

## FRONTIERS / REVIEW RECORD

# The Persuasion Hearing

## Complete faculty-role reviews and submitted-answer audits

All five AI political-scientist reviews for this simulation, in full, followed by all five case-specific assessments of actual AI-modeled learner submissions. These are analytical role reviews, not recruited professors or students.

The reviews assess the preceding version-3 / full-substantive-and-visual-v5 build. The separate faculty-led revision report documents which recommendations were implemented in v6. The original criticisms are preserved here; they do not constitute a fresh endorsement of the revised build.

### Reading map

P1: Comparative institutions and political authority. Original case critique, then submitted-answer assessment for S1.

P2: Democracy measurement and quantitative inference. Original case critique, then submitted-answer assessment for S2.

P3: State, society and political communication. Original case critique, then submitted-answer assessment for S3.

P4: Causal methods and evidence quality. Original case critique, then submitted-answer assessment for S4.

P5: Political concepts, teaching and transfer. Original case critique, then submitted-answer assessment for S5.

Every case field is reproduced. The review method, baseline runtime evidence, ranking and cross-module context are retained for interpretation. Ratings are design judgments, not measured learning effects.

# P1 / Original faculty-role review

AI political-scientist role: comparative institutions, democracy and political authority

Independent inspection of all five current version-3 full configurations, staged learner records and original artwork. No previous reviews or peer verdicts were consulted. This is an AI analytical review, not a credentialed external scholar, human learner trial, native-browser usability observation or fresh complete primary-paper audit. Source-use scores concern the specificity and internal consistency of the supplied explanations, not newly verified bibliographic accuracy. Ratings are judgments on a 1-5 scale, not learning-effect estimates.

## Ratings (1-5 judgments)

|                      |   |
|----------------------|---|
| Content coherence    | 5 |
| Source use           | 5 |
| Political depth      | 4 |
| Challenge            | 5 |
| Evidence sufficiency | 5 |
| Teaching potential   | 5 |

## Strengths

1. The four-cell Salvi map separates opponent-plus-profile from the added-profile contrast and retains the reported correction. Odds, attitude points, participant proportions and ballots are explicitly different quantities.
2. Hackenburg's randomized packages are not silently promoted to evidence that claim count or falsehood causes influence. The local version-ranking sheet is also correctly distinguished from a live rule selecting individual replies.
3. D4's six claims distinguish supported, contradicted and not established, while 'fairest' remains a value judgment. A link to official notices does not verify all sentences that cite them.
4. Official completeness and usable challenge duties differ from a disclosed advocate's legitimate directional role. The institutional rule has costs, powers, a complaint actor and burdens on legitimate use.
5. Westmere couples an intention-to-treat offer contrast with selected finishers and a later language-access failure. Accurate first-screen alternatives do not alone settle renewal incentives or a usable complaint right.

## Problems

### Item 1

### Severity

P2

### Location

revision.transfer T4 / revisionFields.revision

**Evidence**

The first offer experiment has a +0.2 knowledge contrast; the new corrected-guide experiment has +0.4. The record says these are different comparisons and no individual feature is independently varied, but the wording 'corrected encounter' invites attributing the increased contrast to first-screen coverage.

**Why it matters**

Comparing two treatment-effect estimates across separate experiments is not itself a randomized old-versus-corrected-guide contrast. It would undermine the excellent package-versus-mediator lesson if credited causally.

**Recommended repair**

Add a method question asking whether the change between experiments identifies the coverage effect, and require a direct randomized old-guide versus corrected-guide contrast with equivalent invitation/follow-up if that effect is the target. Keep the existing before/after descriptive information.

**Item 2****Severity**

P2

**Location**

revision.decision.fields.defense

**Evidence**

All six sentence classifications, reference reasons, value/fact distinction, an institutional causal chain, rival contrast, two source details and a switch pattern are submitted in one response field.

**Why it matters**

The factual audit may be omitted or buried in prose despite being one of the most concrete tests of whether the learner can check claims. A count gate will not detect missing categories.

**Recommended repair**

Provide six empty claim rows with classification and deciding-reference cells, plus a separate value-judgment note. Keep the connected causal/legal defense as a required argument after the audit, without auto-grading political agreement.

**Item 3****Severity**

P3

**Location**

dataWorkshop.guide and Article A method

**Evidence**

The guide describes an interval as a range compatible with the estimation procedure and the article defines odds briefly, but the task moves quickly between relative odds and additive score points.

### Why it matters

A medium-good novice may preserve the words without grasping why 81.2% relative odds is not 81.2% of people. This is one of the few terms where a small numerical example would add real understanding.

### Recommended repair

Add an unrelated numerical odds example with a clearly different outcome and values, alongside a visual two-condition contrast card. Do not compute an option from the political task or require technical odds conversion from these research results.

### What it teaches

Persuasive power, personal-data use, factual reliability, complete official information, disclosed advocacy and usable rights are separate governance questions. A randomized offer effect concerns all assigned people; voluntary completion is a different selected group. Correct facts can coexist with problematic institutional selection or access.

### What it still does not teach

It does not settle a general speech-regulation policy, identify actual voting effects or establish the causal mediator behind a multi-feature instruction. Its charter supplies official duties rather than investigating how the mandate itself was democratically legitimated.

### Does it stretch learners?

Yes. The source contrasts, sentence verification and all-assigned versus finisher distinction require careful thinking. It is the most research-methods-heavy module; the political licensing decision prevents it from becoming only a statistics worksheet.

### Scaffolding verdict

Enough evidence is supplied for the claim audit and several defensible licenses. The hints do not give numeric answers or a preferred policy. A structured audit table and different-topic odds example would support reading rather than lower the standard.

### Quantitative verdict

The +1 direct profile contrast with interval  $[-2, +4]$  cannot be inferred as exactly zero or as an established positive advantage. The full-interval-above-+3 screen is an explicit decision rule, not an empirical law. Westmere's all-offered +1 support and +0.2 initial knowledge contrast are interpretable; +8 selected finishers is not a completion effect. No precision or ballot effect is invented.

### Design content relation

The bright contemporary digital collage, transparent phone and hearing microphones fit mediation of information into public judgment. It is distinctive from the anime, print-lab and administrative poster styles. War-damage imagery among floating windows can introduce stakes beyond the actual transport/energy record; treat it as broad political media framing, not case evidence.

### Overall verdict

A rigorous political-communication and evidence workshop. It ranks fifth only because my lens prioritizes comparative institutions and these others are closer to a general introductory comparative-politics curriculum; it is not a weak product. I would use it where AI governance, persuasion and causal evidence are the primary goals.

## Ranking

1. democracy-alarm
2. overloaded-state
3. everyone-democracy
4. civic-machine
5. persuasion-hearing

## Cross-module context

1. No P1 fatal content defect was identified in this independent supplied-material review. That is not a certification of all sources, native app behavior or human learning.
2. These are now demanding reading-and-writing inquiries with actual institutional decisions and unfamiliar generated transfer, rather than five quizzes dressed in different art. Source boundaries generally retain a positive finding instead of reducing every conclusion to 'we need more research'.
3. The conceptual MC remains easier than the generative tasks: correctly identifying the balanced or carefully bounded formulation often suffices. Treat objective completion as formative evidence, not the headline measure of challenge or mastery.
4. The strongest common improvement is empty structural scaffolding: actor/action/outcome chains, measurement slots, claim-audit rows and cost ledgers. It should clarify what an adequate argument consists of without populating the mechanism, policy or answer.
5. Same-step rival reasoning is the highest-value shared learning goal but also the easiest place to produce jargon without an actual discriminating prediction. Human review should ask whether a named observation really differs under the two accounts, and whether apparently competing causes could coexist.
6. The many required responses create substantial burden, and current schedules correctly leave timing to a human pilot. I make no minutes-to-complete estimate. Repeated source reconstruction and final restatement may be streamlined without removing the initial commitment, late revision or independent transfer.
7. The original artwork provides five genuinely distinct visual identities. None of those image judgments establishes rendered mobile readability, keyboard access, animation comfort or real Gen-Z enjoyment. A small human browser pilot is still necessary before claims about liking or successful learning.
8. A serious human assessment should distinguish justified retention from automatic changing of mind, legal constraint from policy agreement, arithmetic from denominator understanding, and a completed text field from semantic adequacy.

## P1 / Assessment of S1 submitted answers

### Comparative institutions, democracy and political authority

Separate semantic assessment of S1's five actual submitted response.json files against the current version-3 configurations and original-script first-trial execution records. The independent initial review remains unchanged. S1 is an AI-modeled learner, not a real student. VM completion, word gates and correct objective choices are treated as recording evidence only. No learning gain, native-browser usability or real enjoyment claim is made.

## Recorded baseline execution evidence

### Method

Actual app JavaScript in minimal API/event VM; no native layout or human trial

### All five completed

Yes

### Objective first choices correct

30

### Objective first choices total

30

### Blocked modules

### Interpretation

S1's authored submissions were accepted on all five routes and all six choices per module were correct on first submission. This establishes that the recorded route worked for these submissions, not that every explanation met the semantic rubric. The separate root-authored false-boilerplate Civic negative control also completed; it is not student work and does not establish that the app promised automatic semantic grading.

## Cross-module synthesis

1. S1's actual prose is generally strong and specific: it names dated artifacts, actors, costs, base populations and real source procedures. That judgment is based on the written content, not its thirty correct MC choices or route completion.
2. The most consistent remaining semantic weakness is discriminating causal design. A named rival sometimes concerns another compatible link, or a proposed comparison holds away the feature needed to test its prediction. Targeted method questions are preferable to providing a mechanism or political choice.
3. All five institutional decisions are defensible within the given powers; do not substitute agreement with another reviewer's policy for evaluating reasoning. The uncoded Harrow operational detail is partly an application-design limitation, while Celyna missing-data handling and Orlen chronology are precision repairs to the actual response.
4. The student-model response defeats several tempting shortcuts: it constructs joint distributions, retains review despite delay, explains own-party restraint, avoids exact-zero profile claims and explicitly refuses a causal coverage claim from separate experiments. These are examples of attained reasoning in the produced texts, not evidence of learning gain by real people.
5. An unchanged objective score can coexist with meaningful differences in semantic adequacy. The app explicitly says reasoning awaits human review; the false-boilerplate control illustrates why that distinction must remain visible in completion screens and instructor interpretation.
6. A real pilot should assess whether medium-good learners can produce these connections under ordinary time, motivation and device conditions. These polished AI-modeled submissions do not establish feasibility, satisfaction or transfer for actual seventeen- or eighteen-year-olds.

**Overall evidence judgment**

A very strong response on corrected comparisons, factual auditing, official duties and selection. It explicitly avoids the separate-experiment coverage-effect trap identified in the initial review. The initial advocacy-versus-learning rival is compatible rather than strictly competing, but the later concrete comparisons improve it.

**Source use**

Correct four Salvi cells, direct versus package contrast, agreement odds, correction uncertainty and unaltered human-baseline result. Hackenburg's packages, multiple reply features and fabrication result constrain mediator claims. The cross-study possible world uses populations, interventions and units rather than significance labels alone.

**Core reasoning**

Correctly interprets the version-ranking sheet as ranking whole versions rather than each live reply. Claim/length/ordering differences are not identified as a single mediator. Classifies the six claims correctly and separates 'fairest' from fact verification. The initial learning rival could coexist with an approval objective; the later equal-length claim-feature comparison addresses a more specific link.

**Quantitative reasoning**

Correct +6 package contrast, +1 direct-profile contrast and interval  $[-2, +4]$ , with no percentages-converted or exact-zero claim. The whole-interval-above-+3 screen is unmet without making effects above +3 impossible. Westmere's all-assigned support +1 and initial knowledge +0.2 are correct; the selected +8 finishers and other-household  $-0.75$  average are compatible, not a randomized completion effect.

**Institutional decision**

Package A is correctly costed  $80+40=120$  hours, names inspection/correction/pause and independent clerk powers, retains nonbot access and separates official completeness from disclosed advocacy. It names burdens on legitimate advocates and thinner outcome testing as an actual objection. All six factual categories use the correct deciding references or missing data.

**Late record update**

Retains a representation/access rule without claiming support reduction is necessary. T4's +0.4 knowledge contrast is explicitly not attributed to corrected first-screen coverage across separate experiments. Accessible complaints become a further institutional problem; renewal objective, verified facts and encounter coverage remain distinct.

**Independent transfer**

Proposes an all-assigned early-option comparison, then a language-access comparison after the late record. Keeps nonusers, accessible follow-up, linguistic strata and unmeasured approvals/votes explicit. Specifies a clerk-capacity alternative if translation does not solve access. Selection is appropriately identified as explaining finisher movement rather than erasing the randomized offer contrast.

**Findings in the submitted answers****Item 1****Kind**

strength

**Severity**

none

**Location**

transfer.revision

**Quote**

cannot claim it caused the new 0.4 knowledge contrast because several things changed across comparisons

**Judgment**

This directly avoids the principal causal trap flagged by my independent initial review. It is evidence-specific restraint paired with a positive rights recommendation, not empty caution.

**Feedback**

Retain the direct randomized old-versus-corrected encounter design if claiming a coverage effect.

**Item 2****Kind**

strength

**Severity**

none

**Location**

decision.defense

**Quote**

three is not established because B gives only a ceiling

**Judgment**

Correctly distinguishes an unapproved operating estimate from a contradicted fact. All six factual classifications are right for stated reasons; the value judgment remains a defended normative claim.

**Feedback**

Keep reference reasons visible when these judgments are exported for human review.

**Item 3****Kind**

gap

**Severity**

P2

**Location**

investigation.rival and prediction

**Quote**

truthful information helps people understand benefits ... without a concealed advocacy objective

Judgment

Truthful learning can occur under an approval-oriented objective, and a knowledge-focused version can also persuade. The contrast initially treats compatible objective and learning paths as rivals. Later writing improves the causal specificity, but the initial prediction does not discriminate them by itself.

Feedback

Choose one same-step outcome such as alternative knowledge at a fixed text length; define a reply-feature account versus prior-understanding/processing account and predict an observation on which they differ. An observed approval objective alone does not rule out learning.

Human-review judgment proposed by the AI role

Credit source comparisons, all claim classifications, interval/assignment reasoning, lawful enforcement and the precise T4 update. Revise the initial rival into a same-link discriminating test; do not penalize a defended package A rather than B or C.

P2 / Original faculty-role review

Independent AI political-scientist role: democracy measurement, quantitative methods, causal inference and assessment discrimination. No real credentialed reviewer is represented.

Independent inspection of all five current full-configs, staged learner content, original artwork and the shared JavaScript data-display/recording implementation. No peer reviews or learner answers read before this initial report. Internal consistency and source-boundary review, not a fresh primary-paper audit. No native browser, human participation, measured enjoyment or efficacy evidence. Ratings are exploratory judgments on a 1-5 scale.

Ratings (1-5 judgments)

|                      |   |
|----------------------|---|
| Content coherence    | 5 |
| Source use           | 4 |
| Political depth      | 4 |
| Challenge            | 5 |
| Evidence sufficiency | 5 |
| Teaching potential   | 5 |

Strengths

1. The corrected direct profile comparison is separated from the human-baseline result; relative odds are not converted into a percentage persuaded. Package effects, downstream correlates and a fabrication result are kept distinct.
2. The direct +1 [−2,+4] attitude-score contrast is shown alongside two advocacy-versus-neutral contrasts. The threshold item asks whether the whole interval lies above +3 while explicitly avoiding exact-zero and impossibility conclusions.
3. The official guide's information duty differs from disclosed advocacy. Six sentence audits distinguish supported, contradicted and not established. The unfamiliar offer experiment includes nonusers, a

finisher-selection problem, knowledge versus support and a later language-access complaint.

## Problems

### Item 1

#### Severity

P2

#### Location

dataWorkshop.guide; checks.data-read and data-infer

#### Evidence

The supplied direct interval spans zero and the policy screen stipulates its entire interval exceed +3. The current items principally assess correct reading of that supplied contrast.

#### Why it matters

Students may learn 'crosses zero, be cautious' without understanding why subtracting two point estimates does not supply the uncertainty of their difference or why a policy threshold is not a scientific law.

#### Recommended repair

Add a short authored comparison identifying the necessary direct contrast and distinguishing a stipulated decision threshold from evidence of equivalence. Ask the learner to state what precision/decision information is still needed, without deriving an interval from overlapping separate intervals.

### Item 2

#### Severity

P2

#### Location

revision.transfer.records T2-T3 and lateRecord T4

#### Evidence

All-offered support movement is +1, computed from 100 finishers at +8 and 400 other households at  $-0.75$ . Randomization concerns the guide offer, not finishing; before-invitation interest differs.

#### Why it matters

The offer/finisher contrast is the central quantitative transfer. A learner can quote selection while still attributing the +8 to completing the conversation or assuming prior-interest measurement fully corrects the bias.

#### Recommended repair

Require the weighted recomposition calculation and two explicit estimands: effect of offer among all assigned versus the unestimated effect of completing a dialogue. Have the learner name a design that could identify a bounded completion/use claim and its assumptions, while retaining all households for the offer claim.

### Item 3

#### Severity

P3

**Location**

revision.decision.fields.defense; D4 sentence audit

**Evidence**

One defense box houses six fact classifications with references, a separate value judgment, an institutional causal chain, a same-link rival, a comparison, both sources and a switching rule.

**Why it matters**

The rich audit can be hidden in a general governance paragraph, making it difficult to see whether the student actually distinguished contradiction from absence of evidence.

**Recommended repair**

Add a six-row blank audit grid with status and decisive reference/missing evidence, followed by the integrated political defense. Do not classify the sentences in hints. Retain a separate value-judgment field and human review of the causal chain.

**What it teaches**

Direct component contrasts, effect uncertainty, policy thresholds, randomized packages versus mediators, fact checking versus complete encounters, official versus advocacy mandates, intent-to-offer outcomes and post-assignment selection.

**What it still does not teach**

It does not estimate a local persuasion effect, identify a claim-count mediator or turn a preference outcome into ballots. It introduces relative odds in the paper explanation but the fictional workshop uses score points; transfer between those metrics must be explicitly monitored.

**Does it stretch learners?**

Yes. The sentence audit, direct comparison and unfamiliar all-assigned/completer analysis require genuine work. The core generative chain can still become boilerplate if actors, one causal link and rival outcomes are not assessed.

**Scaffolding verdict**

The different-topic interval worked example is helpful and does not select the political-case option. This module benefits from labeled contrast and audit organizers, not simplified definitions that imply uncertainty means no risk or no duty.

**Quantitative verdict**

Strongest uncertainty and causal-estimand lesson. The provided toy direct contrast is +1 with  $[-2, +4]$ ; it fails the entirely-above-+3 screen without excluding all effects above +3. Transfer offer contrasts are +1 support and +0.2 knowledge, with no precision estimates. Completion is selected;  $100 \times 8 + 400 \times (-0.75) = 500$  gives the +1 all-offered mean. Distinguish these score points from the paper's odds metric.

**Design content relation**

The inspected bright digital/political collage combines phones, official hearing microphones, citations, alternative scenes and campaign discourse. It fits scrutiny of the encountered message. Its technological spectacle should not imply an inevitable or sinister AI effect; the substantive text keeps legitimate accurate advocacy open. No rendered layout or actual preference testing occurred.

## Overall verdict

An unusually responsible AI-persuasion lesson with real statistical depth. Preserve uncertainty and legitimate political tradeoffs, but make the fact audit and offer-effect reasoning visible products rather than buried prose.

## Ranking

1. everyone-democracy
2. democracy-alarm
3. overloaded-state
4. persuasion-hearing
5. civic-machine

## Cross-module context

1. Ranking reflects this role's measurement-and-methods priorities, not empirical comparative efficacy, enjoyment or a declaration that the fifth module is weak.
2. All five are substantive reading-and-writing inquiries, with fixed fictional records rather than outcome-changing roleplay engines. Challenge is already high. Add evidence-organizing scaffolds, not reduced reading or easier conclusions.
3. The data workshops cover distinct concepts: margins/joint distributions, aggregation/weights, selected sample frames and cohorts, inverse conditioning, direct contrasts/uncertainty. This is a coherent progression rather than five cosmetically different copies.
4. Exact plotted values are repeated in accessible tables. That is good multimodal access, but the existing MCs principally demonstrate numerical/table inference, not comprehensive graph literacy. Add one chart-representation critique or student-generated display with an accessible equivalent in each module; do not withhold data from screen-reader users to make a graph harder.
5. Hint restraint is generally good: method questions and different-topic examples do not select an option. Correct-choice prose is still often more bounded and complete than distractors. Replace a subset with balanced credible alternatives plus document/calculation justifications.
6. The application explicitly states that word gates record participation and human review assesses prose. Do not treat completed status, six solved MCs or a high word count as mastery. Exporting rubric-linked evidence is more important than making an automatic semantic-grade claim.
7. Independent human classroom piloting is still needed for actual time, comprehension, engagement, accessibility and learning gain. These AI judgments cannot establish youth preferences or efficacy.
8. Source-use ratings refer to internally precise presentation and transparent boundaries in the current materials. I did not independently retrieve and reread the ten primary articles or reproduce their estimates in this review.

## P2 / Assessment of S2 submitted answers

Democracy measurement and quantitative inference

Separate semantic follow-up written only after saving the independent initial review. Read all five S2 response.json files in full and their unchanged first-trial VM output. Assessed authored reasoning against supplied records and tasks, not a real student, learning gain or native-browser test. No peer reviews used. Quotations below are from AI-modeled S2 responses.

## Execution

### Modules completed

5

### First choice correct

30

### Objective checks

30

### Recording gate failures

0

### Boundary

Original application JavaScript API/event VM accepted all five; this verifies recorded choices and completion, not semantic mastery, human learning or rendered usability.

## Cross module conclusions

1. All thirty S2 objective first choices were correct and all five VM runs completed unchanged. This establishes only recorded completion and MC recognition. There is no before/after learning estimate or real student success rate.
2. S2's actual writing is not generic boilerplate: two joint patterns, an original 12-slot instrument, all six fact classifications, feasible costed decisions and changed transfer recommendations provide independent semantic evidence. It should be acknowledged rather than dismissed merely because recording gates cannot grade it.
3. No numerical calculation errors were found in S2's submitted explanations. The recurring weak points concern operational definitions, assignment/selection, same-outcome rivals and rights-preserving follow-up, which are more advanced than ordinary arithmetic checks.
4. Make the strongest generative products explicit rubric-linked submissions: joint tables, indicator thresholds, cohort/outcome coding, instrument crosswalk and audit grid. This adds scaffolding and diagnosability, not an easier intellectual standard.
5. The root's reported false-prose negative control is consistent with the app's explicit participation-only gate boundary, but I did not run that control or use it as evidence that S2's own writing was false. All semantic judgments above derive from the actual S2 texts and supplied records.

### Semantic verdict

Strong contrast/metric, fact-audit and governance reasoning. No numerical errors were found. The most important improvement is making the competing causal question identical for mechanism and rival, plus protecting usable rights in the proposed language-access experiment.

## Strengths

### Item 1

#### Location

research.crossInference; brief.dataReason

#### Evidence

S2 explicitly says odds cannot be subtracted from attitude points, separates the unchanged human-baseline finding from corrected component uncertainty, and correctly reads +6 [2,10] versus neutral and +1 [-2,4] for the direct profile contrast.

#### Assessment

This is a sound multi-metric, multi-comparison interpretation, not merely 'zero crosses an interval.' It avoids exact-zero, vote and no-privacy-duty inferences.

### Item 2

#### Location

decision.defense

#### Evidence

S2 classifies all six D4 sentences correctly with T/F/B reference reasons: supported, contradicted, not established, supported, contradicted, not established; 'fairest' remains a value judgment.

#### Assessment

The submitted prose really performs the sentence audit. This demonstrates content discrimination beyond the completion gate.

### Item 3

#### Location

decision.choice; transfer.revision

#### Evidence

S2 costs package A at  $80+40=120$  hours, retains disclosed advocacy and alternate access, separates +1 offer support/+0.2 knowledge from finisher movement, and explains why the new +0.4 knowledge contrast is not an isolated ordering effect.

#### Assessment

The institutional mandate and causal-source limits are substantively applied. It retains nonusers and identifies the majority-language complaint barrier after first-screen coverage improves.

## Weaknesses

### Item 1

#### Severity

P2

#### Location

investigation.provisional

**Quote**

The references already contradict some factual claims, so movement is not simply verified learning.

**Assessment**

Inaccurate claims make a pure 'all claims verified' account untenable, but their presence does not by itself show which claims caused the support movement. Verified facts could still drive all of the effect while errors are incidental. The later response recognizes mediator uncertainty, so this is an overstrong intermediate phrasing rather than a persistent wholesale misconception.

**Improvement**

Ask S2 to separate an audit finding about reply content from a causal claim about the source of movement. Require a knowledge/perceived-evidence measure or a comparison that discriminates accurate-learning versus other pathways.

**Item 2****Severity**

P2

**Location**

transfer.mechanism, rival and prediction

**Evidence**

The mechanism addresses omission of the waiver and all-offered outcomes; the rival focuses on prior interest selecting finishers and the +8 finisher association.

**Assessment**

Both are valid concerns and can coexist. Selection into finishing is a rival explanation of +8, not by itself an explanation of the randomized +1 offer contrast. A single contrast should not claim to choose between these accounts for different outcomes. S2 is careful elsewhere, but the task's same-link rivalry remains partly underspecified.

**Improvement**

Require the learner to name one target outcome/link for the rivalry. For +8 use completion-selection alternatives; for the offer effect use two competing mechanisms of the offered guide, with all assigned households retained. Also request the weighted recomposition  $100 \times 8 + 400 \times (-0.75) = 500$  to make the coexistence explicit.

**Item 3****Severity**

P2

**Location**

transfer.test

**Quote**

Randomly assign equally complete versions with a multilingual complaint route or the current majority-language route.

Assessment

The comparison is well specified for assigned outcomes, but the current route is known to be unusable for some households. The proposal does not specify a usable alternate complaint route retained for both arms while language effects are tested. Given the learner's own rights criterion, this is an important protection omission rather than a reason to abandon experimental thinking.

Improvement

Keep equivalent usable complaint access through a staffed/translated fallback in both arms, then vary interface format or discoverability within that protection floor. Have S2 state the burden, missing-follow-up plan and what causal claim the safer comparison still permits.

Completed versus mastered

Strong authored evidence supports the main numerical, source and governance distinctions. A completed badge still cannot prove mediator reasoning, common-outcome rivalry or protection-preserving study design; these need semantic review.

P3 / Original faculty-role review

Independent AI political-science role: political behavior, public opinion and political communication

Initial independent inspection of the five current full configurations, their staged learner material and all five original artwork assets. No peer reviews or learner submissions were inspected before this report. Internal source-fidelity and task-design assessment only: I have not conducted a new primary-paper audit, recruited humans, observed a browser session or measured learning gain. Ratings are exploratory judgments, not empirical outcomes.

Ratings (1-5 judgments)

|                      |   |
|----------------------|---|
| Content coherence    | 5 |
| Source use           | 5 |
| Political depth      | 5 |
| Challenge            | 5 |
| Evidence sufficiency | 4 |
| Teaching potential   | 5 |

Strengths

1. The corrected Salvi direct-profile contrast remains distinct from the human-baseline package effect; relative odds do not become percentages persuaded.
2. Instruction packages, produced claim counts, accuracy and mediation are not conflated. The fabrication finding is not used to dismiss truth as a political value.
3. D4 forces actual sentence-to-reference auditing and distinguishes contradicted facts, unsettled facts and a normative judgment.
4. Official information duties and openly sponsored advocacy have different institutional roles under a common privacy/factual floor.

5. Westmere requires all-assigned contrasts, separates selection of finishers from treatment and makes language-access complaints an enforceable-right question.

**Problems****Item 1****Severity**

P2

**Location**

revision.decision.fields.defense and D4

**Evidence**

Six factual classifications, a value judgment, a causal chain, a same-link rival, both sources and a switch pattern occupy one defense field.

**Why it matters**

Sentence audit omissions may be masked by fluent licensing rhetoric and a correct MC performance.

**Recommended repair**

Provide a six-row blank audit grid: sentence, supported/contradicted/not established, deciding reference or missing record. Follow with a separate normative reason and causal defense. Keep all classifications student-generated.

**Item 2****Severity**

P2

**Location**

P-14/P-29 and transfer T3

**Evidence**

Package randomization and post-treatment features are explained, while selected completers differ in prior interest and receive different experiences.

**Why it matters**

A learner may treat extra words as a proven mediator or prior-interest adjustment as sufficient control of finishing selection.

**Recommended repair**

Require the proposed design to name the randomized component, the estimand and the population retained. Add an unsolved challenge asking which inference is lost if only finishers are followed. Use existing records, not a key-bearing hint.

**Item 3****Severity**

P3

**Location**

dataWorkshop.guide and workedExample

**Evidence**

The unrelated museum example reproduces the direct-contrast and zero-spanning interval structure closely.

**Why it matters**

It is useful support, but optional example use can turn the MC into near-transfer pattern matching even while the political reasoning remains difficult.

**Recommended repair**

Keep the example available but record example use separately, and add a delayed check with a differently arranged interval display requiring an original written contrast interpretation. Do not treat helped first-check success as independent mastery.

**What it teaches**

Capabilities versus rights; component versus package effects; uncertainty; factual auditing; legitimate advocacy versus official-service coverage; actor incentives; selection, offer effects and usable complaint rights.

**What it still does not teach**

It does not establish actual electoral influence, a particular causal mediator, real-world uptake or the effect of a proposed license rule. The source claims and correction are internally handled carefully, but I did not independently verify the primary correction in this review.

**Does it stretch learners?**

Yes. The best challenge is connecting a correct sentence audit and uncertain effect to a defensible enforceable rule rather than shouting that all persuasion is manipulation.

**Scaffolding verdict**

Method hints are appropriately restrained. Blank reference-audit and estimand ledgers would support difficult thinking more effectively than another general reminder that evidence has limits.

**Quantitative verdict**

The exact direct contrast is excellent for preventing an interval-overlap shortcut. Westmere adds substantial transfer: support +1 and knowledge +0.2 in the first offer comparison, with no supplied precision estimate.

**Design content relation**

The surreal digital collage, cables, phone and public microphones represent a mediated public encounter without a villainous robot motif. It is visually exciting but should remain outside precision charts and sentence-audit rows.

**Overall verdict**

My second choice overall and first for political-communication methods. Retain the challenge, but make the six factual classifications and causal estimand independently inspectable.

**Ranking**

1. everyone-democracy
2. persuasion-hearing

3. civic-machine
4. democracy-alarm
5. overloaded-state

### Cross-module context

1. Ranking reflects fit for my political-behavior/communication lens, not a statistical difference in quality. All five have a substantial political core and are preferable to a simple vocabulary quiz.
2. The major cross-module limitation is assessment validity: word thresholds and completed MC checks establish activity attempts, not the semantic quality of a rival, feasible rule or transfer explanation. The teacher rubric acknowledges this correctly; teachers must actually use it.
3. Do not confuse sustained reading with high-friction task search. Preserve the two papers, raw records, counterevidence and writing demand; make multi-part deliverables visible as blank structured ledgers.
4. The art assets are genuinely distinct and relevant. No browser layout, text contrast at runtime, mobile clipping, keyboard operation or accessibility result is asserted here.
5. A real pilot should record unaided and hinted performance separately, use a delayed unfamiliar case and ask medium-good learners to explain their reasoning aloud. These AI reviews cannot establish enjoyment rates, learning gains or population readiness.

## P3 / Assessment of S3 submitted answers

State, society and political communication

Separate semantic follow-up after the initial independent report was saved. I read S3's actual response.json and first VM trial for each of all five modules. S3 is an AI-modeled medium-good learner, not a human. The limited app VM reports all five complete and six correct first choices each; I judged the written explanations independently. No peer reviews, browser observations, source-paper re-audit or learning-gain inference is used.

### Summary

S3's actual prose is generally strong and source-specific, not merely long or cautious. All five numerical explanations are materially correct, all institutional choices are broadly feasible, and the late-record revisions use relevant evidence. The most important shortcomings are incomplete executable survey items in Everyone Says Democracy, an insufficiently identified process-legitimacy comparison in Civic Machine, and a stale stopping rule after the South pilot changes in Overloaded State. These passed the activity gates, demonstrating why completion is not a semantic grade.

### Cross module findings

1. Thirty correct first choices and five complete runs coexist with incomplete open deliverables. That supports the app's explicit claim that gates establish attempts, not mastery; it does not imply the objective checks are defective.
2. S3 is much stronger in these AI-authored responses than a minimal-word submission: actual source details, exact records, numerical bases and legitimate alternative policies appear throughout. These are features of the submitted texts, not evidence of real human learning.
3. The recurring semantic weakness is causal-design inspectability: the treatment, population and outcome can drift between initial explanation, rival, late revision and final proposed test.

4. The most useful support is blank structured response organization plus a short teacher checkpoint. Giving the diagnosis, selecting a preferred policy or removing counterevidence would reduce the valuable challenge.

5. No preference or enjoyment conclusion is inferred from these written answers; that belongs to the separately labelled modeled learner reaction report.

### **Runtime**

#### **Completed**

Yes

#### **Correct first choices**

6

#### **Total checks**

6

#### **Failed gates**

#### **Semantic verdict**

Strong corrected-source reasoning, complete sentence audit and all-assigned numerical transfer. The successive causal tests need a clearer statement of which outcome each rival actually explains.

### **Strengths**

#### **Item 1**

##### **Location**

research.sourceA and crossInference

##### **Evidence**

The response reconstructs the four cells, distinguishes the direct-profile and human-baseline contrasts, and explains how small true effects and uncertain estimates can coexist without subtracting unlike units.

##### **Judgment**

The correction is used appropriately and not turned into exact zero, a vote forecast or a reason to abandon privacy rights.

#### **Item 2**

##### **Location**

decision.defense D4 audit

##### **Evidence**

All six classifications are correct: supported, contradicted, not established, supported, contradicted, not established, each linked to the relevant reference or missing poll. 'Fairest' is separately normative.

##### **Judgment**

This is actual sentence-level verification, not a generic claim that sources should be checked.

**Item 3****Location**

investigation.reason

**Evidence**

Instruction and verification trials are kept separate; claim density is not treated as the isolated mediator, and no undisclosed database access is inferred.

**Judgment**

The revision retains an upstream objective while bounding which downstream links are known.

**Item 4****Location**

decision.choice

**Evidence**

Package A is costed at 80+40 hours, with five-day correction and independent clerk review; official encountered alternatives are distinguished from lawful sponsored advocacy.

**Judgment**

The institutional rule is enforceable and acknowledges burdens and opportunity cost.

**Item 5****Location**

transfer.mechanism/revision and brief.dataReason

**Evidence**

The first all-assigned contrasts are +1 support point and +0.2 knowledge points; the response denies a causal finisher effect and does not attribute the later +0.4 knowledge contrast to coverage alone. The displayed +1 profile interval [-2,4] is not exact zero.

**Judgment**

All arithmetic, units, populations and precision boundaries are materially sound.

**Concerns****Item 1****Severity**

P3

**Location**

transfer.rival and prediction

**Evidence**

The positive account concerns missing waiver presentation and all-offered knowledge, whereas the rival emphasizes selected finishers' +8 movement and prior interest.

**Judgment**

Both are plausible and can coexist; the same-link/same-outcome competition is not fully specified. The proposed randomized presentation contrast does address a coverage effect, but prior-interest selection is not an exclusive alternative to that effect.

**Support without answer**

Require the learner to nominate one outcome and population for the rival test. Ask whether the rival predicts no coverage effect or only explains why the finisher estimate differs; permit coexistence instead of forcing opposite labels.

**Item 2****Severity**

P3

**Location**

transfer.test after T4

**Evidence**

The next test varies translated complaint access and compares an access-barrier account with an interest-only account, while earlier explanations concern first-screen coverage and version incentives.

**Judgment**

The new rights-access question is worthwhile and the learner states that renewal incentives remain unidentified. It should be labelled a newly prioritized link, not presented as if it fully settles the earlier source-of-support rival.

**Support without answer**

Ask the learner to write which earlier uncertainty is deferred, the new question being tested and the observation that could change the current language-access condition.

**Quantitative assessment**

No material numerical error. Offer effects, selected completer changes, package contrasts and component uncertainty are separated correctly.

**Priority fix**

Use a same-outcome/population/step ledger for successive rivals; preserve the strong audit and governance work.

## P4 / Original faculty-role review

Independent AI political-scientist role: comparative public policy, administration, governance and causal inference; not a real professional participant

Independent initial content pressure test of all five version-3 builds through full configs, six staged learner packets and original art inspection. No peer or learner answers were read before this report was saved. The source summaries were assessed for internal coherence and appropriate use, not independently verified against complete primary papers. No native-browser, backend, recruited-human or learning-gain claim is made.

**Ratings (1-5 judgments)**

|                      |   |
|----------------------|---|
| Content coherence    | 5 |
| Source use           | 5 |
| Political depth      | 4 |
| Challenge            | 5 |
| Evidence sufficiency | 4 |
| Teaching potential   | 5 |

**Strengths**

1. The corrected personalization contrast is not turned into zero and is not merged with the preserved human-baseline package contrast. Odds, attitude points, persuaded head counts and votes remain separate.
2. The audit distinguishes false, unsupported and normative statements. Verified listed facts do not prove complete encountered information, and a citation link is not evidence for every sentence.
3. The official guide has a completeness duty while a labelled advocate may legitimately argue for an option. Privacy, truth, disclosure and legitimacy are not collapsed into one ban.
4. The licensing packages name 120 hours, enforcement, correction times, nonbot access and complaints. A monitor cannot approve the route or prohibit all lawful advocacy.
5. Westmere's all-assigned offer contrast, selected finisher contrast and language-access problem are politically meaningful transfer demands. T4 records an upstream incentive without treating that document as its causal effect.

**Problems****Item 1****Severity**

P2

**Location**

D4 sentence audit inside decision.defense

**Evidence**

Six classifications with reference reasons are required inside the same long defense that must also supply a process, rival, causal test, two-source constraints and switching evidence.

**Why it matters**

A learner may produce a strong licensing rationale but omit one or more factual audits, and the app's text gate will not detect the omission.

**Recommended repair**

Provide six blank classification rows with reference/reason fields and a separate value-judgment line, then retain the authored causal defense. Grade distinction quality, not whether the learner agrees with a sponsor.

**Item 2****Severity**

P2

**Location**

First causal process and T4 revision

**Evidence**

Source material and later records explain version objectives, reply features, encountered alternatives, knowledge/support and language access, but several upstream/downstream features remain unassigned together.

**Why it matters**

The demand to generate a mechanism may encourage a neat causal story stronger than the actual package evidence.

**Recommended repair**

Offer an optional blank chain with each link marked observed, proposed or experimentally contrasted. Require the learner to choose one unidentified link and state two different test predictions; do not fill it with the case's mechanism.

**Item 3****Severity**

P3

**Location**

Visual conceptual fit

**Evidence**

The glossy phone and flowing ribbon collage makes persuasion look like an almost literal stream from device to institutions.

**Why it matters**

Its visual metaphor can imply direct mass influence, whereas reach, completion, access and institutional duties are the lesson's central limits.

**Recommended repair**

Keep the collage, but pair it with an exact small assigned/start/finish display using the stated Westmere counts, labelled fictional. Add a distinction between information access and support measurement rather than additional ominous imagery.

**What it teaches**

Evaluate a named contrast, not an alarming headline; truthful statements, complete encounters and legitimate advocacy are distinct institutional questions. Governance rules need authority, enforcement, rights, burden and evidence; population reach and selected-user influence answer different questions.

**What it still does not teach**

It does not isolate claim count, errors or profile access as universal causal mediators, estimate a voluntary election effect, or establish that a specific licensing rule is optimal. A formal floor still needs usable complaint implementation.

**Does it stretch learners?**

Yes, especially for quantitative uncertainty, six-sentence source audit and offer-versus-finisher transfer. The strongest political reasoning is defending differentiated official and campaign duties, not simply saying that AI can persuade.

**Scaffolding verdict**

The museum examples explain method without naming the political option; demand remains high. Separate blank evidence organizers would be more useful than easier checks.

**Quantitative verdict**

Well-designed direct contrasts and zero-crossing interval interpretation. Westmere permits exact descriptive +1 support and +0.2 knowledge offer contrasts, but supplies no precision estimate and no causal isolation of the newly corrected feature.

**Design content relation**

Y2K/digital collage gives a distinctive contemporary media encounter. It should not substitute visual drama for observed public reach; the exact graphs must remain clearly separate.

**Overall verdict**

I liked the applied licensing challenge and careful scope. It is a strong pilot candidate with a separate claim-audit surface and visible observed-versus-inferred chain.

**Ranking**

1. overloaded-state
2. civic-machine
3. democracy-alarm
4. persuasion-hearing
5. everyone-democracy

**Cross-module context**

1. This ranking reflects one explicitly modeled administrative/governance lens, not a population preference or quality league table. Everyone Says Democracy ranks fifth partly because of a specific repairable post-answer defect, not shallow political content.
2. All five have enough supplied political content for demanding medium-good learners: institutions appear as powers, incentives, legal constraints and public opportunities, not just topic decoration. They need human criterion-level review to test whether the content was actually learned.
3. The class tasks are rigorous evidence-reasoning practicums with branching/commitment features, not full dynamic political strategy games. Do not equate a saved choice with a simulated policy outcome.
4. The recurring weakness is deliverable bundling: long questions contain many distinct requirements in very few answer fields. Preserve sustained reading, but provide optional blank ledgers and mark which

analytical links need revision.

5. Method hints generally preserve productive difficulty. Wrong-answer feedback must explain the actual selected error; a correct key plus stale affirmative feedback is not an adequate assessment system.

6. No learning gain, real enjoyment distribution, teen usability, native accessibility or backend robustness follows from these AI judgments. A facilitated human pilot should measure completion paths, quality of transfer, time/effort and delayed unfamiliar reasoning.

## P4 / Assessment of S4 submitted answers

### Causal methods and evidence quality

Separate semantic audit conducted only after the independent initial scholar review was saved. Read all five actual S4 authored response.json files, their saved staged commitments and five first-trial VM records. Commitment fields match the final response fields in all five; this preserves the recorded initial answers but is not proof of a human participant or unobservable reading behavior. The limited actual-JavaScript API/event VM reports five completed routes, thirty correct first-attempt objective choices and no locked gates. None of those facts establish prose mastery, native usability or real learning gain.

### Overall verdict

S4 produced substantively strong, source-connected and numerically correct work rather than generic threshold-filling prose. The main residual weakness is causal-test feasibility and identification: several proposed experiments could conflict with the rights the same learner defends, or change more than the purported link. All-correct objective checks coexist with meaningful unfinished research-design reasoning.

### Cross module findings

1. Thirty correct first objective choices coexist with several incomplete causal designs. This is direct evidence that automated completion and recognition scores are not a sufficient semantic assessment, even for this strong AI-authored profile.
2. The learner does actual numerical reasoning and changed-location transfer in all five. It often supplies exact costs, authority and protected outcomes; there is no basis for claiming the tasks are empty or impossible from this authored record.
3. Rights are reasoning constraints for research design too. A learner can correctly defend nomination and complaint rights, then inadvertently violate them in the control condition. This is an excellent further stretching question, not grounds for making the lesson easier.
4. Matched/phased comparative designs need explicit assignment and selection assumptions. 'Hold comparable' is useful starting language but not sufficient causal identification.
5. The Everyone Says Democracy stale wrong-answer messages were not exposed by S4 because its first choices were all correct. That specific content defect remains important despite a clean successful route.
6. S4 is an AI modeled learner with unusually articulate, source-aware prose. These responses cannot establish the typical performance, effort, vocabulary difficulty or enjoyment of real medium-good students.

**Runtime**

|                         |      |
|-------------------------|------|
| Completed               | True |
| Objective first attempt | 6/6  |
| Failed gates            | 0    |

**Semantic verdict**

Strong contrast, sentence-audit and institutional-duty understanding; remaining concerns are ethical comparison feasibility and ensuring the experimental design identifies one chosen link.

**Strengths****Item 1****Location**

research and investigation.provisional/reason

**Evidence**

S4 differentiates corrected G1/G0 from G1/H0, odds from score points, stored profiles from volunteered details and whole-version ranking from live reply choice. It narrows from an upstream objective account to explicit learning-versus-volume credibility processes.

**Judgment**

Substantive source and mechanism reasoning; avoids declaring errors the cause of persuasion.

**Item 2****Location**

decision.choice/defense

**Evidence**

All six D4 classifications and deciding references are correct; 'fairest' is a value judgment. Package A sums to 120 hours, retains lawful advocacy, names five-day correction, independent clerk and small-advocate/launch burdens.

**Judgment**

Meets important deliverables that a length gate alone would not check.

**Item 3****Location**

transfer.mechanism/revision

**Evidence**

S4 computes +1 support and +0.2 knowledge offer contrasts, keeps selected finishers separate, does not attribute the new +0.4 knowledge contrast to ordering alone, and identifies inaccessible complaints after first-screen coverage repair.

**Judgment**

Genuine population and encountered-rights transfer, including appropriate uncertainty limits.

**Item 4****Location**

brief.dataReason

**Evidence**

+6 [2,10] is named as a broad attitude package and +1 [−2,4] as the direct profile contrast; no exact-zero or vote conclusion follows.

**Judgment**

Correct quantitative scope and a differentiated governance condition.

**Limitations****Item 1****Severity**

P2

**Location**

transfer.test

**Evidence**

The learner proposes randomizing chosen-language complaint access versus the existing majority-only path, although complaint usability is a defended right; it supplies a paper-access fallback but does not establish that fallback's complaint language/access.

**Issue**

A real official-service test cannot knowingly deny a usable complaint right merely to obtain a clean contrast. Nonbot information access is not necessarily accessible independent complaint access.

**Support preserving demand**

Ask for an ethically lawful design: two usable complaint formats, a nonbinding prototype, or a justified rollout with an accessible independent route for all. Require the learner to specify what causal component remains identified.

**Item 2****Severity**

P2

**Location**

transfer.prediction and brief.nextStudy

**Evidence**

Complete initial coverage is compared with the old sequence, or an accessible complete-guide offer with the current invitation. Selection reasoning predicts similar all-assigned contrasts despite large finisher means.

**Issue**

Coverage/access interventions can themselves change who starts/finishes, so an all-offer difference does not uniquely identify omission-framing rather than engagement. The broad invitation test is useful but is a package estimand, not a single-link test.

**Support preserving demand**

Require a named package estimand plus intermediate option encounter/use measures, or hold accessible invitations fixed and vary one lawful display feature. Explain how selection and framing could coexist.

**Item 3****Severity**

P3

**Location**

initial investigation.rival

**Evidence**

Support-oriented version selection is contrasted with accurate information changing beliefs.

**Issue**

An upstream selection objective can select accurate persuasive information; those are not inherently exclusive processes at one link. The later accurate-learning versus volume-credibility rival is stronger.

**Support preserving demand**

Ask the learner to specify the precise link and shared outcome in each account before testing; reward its later narrowing rather than initial certainty.

**Instructor follow up**

Redesign your complaint experiment so every household retains a usable challenge route, and state whether it tests complaint usability, engagement, knowledge or the sponsor's objective.

## P5 / Original faculty-role review

AI political scientist role: political concept formation, research-method reasoning, and teaching/transfer assessment

Independent inspection of all five latest version-3 content configurations, staged learner materials and original artwork. This is an AI analytical role, not a credentialed human scientist. I have not recruited students, measured learning gains, used a native browser or independently reread all primary papers. Source judgments concern the internal accuracy, coherence and boundaries of the explanations supplied by the builds; the citations require a separate primary-paper audit. Initial report written before inspecting learner answers or peer reports.

**Ratings (1-5 judgments)**

|                      |   |
|----------------------|---|
| Content coherence    | 5 |
| Source use           | 5 |
| Political depth      | 5 |
| Challenge            | 5 |
| Evidence sufficiency | 5 |
| Teaching potential   | 4 |

**Strengths**

1. The corrected added-profile contrast is separated from the bot-versus-human contrast. Uncertainty is not converted to exact zero, and odds are not a percentage of persuaded people.
2. Randomized packages, subsequently observed claim counts, verification and factual accuracy remain distinct. The learner must audit six sentences using actual references, not trust their linked citations.
3. Official information and disclosed advocacy have different duties while sharing factual/privacy floors. This makes political permission and democratic value more than generic suspicion of influence.
4. Westmere forces all-assigned offer effects, selected finishers, earlier interest, encountered alternatives and language-access complaints into the same bounded account.

**Problems****Item 1****Severity**

P2

**Location**

papers[0].method and contrastTable

**Evidence**

The source uses relative odds while the fictional workshop uses attitude-score points. Both are explicitly distinguished, but the odds explanation has no small numerical interpretation exercise.

**Why it matters**

A learner can memorize 'not a percentage persuaded' without understanding what the positive relative-odds contrast does mean. This is a serious conceptual gap when comparing the actual studies.

**Recommended repair**

Add an optional different-topic binary-odds illustration showing probability, odds and an odds ratio, clearly label it as an analogy rather than a reconstruction of the ordinal model. Test which units can be compared and require a source-specific contrast description, not cross-study subtraction.

**Item 2****Severity**

P2

**Location**

revision.decision.fields.defense

**Evidence**

The six factual classifications, value judgment, process chain, rival, contrast, both studies and switch pattern all occupy one text field with a 20-word recording minimum.

**Why it matters**

A learner can pass recording while never conducting the sentence audit. This hides a distinct critical-reading performance in a long policy paragraph.

**Recommended repair**

Use six learner-authored audit rows with status and deciding reference/missing evidence, followed by the value judgment and separate mechanism/test defense. Human review assesses reasons; completeness checks may count all six rows but must not grade meaning automatically.

**Item 3****Severity**

P2

**Location**

dataWorkshop workedExample and transfer T2/T4

**Evidence**

The interval exercise is carefully taught, while Westmere's  $+1/+0.2$  then  $+1/+0.4$  contrasts deliberately have no precision estimate.

**Why it matters**

Students may inadvertently transfer the earlier confidence-interval conclusions to later exact descriptive sample means, treating a changed knowledge contrast as a securely identified improvement across experiments.

**Recommended repair**

Before the final transfer test, ask the learner to mark each quantity as descriptive sample contrast or estimate with uncertainty and identify whether comparisons belong to the same experiment. Do not invent precision or claim a correction effect from the two separate studies.

**What it teaches**

Define a treatment contrast, distinguish package from component effects, interpret uncertainty, audit factual claims independently of attribution, trace incentives and encounters, preserve citizen challenge, and distinguish the effect of an offer from selected-user movement.

**What it still does not teach**

It does not identify the mechanism of opinion movement, prove intentional deception, estimate behavior or votes, measure enduring learning, or equate a rights judgment with a statistically clear capability effect.

**Does it stretch learners?**

Yes, with the highest statistical-language load. The evidence is sufficient, but many distinct performances are concentrated in two response fields. Difficult thinking is productive; difficulty finding the required audit inside a paragraph is not.

**Scaffolding verdict**

The different-topic interval example is good procedural support and does not name the live political answer. Add unit/contrast organization and a structured factual audit. Keep the correction and the uncertain effect rather than simplifying the substantive evidence.

**Quantitative verdict**

Excellent direct-contrast and threshold reasoning. Add source-unit understanding and explicit descriptive-versus-precision classification to connect the taught graph with the later offer ledger. Actual graph literacy, beyond this cued exercise, still needs a new display and human pilot.

**Design content relation**

The surreal chrome phone, cables, political images and hearing microphones clearly connect digital communication to public power. It is visually distinctive, though dramatic flowing imagery can suggest automatic powerful influence; use an optional art-versus-evidence reading prompt to question that implication.

**Overall verdict**

Rigorous and worthwhile, especially for stronger students, but it requires the most careful facilitation and structured response architecture. Its complexity should be made navigable, not diluted.

**Ranking**

1. democracy-alarm
2. overloaded-state
3. civic-machine
4. persuasion-hearing
5. everyone-democracy

**Cross-module context**

1. This ranking concerns current teaching readiness under my concept/method lens, not real enjoyment or measured efficacy. Everyone would move upward after its feedback defect is repaired.
2. All five provide sustained reading, political actors, constrained decisions and generative transfer. They are not empty multiple-choice games. Their strongest challenge resides in the authored rival, bounded positive finding, feasible decision and discriminating test.
3. The recurring weakness is compound response architecture. Split and label dimensions while retaining the whole intellectual demand. Word minima are explicitly participation gates and should stay distinct from semantic quality.
4. The hints I inspected generally direct method rather than supply political answers. Do not make them more revealing; target denominators, contrasts, units, locations and missing records. Post-submission feedback needs its own complete wrong-option audit.

- 5. The five transfers are meaningfully altered cases, but remain scaffolded near transfer. A delayed, less-cued unfamiliar dossier is needed before claiming independent generalization.
- 6. Graph/table activities are concrete and politically relevant. Claims of general graph literacy require unfamiliar encodings or mislabeled displays, not only recognition of formulas beside a worked example.
- 7. A human pilot should retain the challenging reading, record where a learner becomes stuck, separately score content/process/quantitative/transfer performances, and include a delayed new-case task. No completion percentage or AI persona rating is evidence of real learning gain.
- 8. The artwork is relevant and differentiated. It could teach political visual rhetoric through optional interpretation prompts, while remaining separate from documentary evidence. I did not assess contrast, mobile rendering, keyboard behavior or assistive technology.

P5 / Assessment of S5 submitted answers

Political concepts, teaching and transfer

Separate semantic audit written after the independent p5 review was saved. I read S5's five actual submitted response.json files and all five corresponding runtime/trials/s5/\*-first.json records. This audits AI-modeled learner texts, not human learning or enjoyment. The runtime is an API/event VM execution of original app JavaScript, not native-browser use. Each inspected first trial completed with six correct objective choices and no recorded failed gate. Those flags establish completion and keyed-choice success only; this report separately judges authored content.

Cross module conclusion

S5 produces strong, source-grounded and often genuinely generative reasoning in all five. This is useful evidence that the material contains enough information for this modeled response, not that medium-good real students will produce it unaided or enjoy it. The remaining weaknesses show why human semantic review is essential: alternative unknown facts or heterogeneous components can be mislabeled causal rivals; causal comparisons may condition on later outcomes; a weekly excess can be overstated as stock growth; a plan may omit ready-to-use items; and useful policy tests can fail to discriminate proposed mechanisms. None was exposed by the keyed 6/6 choice results. Preserve challenge and use neutral comparison/response organizers rather than give more of the answer.

Runtime

|                |      |
|----------------|------|
| Completed      | True |
| Correct checks | 6    |
| Total checks   | 6    |
| Failed gates   | 0    |

Semantic verdict

Strong claim audit, contrast interpretation and governance reasoning. The final transfer test is a valuable language-access test but not a clean rival-mechanism discriminator; some cross-experiment language should remain explicitly descriptive.

**Strengths****Item 1****Location**

research and investigation.reason

**Example**

S5 names the four opponent/profile cells, keeps the corrected direct contrast separate from the human baseline, distinguishes odds from attitude points and identifies instruction packages rather than a proven claim-count mediator.

**Judgment**

Source boundaries and units are substantively respected. P-29's two tests are not merged into a single verification-causes-persuasion finding.

**Item 2****Location**

decision.defense

**Example**

The six statuses are supported, contradicted, not established, supported, contradicted, not established, with T/F/B or missing-poll reasons. 'Fairest' receives a value criterion rather than fact status.

**Judgment**

This is a completed authored factual-reading audit, not only multiple-choice recognition. It distinguishes linked attribution from support.

**Item 3****Location**

decision.choice and transfer.revision

**Example**

Package A uses 80 + 40 hours and independent correction/complaint authority. Official alternatives/challenge duties remain distinct from disclosed campaign advocacy. T4 adds usable multilingual complaints and examines renewal incentives.

**Judgment**

Governance is feasible and rights-based without presuming all advocacy or opinion movement is illegitimate.

**Item 4****Location**

transfer.mechanism and brief.dataReason

**Example**

S5 calculates the offer contrasts +1 support and +0.2 knowledge; 100 finishers at +8 and 400 others at -0.75 can yield the 500-person +1 mean. The direct profile interval -2 to +4 does not establish a clear positive advantage or exact zero.

**Judgment**

The arithmetic, population and uncertainty are right. Selected finishing is not called randomized and no ballot effect is invented.

**Limitations****Item 1****Severity**

P2

**Location**

transfer.test

**Example**

The final test compares complaint-language access; an access account predicts better challenge use/fairness, while an objective-driven support account may persist without those changes.

**Judgment**

Those accounts concern different response dimensions and may both operate. This is a useful policy test, but it does not force competing predictions for one specified outcome or isolate the objective-versus-access mechanism. The learner partly acknowledges that it cannot isolate the renewal incentive.

**Item 2****Severity**

P2

**Location**

investigation.reason and transfer.prediction

**Example**

Knowledge-linked movement is contrasted with support movement without comparable knowledge gains; omission predicts possibly lower Standard support.

**Judgment**

A factual knowledge test does not capture all legitimate reasons or value learning, and added waiver information need not lower support. Retain uncertainty and define the proposed process more tightly; support change without factual-score change is not automatically manipulative persuasion.

**Item 3****Severity**

P3

**Location**

transfer.revision

**Example**

S5 describes knowledge improving from +0.2 to +0.4 across the old and new offer comparisons while correctly saying the cause is not isolated.

**Judgment**

These are descriptive contrasts in separate fictional randomized comparisons with no precision estimates. The wording should say the sample contrast is larger, not establish a statistically reliable knowledge improvement or a correction effect.

**Support retaining difficulty**

Require one outcome, one manipulated link and opposite or meaningfully different predictions from the two accounts. Keep a separate access-rights policy test if that is the decision goal. Ask for labels distinguishing descriptive contrasts, interval estimates and normative duties; no additional answer-revealing hint is needed.

**Mastery claim allowed**

Several demanding performances are evidenced, but a valuable governance proposal is not automatically a well-identified psychological-mechanism test. All-correct choices do not resolve that limitation.