Artificial Intelligence, Teacher Tasks and Individualized Pedagogy

Bruno Ferman (EESP-FGV)
Lycia Lima (FGV-CLEAR)
Flavio Riva (EAESP-FGV)

March 22, 2021

Roadmap

- **Artificial Intelligence (AI)**...
  - shifts limits of which **tasks** can be automated;

- **Artificial Intelligence (AI)**...
  - shifts limits of which **tasks** can be automated;
  - re-allocates tasks between **human labor** and **technology** in labor markets;

- **Artificial Intelligence (AI)**...
  - shifts limits of which **tasks** can be automated;
  - re-allocates tasks between **human labor** and **technology** in labor markets;
  - revives debates on what **should** be automated.

- **Artificial Intelligence (AI)**...
  - shifts limits of which **tasks** can be automated;
  - re-allocates tasks between **human labor** and **technology** in labor markets;
  - revives debates on what **should** be automated.
- In education, ongoing controversy on **automated writing evaluation (AWE)** systems...
  1. natural language processing;
  2. machine learning algorithms.

- Supporters of AWE argue that...
  - AWE systems may relax teachers' **time constraints**.
  - They may also help **human capital constraints**.
- Critics...
  - AWE is "*blind to meaning*" and **cannot emulate human behavior**.

1. How are these *ed* techs **incorporated into instruction**?
2. What are the effects on **students' outcomes** that proxy for learning?

- We study **two** *ed* techs designed to improve scores in the **argumentative essay** of the National Secondary Education Exam (ENEM).

- We study **two** *ed* techs designed to improve scores in the **argumentative essay** of the National Secondary Education Exam (ENEM).

- Focus on bottlenecks to effective pedagogy in **public schools**.

- We study **two** *ed* techs designed to improve scores in the **argumentative essay** of the National Secondary Education Exam (ENEM).
- Focus on bottlenecks to effective pedagogy in **public schools**.
- They use different combinations of **artificial** and **human intelligence**:
  1. Pure AWE *ed* tech: standard AWE system;
  2. Enhanced AWE *ed* tech: timely "supervision" step by human intelligence.

Roadmap

- Key determinant of access to **post-secondary education (PSE)** in Brazil:

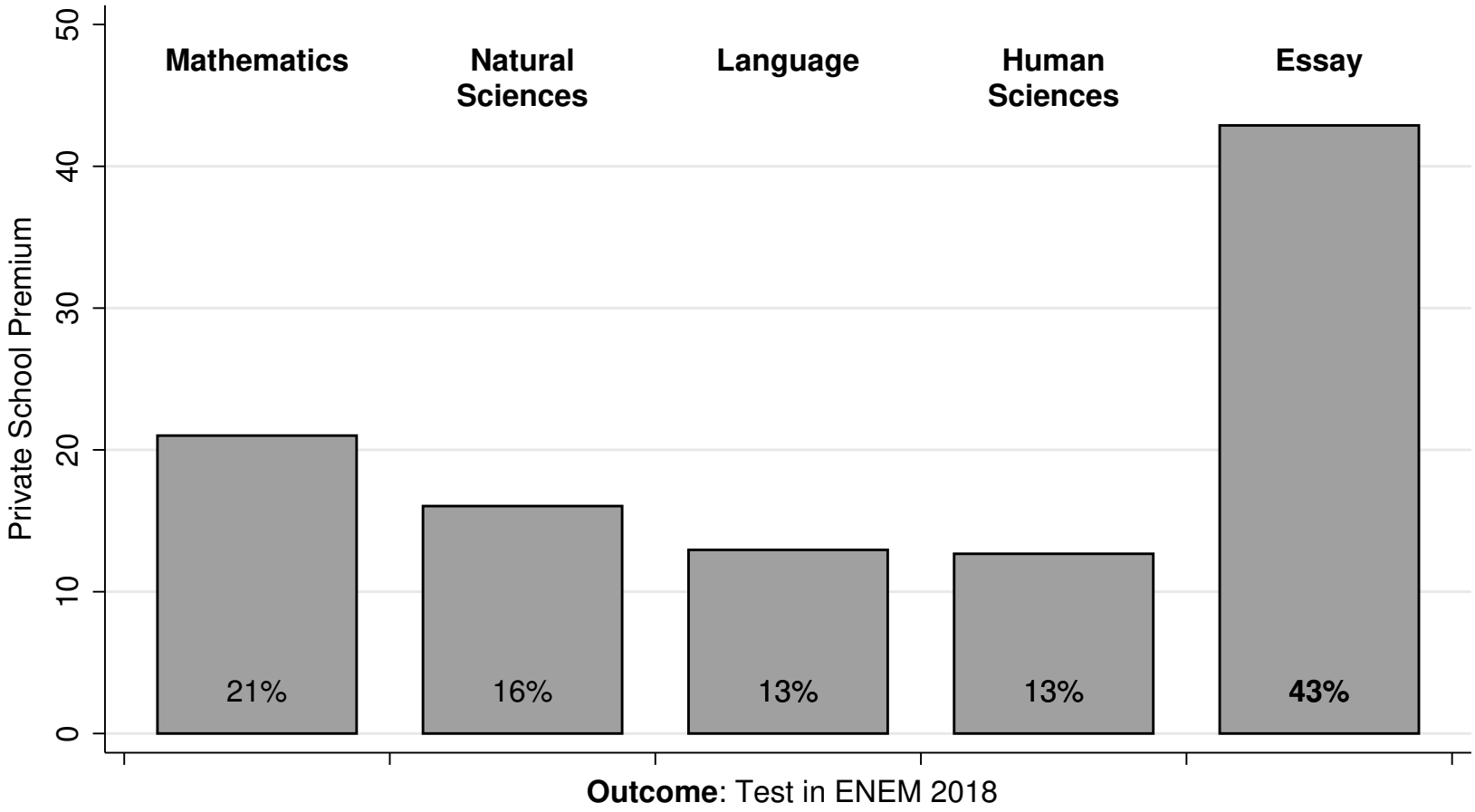
,

- Key determinant of access to **post-secondary education (PSE)** in Brazil:

  ,

- Large differences between quality in **public** and **private schools** is apparent.

Figure: Private School Premium is Particularly Large in the Essay



Essay Topic — 2019

"The Democratization of Access to Cinema in Brazil"

Syntactic Skills

Formal written norm + and a "fluid" text built on **argumentative connectives** within and across paragraphs.

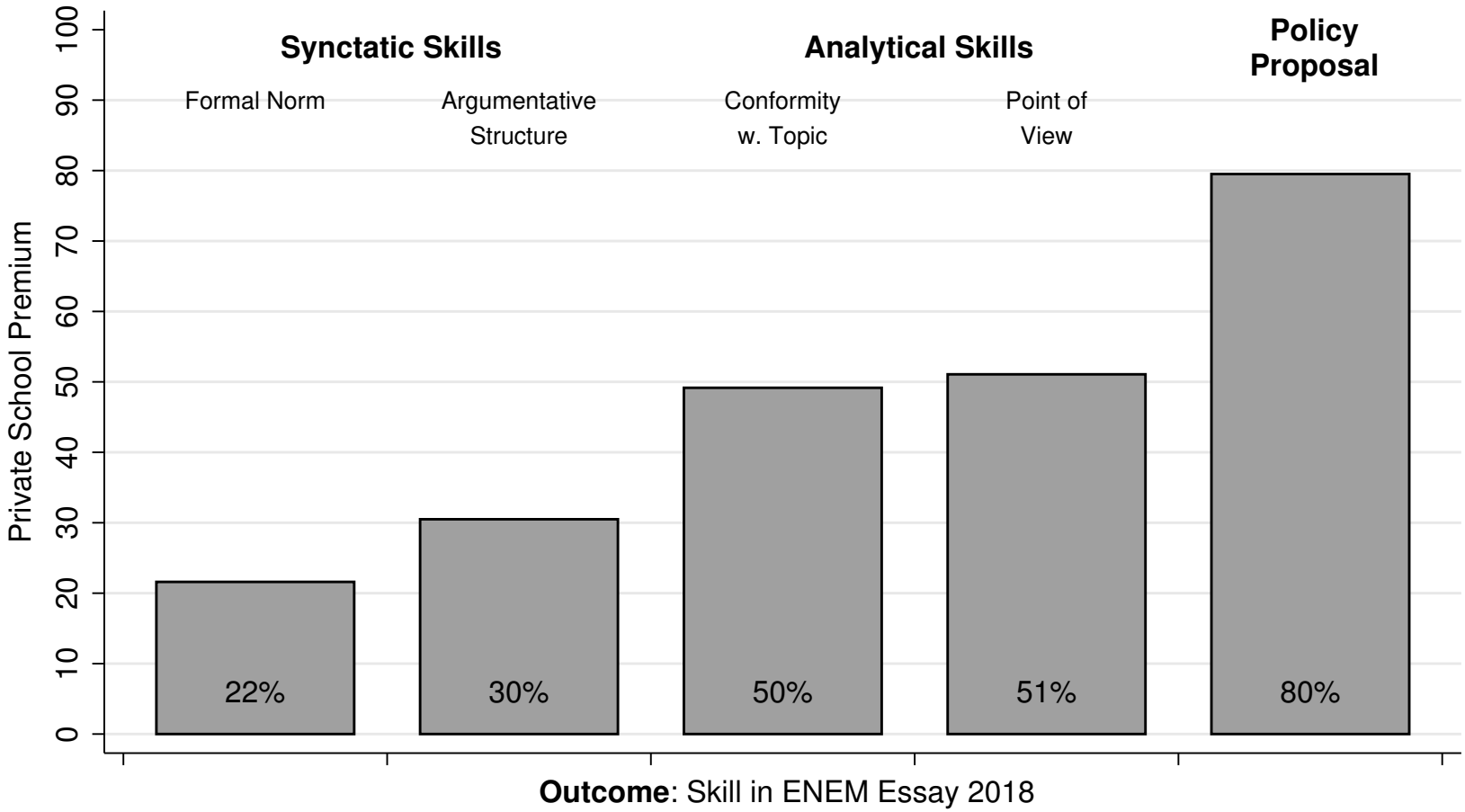
Analytical Skills

Ability to interpret and use information from the motivating elements + from knowledge acquired throughout the schooling process.

Policy Proposal Skills

Intervention consistent with the thesis developed in the essay.

Figure: Private School Premium is Increasing in Skill Sophistication



1. Broad range of writing skills captured in ENEM.

1. Broad range of writing skills captured in ENEM.
2. Large differences between public and private students.

1. Broad range of writing skills captured in ENEM.
2. Large differences between public and private students.
3. Large potential number of beneficiaries from the *ed* techs.

- **Field experiment** in Espirito Santo, Brazil, in 2019, 178 schools (app. 19,000 students):
  1. **Pure** AWE arm: 55 schools
  2. **Enhanced** AWE arm: 55 schools
  3. Control Arm: remaining 68 schools

- **Field experiment** in Espirito Santo, Brazil, in 2019, 178 schools (app. 19,000 students):
  1. **Pure** AWE arm: 55 schools
  2. **Enhanced** AWE arm: 55 schools
  3. Control Arm: remaining 68 schools
- 5 writing activities (ENEM training essays) throughout the year.

Roadmap

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Estimation

We fit the following regression to study effects on the primary outcome:

$$Y_{ise} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \varepsilon_{ise}$$

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Estimation

We fit the following regression to study effects on the primary outcome:

$$Y_{ise} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \varepsilon_{ise}$$

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Estimation

We fit the following regression to study effects on the primary outcome:

$$Y_{ise} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \varepsilon_{ise}$$

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Estimation

We fit the following regression to study effects on the primary outcome:

$$Y_{ise} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \varepsilon_{ise}$$

Primary Outcome — Did AWE Systems Improve ENEM Writing Scores?

i. Official ENEM 2019 essay scores (in total, and *per* skill);

ii. "Unofficial" ENEM 2019 essay scores (in total, and *per* skill).

Estimation

We fit the following regression to study effects on the primary outcome:

$$Y_{ise} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \varepsilon_{ise}$$

Estimation

The models for mechanisms and secondary outcomes are similar:

$$Y_{is} = \tau_{\text{ITT}}^{\text{Enhanced}} W_s^{\text{Enhanced}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}_{is}' \mathbf{\Pi} + \xi_{is}$$

Inference

We present p-values:

1. based on standard errors clustered at the strata level — **??**
2. based on randomization inference using the protocol and 1,000 placebos
3. adjusted for multiple hypothesis testing — **?**

and summary indexes based on **?**.

Roadmap

Figure: **95% + of Teachers** Used the Platform to Assign Essays

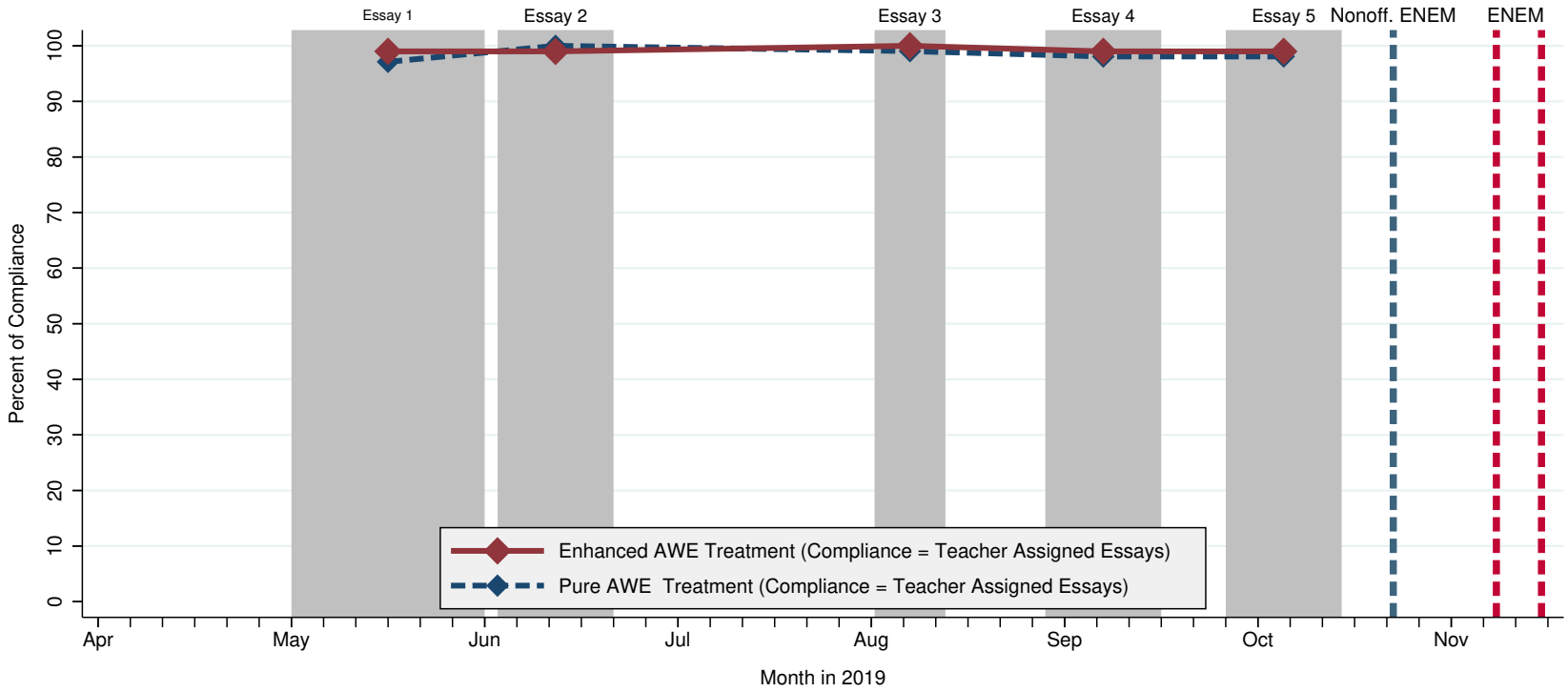


Figure: **75-80% of Students** Used the Platform to Submit Essays

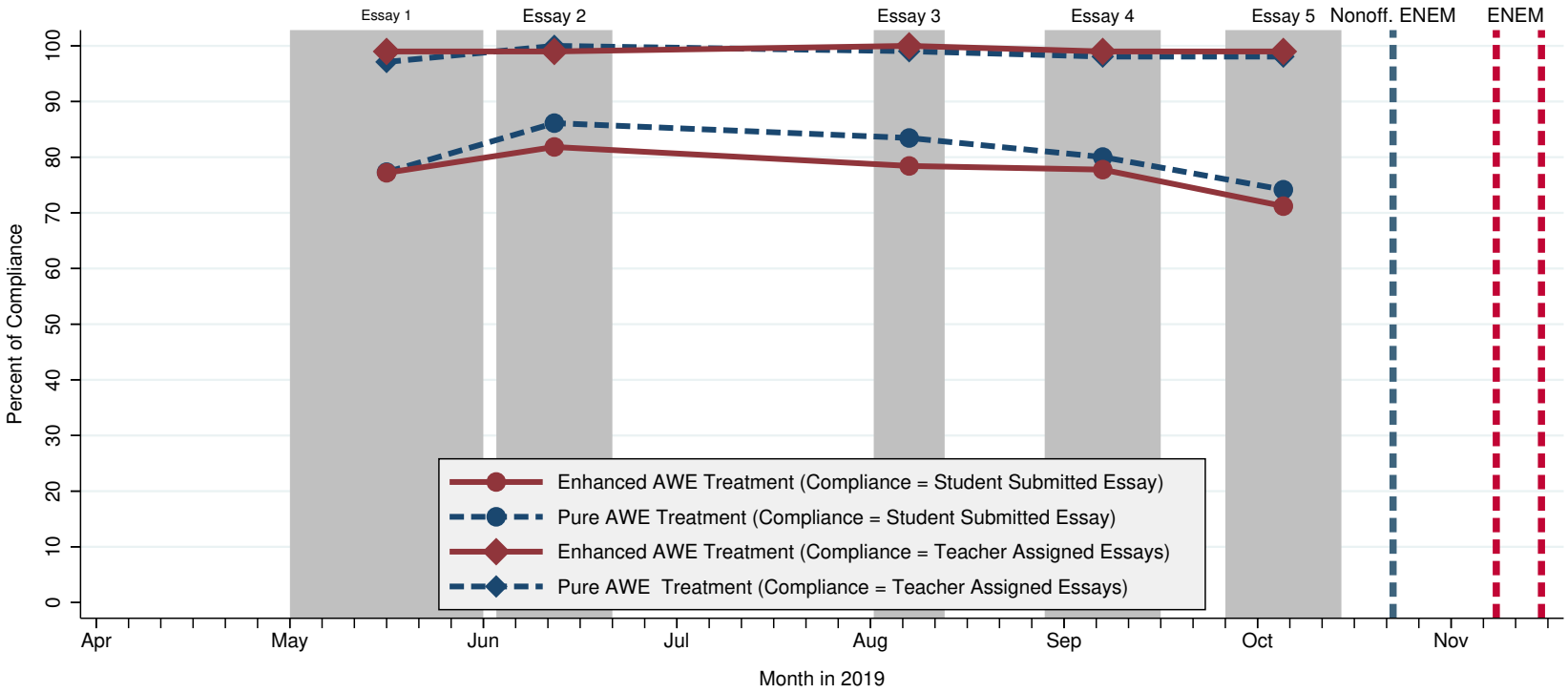


Figure: ITT Effects on ENEM Writing Scores

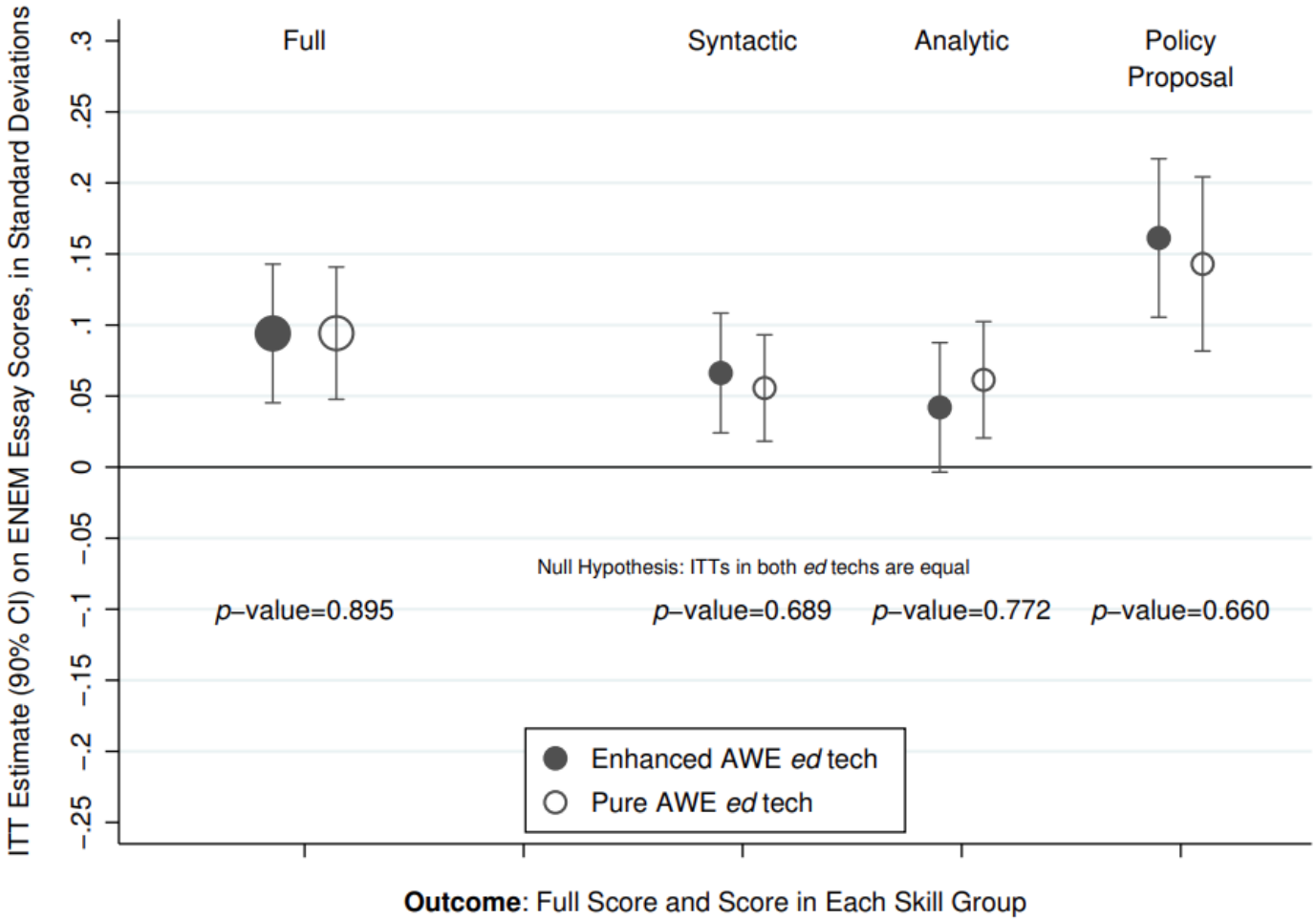


Figure: *ed* techs Had **Positive Impacts on the Full Score**

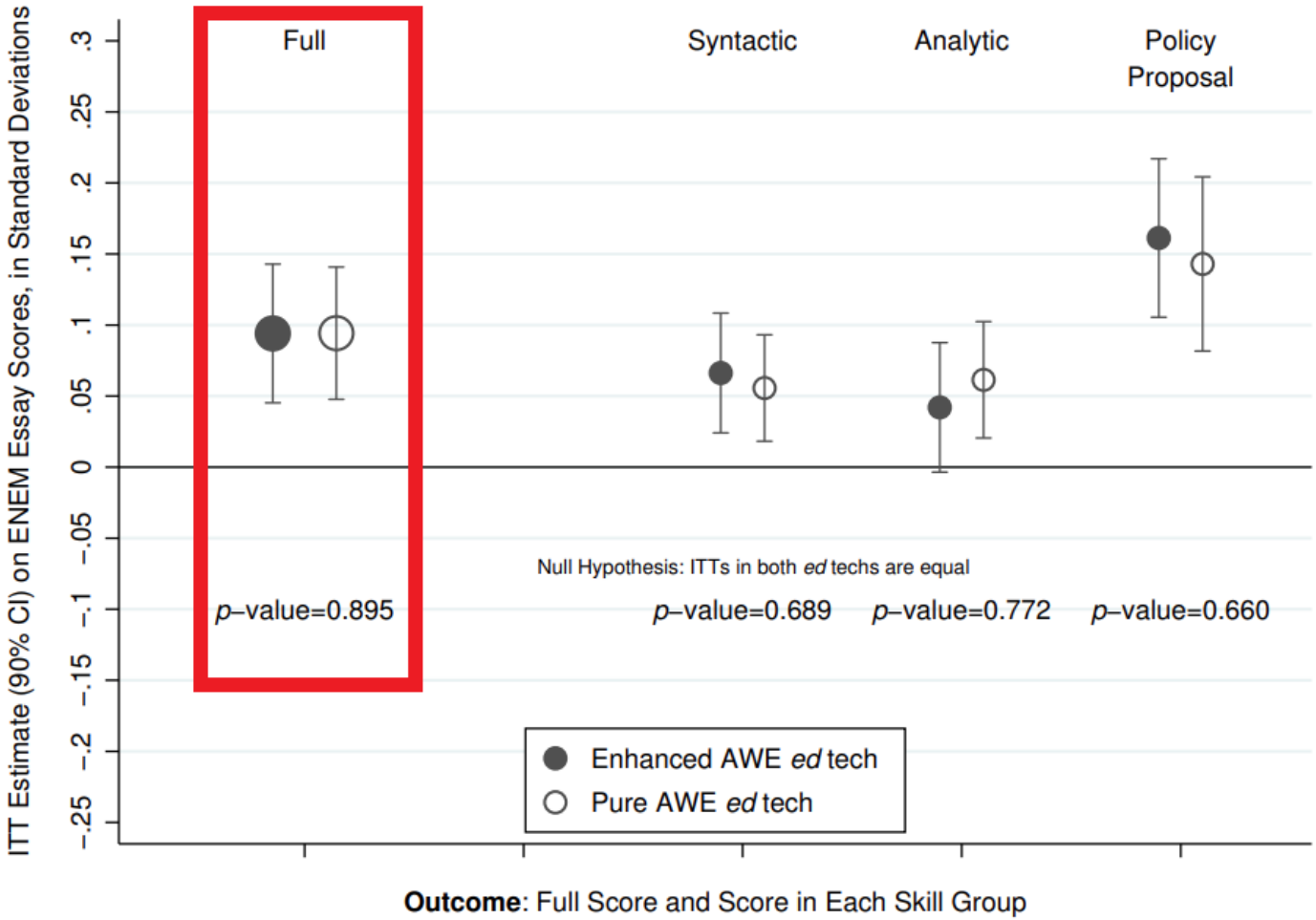


Figure: *ed* techs Had **Very Similar** Impacts on the Full Score

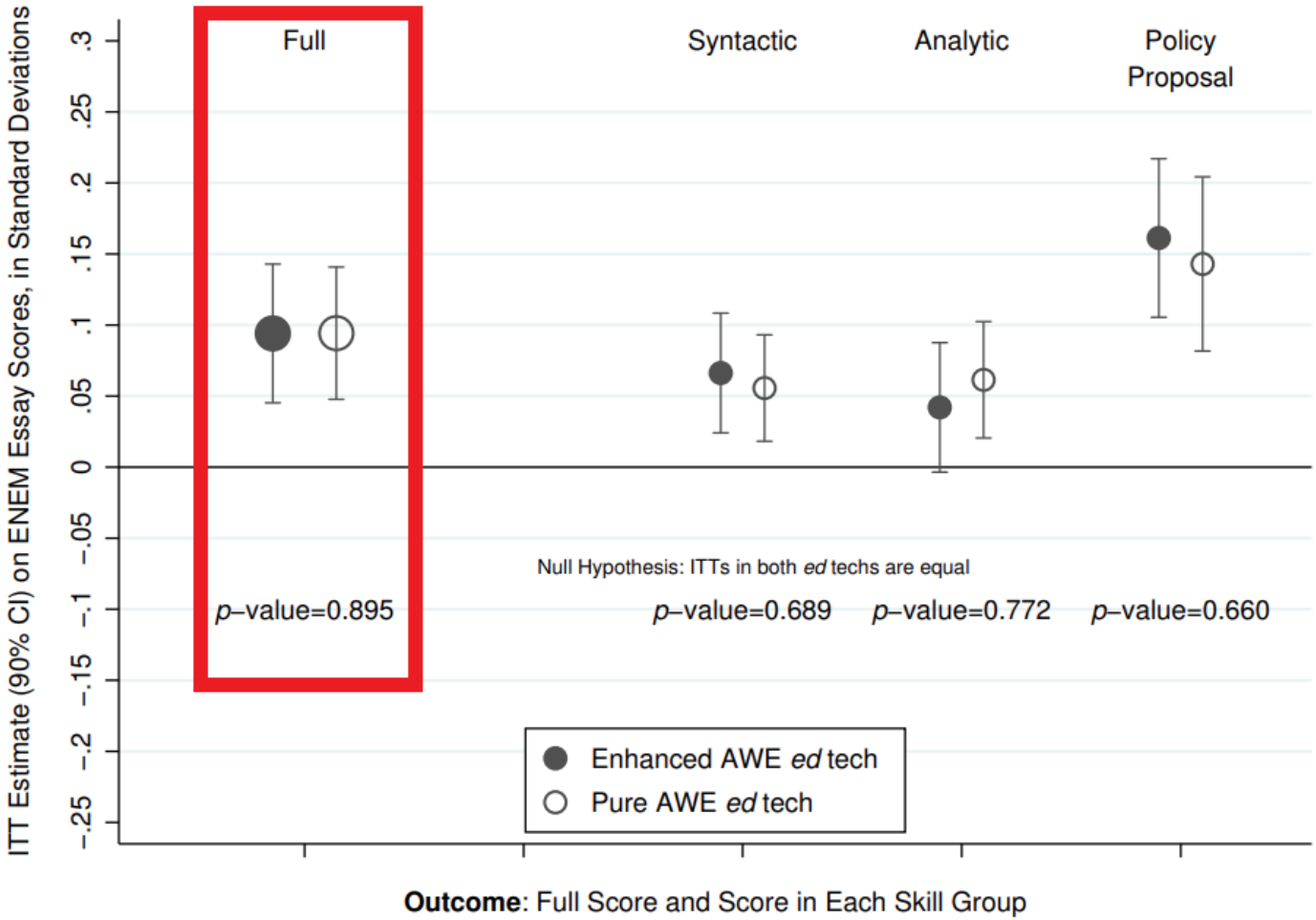


Figure: Effects Were Channeled By **Very Similar Effects on All Scores**

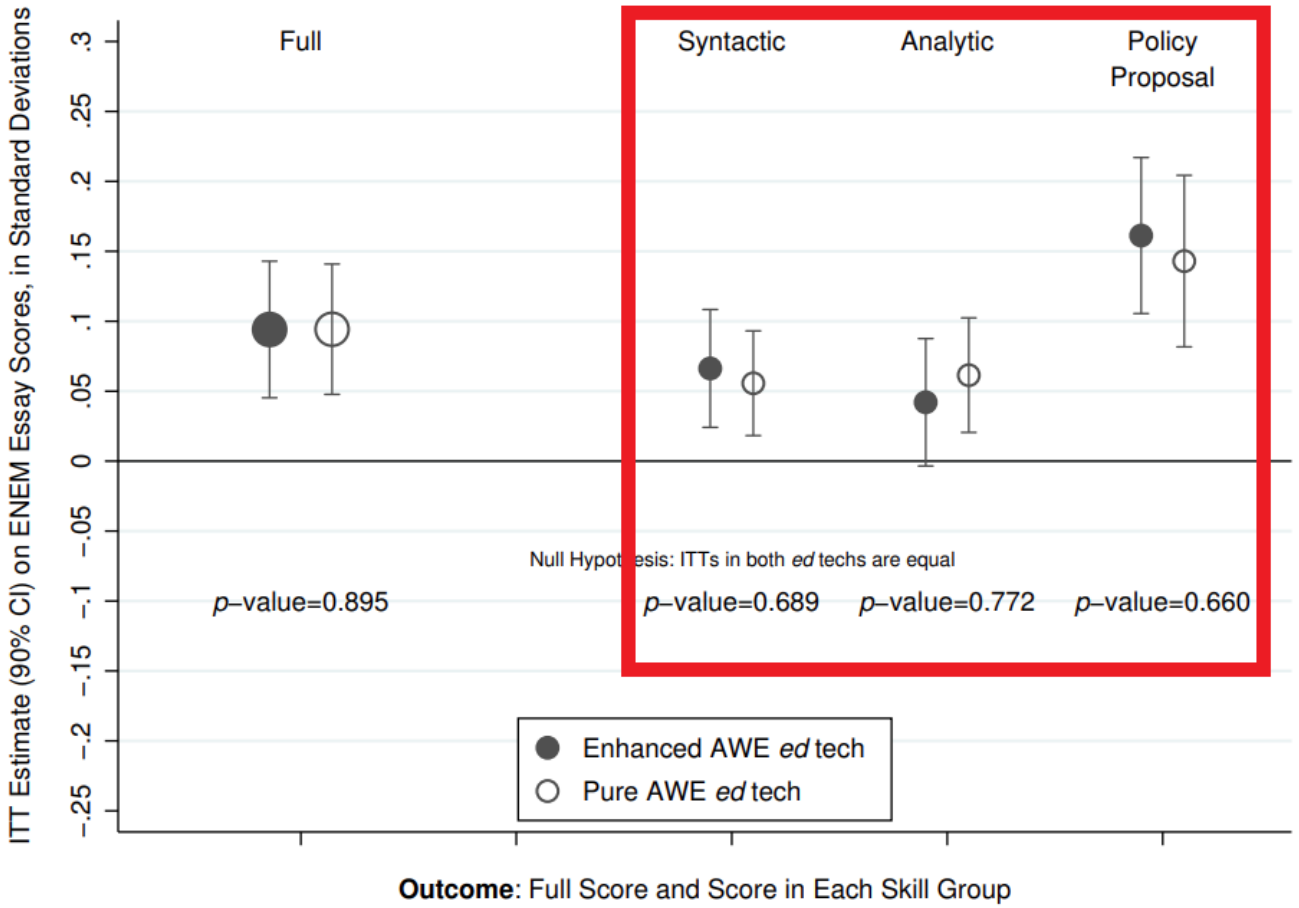


Figure: ITT Effects on Training, Feedback and Individualized Pedagogy

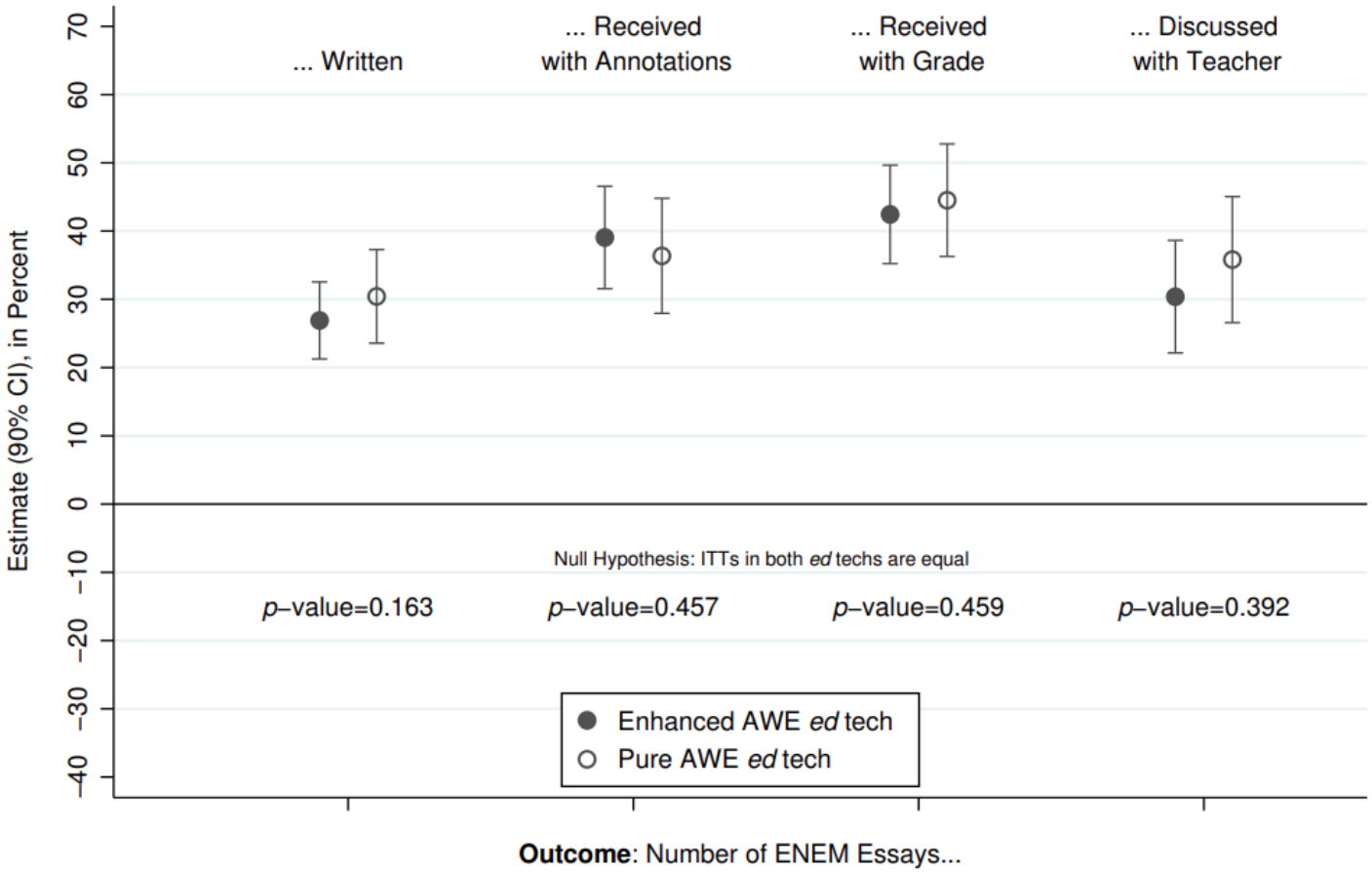


Figure: *ed* techs Had **Very Similar Impacts on Training and Feedback**

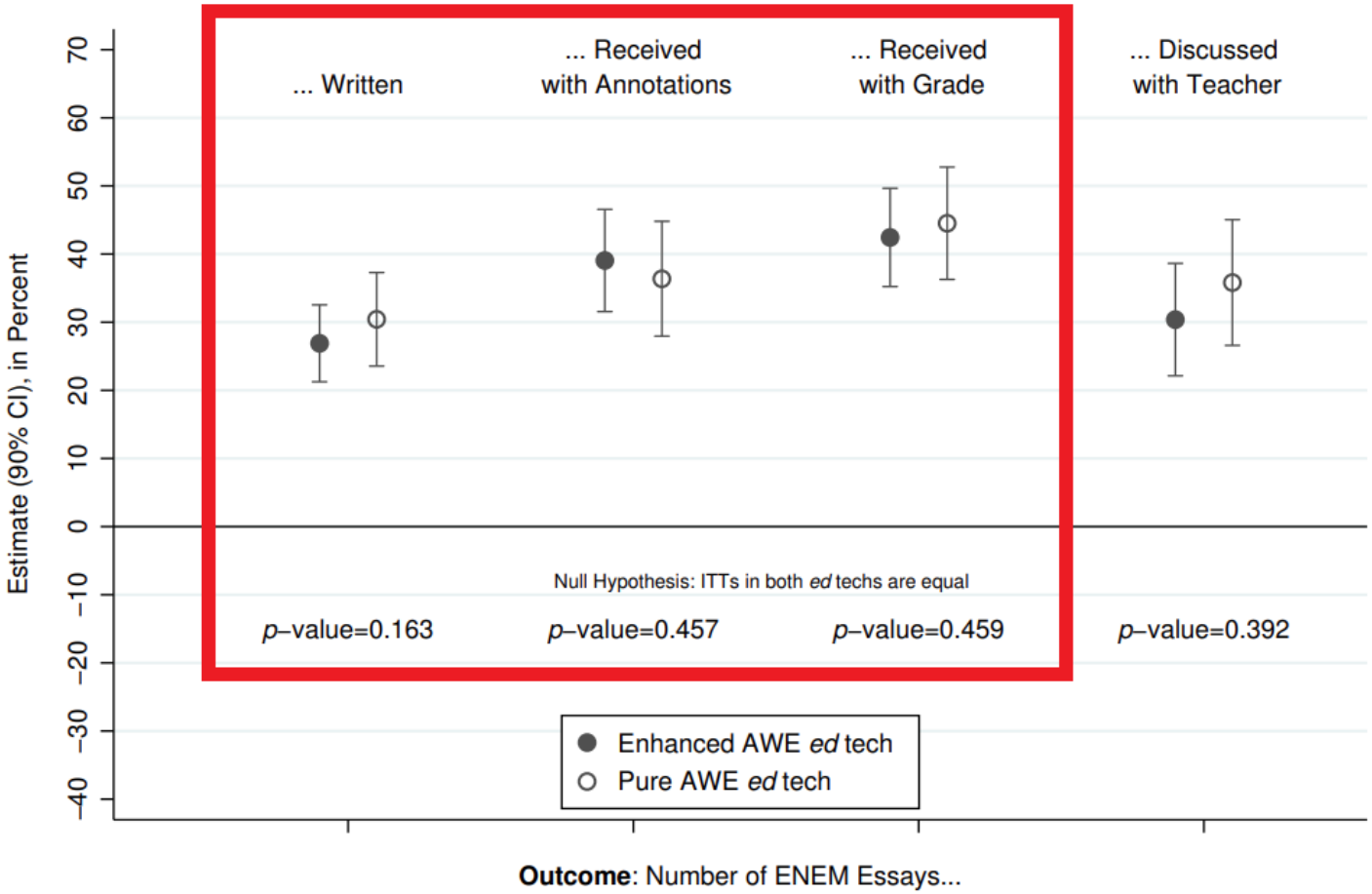


Figure: *ed* techs Induced **Very Similar Shifts to Nonroutine Tasks That Supported the Individualization of Pedagogy**

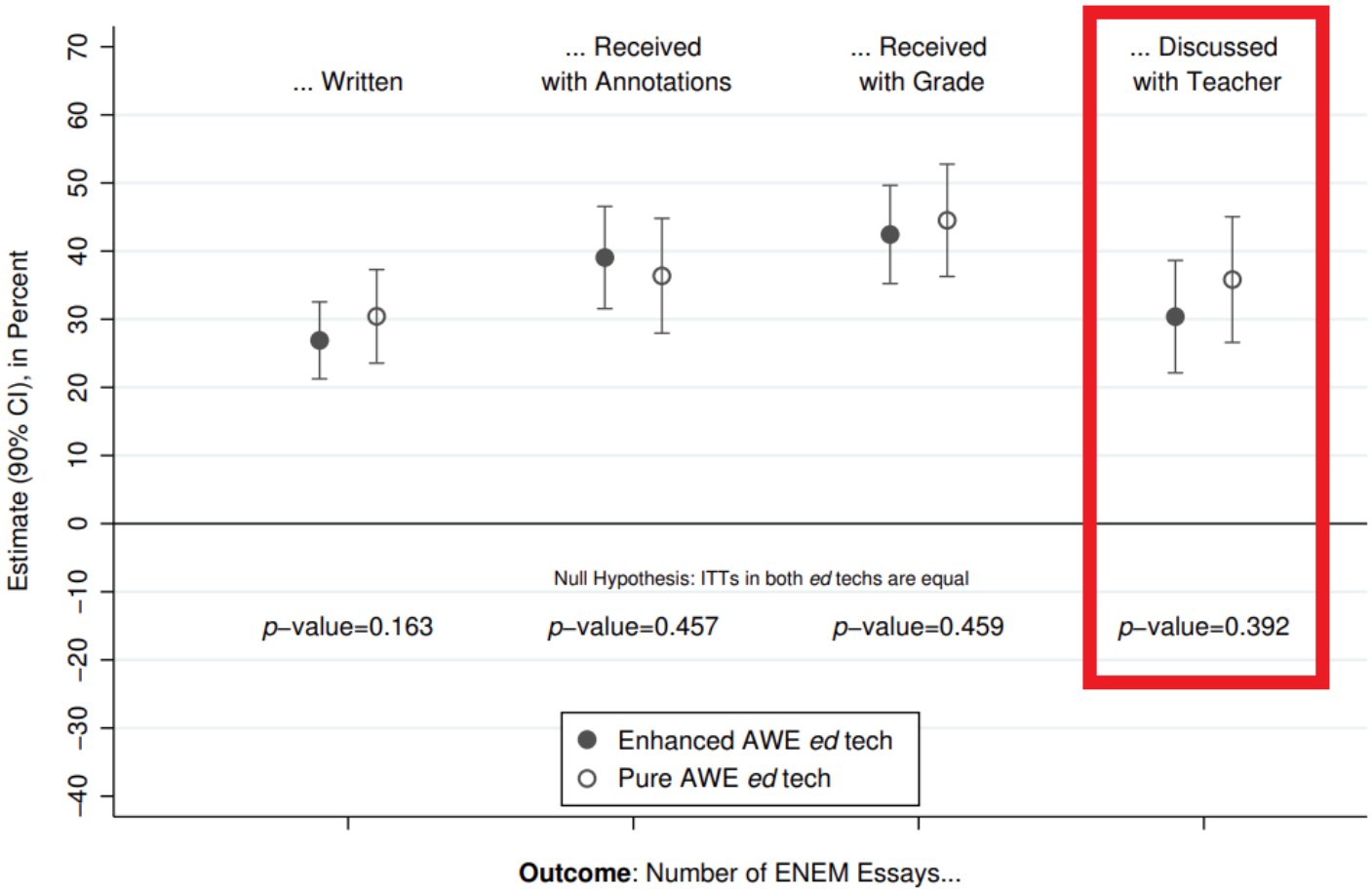


Figure: *ed* techs **Positively Impacted the Quality of Feedback**

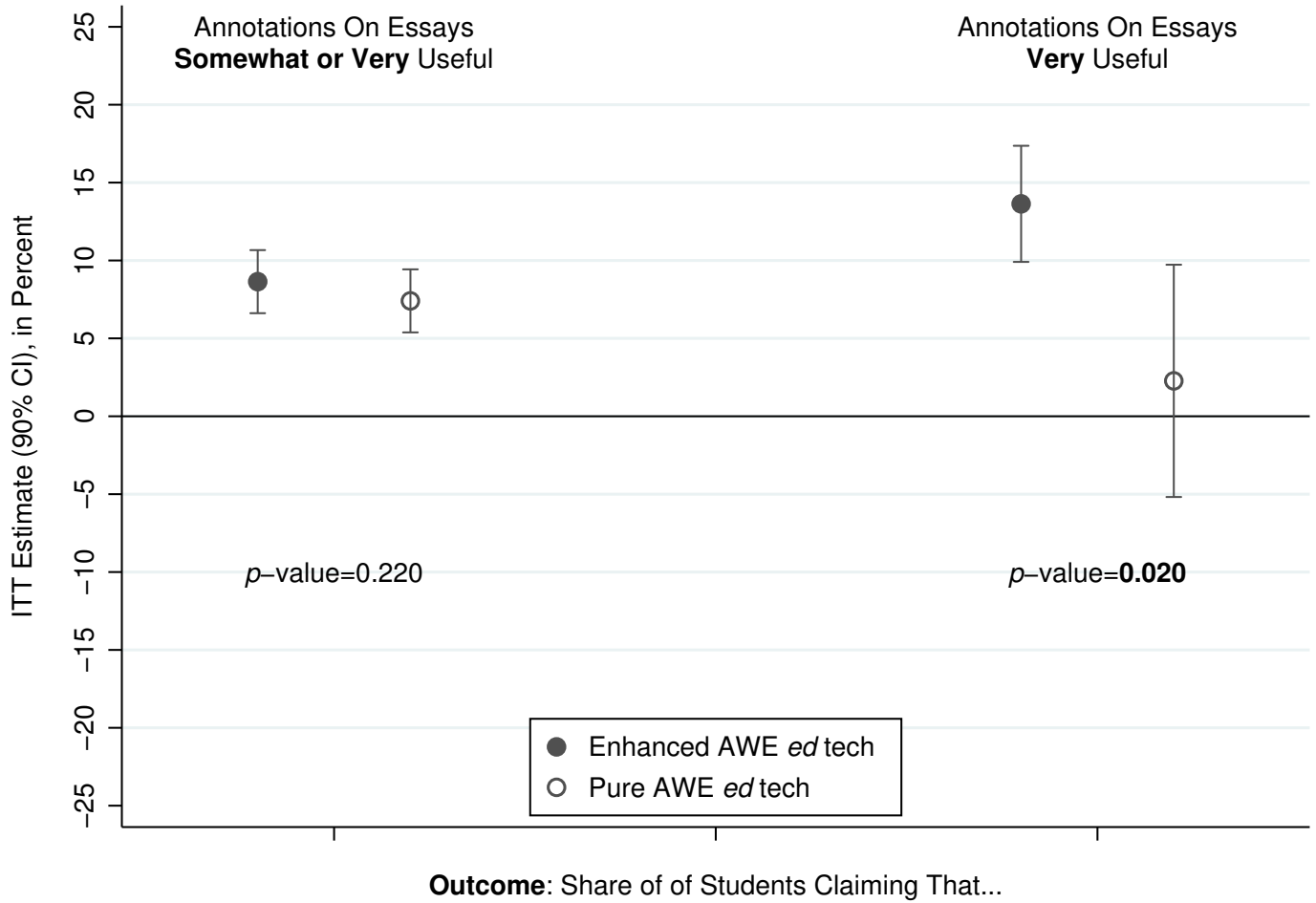
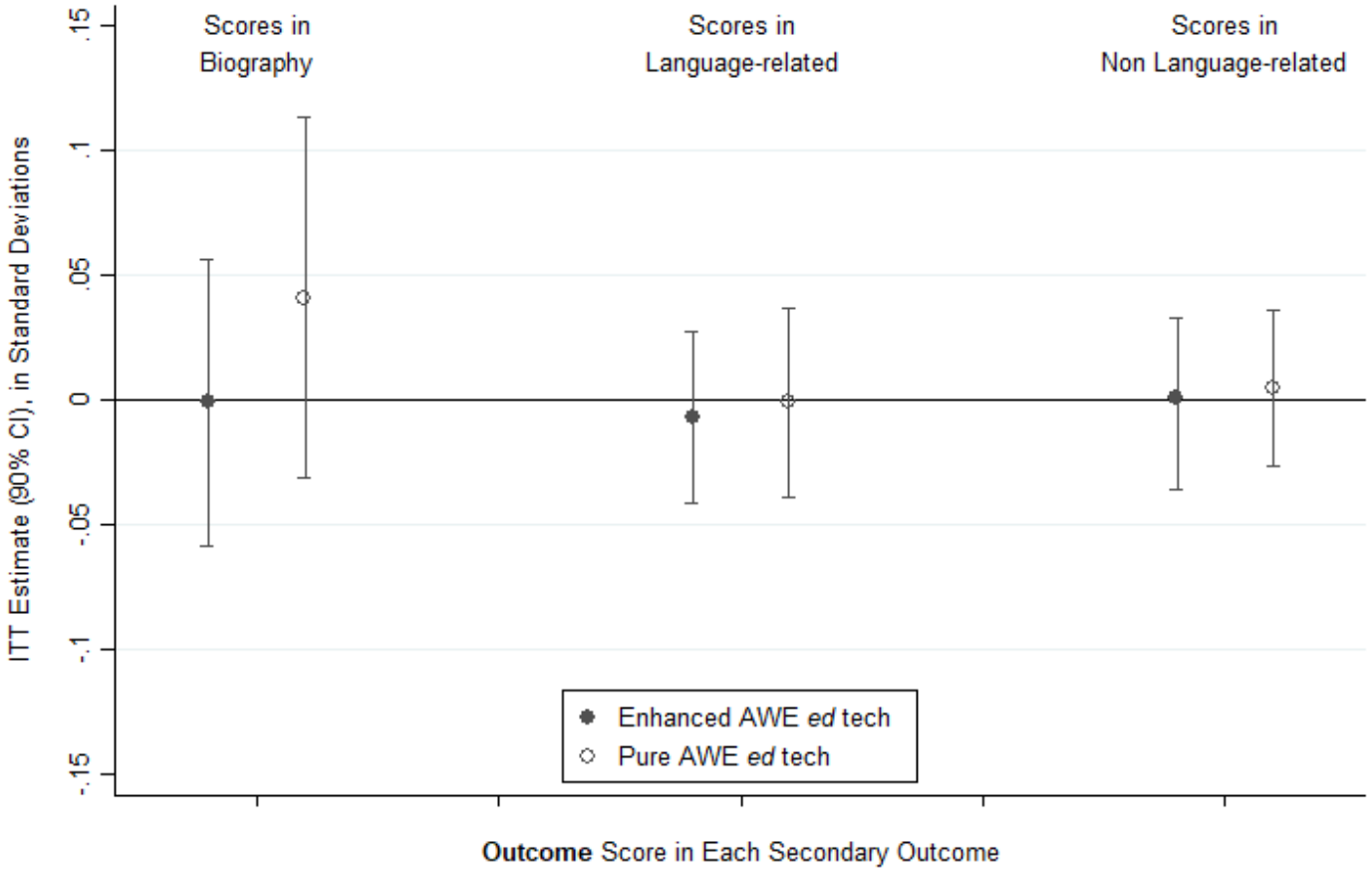


Figure: **No Evidence of Adverse Effects or Topic Complementarity**



Roadmap

- **Positive effects** of both *ed* techs on writing scores.

- **Positive effects** of both *ed* techs on writing scores.

- Despite improvements in perceived quality, additional inputs from **human graders** did not improve effectiveness.

- **Positive effects** of both *ed* techs on writing scores.
- Despite improvements in perceived quality, additional inputs from **human graders** did not improve effectiveness.
- Positive effects even in skills AI arguably falls short in evaluating.

- **Positive effects** of both *ed* techs on writing scores.
- Despite improvements in perceived quality, additional inputs from **human graders** did not improve effectiveness.
- Positive effects even in skills AI arguably falls short in evaluating.
- Suggestive evidence of complementarity: teachers' tasks shift toward **nonroutine** tasks (interpretation of essays and personal interaction about writing quality).

1. Control $\iff$ Pure AWE — **first impact evaluation** that relies on a large sample and a credible research design

1. Control $\iff$ Pure AWE — **first impact evaluation** that relies on a large sample and a credible research design

2. Control $\iff$ Enhanced AWE — **efficacy trial** of an *ed* tech tailored to overcome bottlenecks of pedagogy in public schools (human capital and time constraints)

1. Control $\iff$ Pure AWE — **first impact evaluation** that relies on a large sample and a credible research design

2. Control $\iff$ Enhanced AWE — **efficacy trial** of an *ed* tech tailored to overcome bottlenecks of pedagogy in public schools (human capital and time constraints)

3. Pure AWE $\iff$ Enhanced AWE:
   - **no added value** of human graders, which are costly
   - question of **perfect task emulation** (or substitutability) bypasses the **complementarities** based on expected comparative advantages between AI and human labor

Luckin et al, 2016, Intelligence unleashed: An Argument for AI in education.

"*AI-powered tools will serve as a catalyst for the transformation of the role of the teacher* [...] *allow*[*ing*] *teachers to devote more of their energies to the creative and very human acts that provide the ingenuity and empathy to take learning to the next level*."

Thank You!