상태를 비교적 정확하게 판단하기 때문입니다.

## 확신하고도 틀린 순간이 가장 값집니다

확신하고도 틀린 경험은 피해야 할 쓰라린 순간처럼 보이지만 오히려 이것은 발전을 위한 아주 중요한 계기가 됩니다. 연구결과로도 이 사실이 뒷받침되고 있습니다. Butterfield 와 Metcalfe(2001)는 강하게 확신했다가 틀린 답을 확신 없이 찍었다가 틀린 답보다 훨씬 더 잘 고친다는 '과교정 효과(hypercorrection effect)'를 밝혀냈고, Metcalfe 와 Finn(2012)은 이 현상이 아동에게서도 나타남을 확인했습니다. 맞다고 믿었는데 틀렸다는 놀라움이 주의를 끌어당겨 그 순간의 정답을 깊이 새기는 것입니다. 우리가 맞은 문제는 잘 기억못하지만 틀린 문제는 간간이 기억이 나는게 이런 상황입니다. 그런데 사람은 틀리고 나면 처음부터 자신이 없었던 것처럼 기억을 고쳐 쓰기 쉽습니다. 미리 종이에 '4 점'이라고 적어 둔 숫자가 이 유익한 놀라움을 붙잡아 두는 셈이니, 이것이 확신도를 미리 적게 하는 또 하나의 이유입니다.

**그래서 저는 확신하고 틀린 문제를 부끄러운 실수가 아니라 학급 전체가 함께 들여다보는 귀한 배움의 자료로 씁니다. 물론 이는 학급의 심리적 안전감이 뒷받침될 때만 가능합니다. 그리고 틀린 문제를 허용하는 교실의 규범도 필요합니다. 이 규범이 만들어졌을 때 틀렸다는 사실을 드러낼 수 있습니다.**

**틀렸다는 사실을 드러내는 것이 안전하지 않은 교실에서는 학생들이 상처받지 않으려고 처음부터 확신도를 낮게 적기 시작합니다. 일종의 방어기제를 작동하게 되는 것이죠.**

## 이 요건이 빠지면 교실에서 어떤 모습이 보이는가

확신도 요건이 약한 학생들에게서 반복적으로 나타나는 다섯 가지 신호와, 그때 제가 건네는 '다음 한마디'를 정리하면 〈표 4〉와 같습니다. 이 가운데 교사가 가장 놓치기 쉬운 것은 넷째 신호입니다. AI의 답을 골라 맞혔다면 겉으로는 문제가 없어 보이지만, 갈등 없이 베낀 학생은 다음번에 AI가 그럴듯한 오답을 내놓을 때 그대로 따라 틀리기 쉽습니다.

**〈표 4〉 확신도 요건이 약할 때 나타나는 신호와 교사의 다음 한마디**

| 교실에서 보이는 신호 | 무엇이 빠져 있는가 | 교사의 다음 한마디 |
|---|---|---|
| 모든 문제에 똑같은 확신을 보임 | 문제 간 변별 | '이 중에서 제일 불안한 문제 하나만 골라 볼래?' |
| 틀리고도 전혀 놀라지 않음 | 확신의 자각 | '풀기 전에 몇 점짜리 확신이라고 생각했는지 기억나?' |
| 확신도에 늘 가운데 점수만 적음 | 판단 자체를 미룸 | '맞는 쪽에 가까워, 틀리는 쪽에 가까워? 둘 중 하나만 골라 보자.' |
| 두 답이 다른데 갈등 없이 AI 선택 | 내 판단(ⓑ)의 근거 | 'AI 답은 잠시 가리고, 네 답의 근거를 한 줄만 적어 보자.' |
| 확신 정도를 물으면 자기 답만 반복함 | 확신과 답의 구분 | '답 말고, 그 답에 내기를 건다면 얼마나 걸 수 있겠어?' |

## 수학 교실이라서 유리한 점

확신도를 다루기에 수학 교실은 다른 교과보다 꽤 유리한 편입니다. 구한 해를 원래 식에 다시 대입하는 검산, 극단적인 값을 넣어 보는 확인, 답의 크기를 어림하거나 단위를 따져 보는 일처럼 막연한 확신을 정답지 없이 시험해 볼 도구가 여럿 있기 때문입니다. 그래서 저는 확신도와 검산을 한 묶음으로 가르칩니다. **확신도를 낮게 적은 문제일수록 검산을 먼저 하게 하고, 내 확신도도 높은데 AI의 답과 다른 문제는 두 답을 모두 원래 식에 대입해 보게 합니다. 이렇게 하면 확신도 칸은 기분을 적는 칸이 아니라 무엇부터 검증할지 순서를 정하는 칸이 됩니다.**

그래서 저는 수업에서 '확신도'와 '검산'을 한 묶음으로 엮어 가르칩니다. 확신도를 낮게 적은 문제일수록 검산을 가장 먼저 해 보게 하고, 내 확신도도 높은데 AI의 답과 다른 문제는 두 답을 모두 원래 식에 대입해 보게 합니다. 이렇게 연결하면 확신도를 적는 칸은 단순한 기분을 적는 칸이 아니라, '무엇부터 검증할지 순서를 정하는 칸'이 됩니다. 앞서 확신도가 나머지 요건들에 우선순위를 매기는

요건이라고 말씀드렸는데, 수학 교실에서는 그 우선순위가 이처럼 구체적인 검산 행동으로 나타납니다.

또 수학의 풀이는 한 줄 한 줄 단계로 나뉘므로, 단계별로 끊어서 물으면 학생들은 '식을 세운 단계는 자신 있는데 이항한 단계가 불안하다'처럼 훨씬 정확하게 답합니다. AI 의 풀이도 '몇 번째 줄부터 확신이 떨어지는지'를 묻게 하면 판단이 날카로워집니다. 다만 이유를 설명하거나 증명하는 문제에서는 확신도를 정답 확률로만 좁게 보지 않도록 '이 풀이를 짝에게 설명해 납득시킬 만큼 확신하는가?'라고 묻습니다.

## 남는 문제 — 이번 회의 연구 근거가 가장 얇습니다

이번 회차는 연재 전체에서 가장 새로운 제안을 담은 만큼 학술적 근거도 가장 얇다는 점을 솔직하게 적어 두어야겠습니다. 첫째, 확신도를 두 칸으로 나누어 적게 하는 수업 방식의 효과를 직접 검증한 선행 연구를 저는 아직 찾지 못했습니다. 아마도 추후에 제가 연구해야 할까요? Lee 와 동료들(2025)은 성인 지식노동자 대상 설문이었고 Steyvers 와 동료들(2025)은 성인이 객관식 문제에 대한 AI 의 답변을 판단하는 실험이었습니다. 수학 서술형 풀이처럼 호흡이 긴 과제와 청소년 학습자에게서도 같은 효과가 나타날지는 앞으로 확인해야 할 열린 질문입니다. 둘째, 확신도를 매번 묻다 보면 확신 자체가 위축될 수 있습니다. 셋째, 숫자를 매기는 행위가 깊은 판단을 대신해 버려 기록 칸이 형식적인 절차가 될 위험이 있습니다.

넷째, 학생들이 확신도 기록을 평가로 오해하면 실제 확신이 아니라 교사가 보고 싶어 할 것 같은 점수를 적습니다. 그래서 학기 초에 어떤 평가에도 반영하지 않는다고 분명히 밝히지만, 오해가 단번에 사라지지는 않습니다.

다섯째, 과신을 무조건 고쳐야 할 대상으로만 볼 수 있는지도 조심스럽습니다. 약간의 근거 없는 낙관은 어려운 문제 앞에서 버티게 해 주는 힘이 되기도 합니다. 제가 보정보다 변별을 먼저 다루자고 한 이유도 여기에 있습니다. 자신감의 크기는 건드리지 않더라도 어느 문제가 유독 위험한지 가려내는 눈은 키워 줄 수 있습니다.

이러한 한계에도 이 요건을 빼놓지 않는 이유는 하나입니다. 무엇을 검증할지 정하려면 지금 내 생각의 어디가 불안한지를 먼저 알아차려야 하는데, 확신도 말고는 그 지점을 비춰 주는 도구를 제가 아직 찾지 못했기 때문입니다.

## 다음 회로 넘어가며

**이번 회의 이야기를 한 문장으로 줄이면 이렇습니다. AI 와 함께 배우는 교실에서 확신은 하나가 아니라 둘이며, 교사가 눈여겨봐야 할 것은 각 확신의 크기가 아니라 '두 확신 사이의 간격'입니다. '맞는 것 같아서요'라는 대답은 그 두 확신과 간격이 하나로 뭉개진 상태를 보여 주는 말이었고, 확신도를 두 칸으로 나누어 묻는 일은 뭉개져 있던 감각을 다시 또렷하게 펼쳐 보이는 작업입니다.**

**물론 확신도는 그 자체로 완결되는 요건이 아닙니다. 확신도가 검증을 어디에 걸어야 할지 정해 주는 나침반이라면, 다음 회에서 다룰 여섯 번째이자 마지막 요건인 '결과'는 실제로 어떻게 검증하고 그 결과를 자기 지식 구조에 어떻게 되먹일지를 다룹니다. 아울러 3 부를 마무리하면서 여섯 요건이 왜 낱개의 목록이 아니라 하나로 맞물려 돌아가는 묶음인지도 함께 정리하겠습니다.**

## 생각해 볼 문제

| 구분 | 생각해 볼 문제 | 생각의 실마리 |
|---|---|---|
| 문제 1 | 우리 반 학생이 '맞는 것 같아서요'라고 말할 때, 그 확신은 자기 지식 구조에 기대고 있을까요, 아니면 AI 문장의 유창성에 기대고 있을까요? 수업 중에 이 둘을 가려낼 질문 하나를 만든다면 무엇일까요? | 〈그림 2〉와 〈표 4〉의 '다음 한마디'를 참고하되, 내 교과와 학생 수준에 맞는 말로 바꾸어 봅니다. |
| 문제 2 | 두 칸 기록지를 도입한 뒤 모든 문제에 2 점만 적는 학생이나 확신도를 평가로 오해하는 학생이 생긴다면, 기록지 양식이나 학급의 약속을 어떻게 바꿀 수 있을까요? | 심리적 안전감, 평가에 반영하지 않는다는 원칙, 근거 한 줄 쓰기가 각각 어떤 역할을 하는지 함께 따져 봅니다. |
| 문제 3 | 이 글은 AI 의 답을 보기 전에 내 답과 확신도를 먼저 적게 합니다. 반대로 AI 의 답을 먼저 보여 주는 편이 나은 과제나 학생도 있을까요? 있다면 그 기준은 무엇일까요? | 자기 판단을 세울 최소한의 배경지식이 없을 때 이 순서가 학생에게 어떤 어려움이 되는지 떠올려 봅니다. |

※ 정해진 답이 있는 문제가 아닙니다. 연수나 협의회에서 동료 교사와 함께 이야기 나누는 출발점으로 활용해 주세요.

## 참고 문헌


Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. Personality and Social Psychology Review, 13(3), 219–235.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(6), 1491–1494.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences, 116(39), 19251–19257.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. Learning and Instruction, 22(4), 271–280.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 3(4), 552–564.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. Journal of Educational Psychology, 92(1), 160–170.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126(4), 349–370.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. Psychonomic Bulletin & Review, 15(1), 174–179.

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. Learning and Instruction, 22(4), 253–261.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. Psychology of Learning and Motivation, 26, 125–173.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, 4(1), 33–45.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7, 221–231.

## 부록 1. 두 칸 확신도 기록지

반드시 ①②단계를 마친 뒤에 AI 의 답을 확인합니다. 이 기록지는 어떤 평가에도 반영하지 않습니다.

| 학년 반 번호 | 학년 반 번 | 이름 | |
|---|---|---|---|
| 단원 | | 날짜 | 월 일 |

| 확신도 점수 기준 (해당하는 점수에 ○ 표시) | | | |
|---|---|---|---|
| 1 점 거의 모르겠다 | 2 점 자신 없는 편이다 | 3 점 자신 있는 편이다 | 4 점 확실하다 |

| 문항 | 단계 | 답 | 확신도 | 근거 한 줄 |
|---|---|---|---|---|
| 1 | ①② 내 판단 ⓑ | | 1 2 3 4 | 나는 ___ 때문에 이만큼 확신한다. |
| | ③ AI 판단 ⓐ | | 1 2 3 4 | AI 답을 ___ 때문에 이만큼 믿는다. |
| | ④ 차이 ⓐ–ⓑ | ( ) AI 쪽 / 균형 / 내 쪽 | **두 답이 같은가?** | 같다 / 다르다 |
| | ⑤ 정답 확인 | 정답: 예상과 같았는가? 같았다 / 달랐다<br>놀란 점이나 새로 알게 된 점: | | |
| 2 | ①② 내 판단 ⓑ | | 1 2 3 4 | 나는 ___ 때문에 이만큼 확신한다. |
| | ③ AI 판단 ⓐ | | 1 2 3 4 | AI 답을 ___ 때문에 이만큼 믿는다. |
| | ④ 차이 ⓐ–ⓑ | ( ) AI 쪽 / 균형 / 내 쪽 | **두 답이 같은가?** | 같다 / 다르다 |
| | ⑤ 정답 확인 | 정답: 예상과 같았는가? 같았다 / 달랐다<br>놀란 점이나 새로 알게 된 점: | | |
| 3 | ①② 내 판단 ⓑ | | 1 2 3 4 | 나는 ___ 때문에 이만큼 확신한다. |
| | ③ AI 판단 ⓐ | | 1 2 3 4 | AI 답을 ___ 때문에 이만큼 믿는다. |
| | ④ 차이 ⓐ–ⓑ | ( ) AI 쪽 / 균형 / 내 쪽 | **두 답이 같은가?** | 같다 / 다르다 |
| | ⑤ 정답 확인 | 정답: 예상과 같았는가? 같았다 / 달랐다<br>놀란 점이나 새로 알게 된 점: | | |

## 부록 2. 단원 정리: 나의 보정 곡선

한 단원 동안 쓴 기록지를 모아 내 확신도(ⓑ) 점수별로 문항 수와 맞힌 문항 수를 세어 봅니다. 정답률을 막대그래프로 그렸을 때 점수가 높을수록 막대도 높아지면 보정이 잘된 것입니다.

| 학년 반 번호 | 학년 반 번 | 이름 | |
|---|---|---|---|
| 단원 | | 날짜 | 월 일 |

| 내 확신도(ⓑ) | 1 점 | 2 점 | 3 점 | 4 점 |
|---|---|---|---|---|
| 문항 수 | | | | |
| 맞힌 문항 수 | | | | |
| 정답률(%) | | | | |

| 정답률 막대그래프 | | | | |
|---|---|---|---|---|
| 100% | | | | |
| 75% | | | | |
| 50% | | | | |
| 25% | | | | |
| 확신도 | 1 점 | 2 점 | 3 점 | 4 점 |

| 돌아보기 | 확신이 높았는데 틀린 문항은 무엇이었고, 그 문항에서 무엇을 배웠는가? |
|---|---|
| 다음 단원 다짐 | 다음 단원에서 내 확신도를 매길 때 꼭 확인할 것 한 가지: |

## 부록 3. 기록지 작성 예시

본문 〈표 3〉의 세 예시를 기록지에 옮겨 적은 모습입니다. 파란 글씨가 학생이 쓰는 부분이며, 매긴 점수는 동그라미 숫자로 표시했습니다.

| 문항 | 단계 | 답 | 확신도 | 근거 한 줄 |
|---|---|---|---|---|
| 예시 1 | ①② 내 판단 ⓑ | x = 3 | 1 ② 3 4 | 이항할 때 부호가 헷갈려서 |
| | ③ AI 판단 ⓐ | x = −3 | 1 2 3 ④ | 풀이가 단계별로 자세해서 |
| | ④ 차이 ⓐ−ⓑ | ( +2 ) AI 쪽 | **두 답이 같은가?** | 다르다 |
| | ⑤ 정답 확인 | 정답: x = 3 예상과 같았는가? 달랐다<br>설명이 자세하다고 해서 무조건 맞는 것은 아니었다. | | |
| 예시 2 | ①② 내 판단 ⓑ | 12 | 1 2 3 ④ | 넓이 공식을 두 번 확인해서 |
| | ③ AI 판단 ⓐ | 12 | 1 2 3 ④ | 내 답과 결과가 같아서 |
| | ④ 차이 ⓐ−ⓑ | ( 0 ) 균형 | **두 답이 같은가?** | 같다 |
| | ⑤ 정답 확인 | 정답: 12 예상과 같았는가? 같았다<br>AI 는 나와 다른 공식을 썼다. 두 방법이 모두 성립하는 이유를 찾아봐야겠다. | | |
| 예시 3 | ①② 내 판단 ⓑ | 모르겠음 | ① 2 3 4 | 무엇을 구할지부터 막막해서 |
| | ③ AI 판단 ⓐ | 5 | 1 ② 3 4 | 풀이의 중간 단계가 이해되지 않아서 |
| | ④ 차이 ⓐ−ⓑ | ( +1 ) 둘 다 낮음 | **두 답이 같은가?** | 다르다 |
| | ⑤ 정답 확인 | 예상과 같았는가? 달랐다<br>구해야 하는 것이 넓이가 아니라 둘레였다. 문제를 다시 읽는 것부터 시작하자. | | |

※ 예시는 설명을 돕기 위해 구성한 것입니다. 예시 1 처럼 두 답이 다른데 차이가 AI 쪽으로 크게 기울어 있다면 교사가 먼저 다가갈 신호입니다.