

FRONTIERS / REVIEW RECORD

Everyone Says Democracy

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	5
Challenge	5
Evidence sufficiency	5
Teaching potential	5

Strengths

1. The linked A-D records separate calling an act democratic, approving it, reporting importance and choosing its sponsor. They include positive own-party restraint, so students are not taught that citizens invariably excuse their side.
2. C remains a real ambiguity: a strong importance report with little rating response can reflect attention, belief, implementation expectation or another link. The task does not announce that this must be indifference.
3. Cedar/Doran compares actual removal, compliance and appeal powers while holding elected appointment alike. That is much stronger institutional reasoning than declaring every elected judicial appointment undemocratic.
4. The source explanation preserves Wunsch's narrower majoritarian/judicial relationship despite failure of generalized matching, and Claassen's coherence-versus-coverage and equivalence limits. The two studies actually change the proposed measurement design.
5. The unfamiliar Sable record changes Ari's classification under a party cue, while Bea's restraint survives. This forces an update beyond a simple policy-tradeoff account.

Problems

Item 1

Severity

P2

Location

revision.decision.fields.choice / 12-slot pilot

Evidence

The plan simultaneously covers four institutions, six response targets, a paired profile costing two slots, usable response scales and two contexts. The instruction says a question may serve an overlap but does not demonstrate how to count that overlap without double-barrelled wording.

Why it matters

A novice can technically stay inside twelve slots by writing a compound question that actually collapses recognition, approval and importance - the very error the module is meant to teach.

Recommended repair

Provide a blank twelve-slot budget and a different-topic contrast between one precise question usable for two stated analytical purposes and one double-barrelled question requiring separate responses. Require each response target to be labeled; keep the conceptual tradeoff.

Item 2**Severity**

P2

Location

revision.transfer ST-03 and revisionFields.revision

Evidence

Ari changes the democratic-status classification with the sponsor, but credibility and expected implementation are unmeasured. The record makes party cueing vivid and the rubric warns against automatic motive inference, yet the term 'same action' can tempt an inference that sponsor-driven motivated reasoning has been identified.

Why it matters

Identical legal text is not necessarily identical perceived implementation. A plausible cognitive account needs to distinguish classification of authority from expectations about whether a promise will be carried out.

Recommended repair

In the next-test prompt explicitly request a separate expected-implementation or sponsor-credibility item alongside the status judgment, plus randomized/counterbalanced order. Ask what result would favor a rational credibility account versus a sponsor-dependent institutional standard; do not label Ari's motive in advance.

Item 3**Severity**

P3

Location

revision.decision.fields.defense and final brief

Evidence

The four-person linked packet and distinct 200-person table are carefully labeled but both feature Vale, recognition and choice. The same name can make their boundary harder to remember during synthesis.

Why it matters

A learner may attach the constructed four-person pathways to the 200-person proportions and claim those mechanisms were observed for the whole group.

Recommended repair

Use a persistent contrasting dataset tag - Linked interviews A-D versus Separate 200-person exercise - whenever either appears in an export, final brief reference or data panel. No new explanation of the answer is required.

What it teaches

Political meanings, support, recognition, importance and action can diverge. A measurement battery has to trade coherence against concept coverage and context interpretation. Institutions are powers and protections, not attractive names. Political restraint is an outcome worth explaining alongside tolerance.

What it still does not teach

Drafting a survey does not validate a scale; association of measured understandings with randomized profile responses does not identify a course effect; hypothetical choices do not reveal real votes or party-board decisions. The task should keep those boundaries without discouraging bounded political inference.

Does it stretch learners?

Very strongly. Designing usable questions under a finite slot budget and explaining an ambiguous person require more than the standard 'voters are hypocritical' narrative. The late transfer asks for an actual model change and a credible education test.

Scaffolding verdict

The glossary and linked-record structure are adequate. A slot counter and question-target ledger would make the measurement task more doable without supplying substantive choices. It is likely the most linguistically intricate of the five, because everyday words such as approval and important have precise analytical jobs.

Quantitative verdict

The reversed-conditional problem is good: 72/120 within recognized-disapproved respondents differs from 32/136 among Vale choosers. Linked counts permit description of overlap, not identification of a lesson's effect. The open task correctly demands two possible processes and a discriminating comparison.

Design content relation

The anime/manga civic scene and campaign posters provide a distinctive youth-facing identity without making the political task babyish. Blank speech balloons leave meanings to the reader, which fits the subject. The monumental judicial imagery risks implying that all of democracy is courts; the four-institution survey requirement counterbalances it.

Overall verdict

An excellent concept-and-choice workshop, my third choice. Its greatest risk is question-design overload, not lack of depth. I would look especially closely at whether student-created questions really separate targets rather than merely name them.

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

Strongest evidence of distinction among meanings, approval, importance and choice, including positive own-party restraint. The prospective tradeoff-versus-disbelief comparison needs a more explicit factorial design; the late revision and subsequent sponsor test are substantially better.

Source use

Accurately reconstructs Claassen's factor-analysis/battery coverage tradeoff and Wunsch's randomized candidate features versus measured understandings. Preserves the attribute-specific majoritarian/judicial relationship after generalized matching fails. Neither source is presented as a randomized education effect or real-election forecast.

Core reasoning

A and D are linked to tolerating choice and restraint; B's later classification changes narrow a stable-meaning account. C's missing attention and implementation beliefs remain open. The response explicitly proposes counterbalanced description order and expected implementation rather than assigning an unobserved motive.

Quantitative reasoning

Correct conditional $72/120=60\%$, total $136/200=68\%$, and reverse conditional $32/136\approx24\%$. Keeps disapproval distinct from rejection and proposes two plausible mechanisms without treating table association as a lesson effect.

Institutional decision

A feasible twelve-slot budget covers four institutional components, separate response targets and a two-slot matched candidate pair. It includes knowledge of actual powers and usable options, while admitting depth sacrifices and no validated cross-country ranking. Cedar/Doran powers and both parties' real amendment/nomination authorities are correctly distinguished. Some question wording remains an outline rather than a ready-to-field item.

Late record update

Narrowing the tradeoff explanation is justified by Ari's changed status classification, not just changed choice. The response retains Bea's restraint and keeps perceived credibility as an unmeasured rival. It distinguishes appointment from unlawful removal or suspended coverage.

Independent transfer

The original disbelief rival is credible and the late sponsor/order/credibility test is useful. The prospective prediction holds heating offers alike while predicting a subsidy-advantage mechanism that would need an unequal-subsidy comparison as well. A delayed unfamiliar education comparison is properly distinguished from taught recall, although a null unfamiliar choice cannot uniquely identify rehearsal.

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

strength

Severity

none

Location

transfer.revision

Quote

Ari now changes the action's democratic classification when the sponsor changes, rather than only accepting an unchanged recognized cost.

Judgment

The learner locates exactly which analytical link changes. This is independent conceptual transfer rather than copying the earlier 'tradeoff' diagnosis.

Feedback

Retain separate status, approval, expected implementation and choice measures in the next test.

Item 2**Kind**

gap

Severity

P2

Location

transfer.prediction

Quote

Hold party and heating offers alike ... simple subsidy tradeoff predicts switching mainly when the subsidy advantage disappears.

Judgment

Once heating offers are held alike, the proposed design contains no varying subsidy advantage to test the second prediction. Moreover, a policy tradeoff can itself vary with expected implementation because the institutional cost is probabilistic. The two accounts are not cleanly separated as written.

Feedback

Specify a two-by-two comparison varying subsidy advantage and credible implementation separately, with party held constant and distinct recognition/approval/choice outcomes. Predict each cell under explicitly defined accounts and acknowledge overlapping responses rather than forcing total exclusivity.

Item 3**Kind**

gap

Severity

P3

Location

decision.choice items 6-8

Quote

Item 6 asks approval on a five-point scale; 7 asks importance on a five-point scale

Judgment

The targets are named, but the exact objects and response anchors are not written as deployable questions. Context makes the intended action inferable, yet someone implementing the pilot could ask approval of Vale, democracy or the court action instead.

Feedback

Write the exact object in each item, including whether importance means intrinsic institutional importance or weight in this choice. Supply endpoint labels. Keep the twelve-slot tradeoff and acknowledge unvalidated measurement.

Item 4

Kind

precision

Severity

P3

Location

transfer.test

Quote

Improved taught recall without changed unfamiliar choices suggests rehearsal rather than stronger restraint.

Judgment

The word 'suggests' is appropriately bounded. Such a pattern is compatible with rehearsal, but also insufficient course intensity, weak transfer, preexisting ceiling or noisy choice measurement; it does not identify rehearsal uniquely.

Feedback

State one additional rival to rehearsal and a check that distinguishes it, while retaining the delayed unfamiliar test.

Human-review judgment proposed by the AI role

Credit the construct distinctions, positive restraint, precise late-record link, feasible measurement budget and denominators. Require a sharper prospective factorial prediction and fully worded question objects before calling the measurement plan operationally complete.

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	5
Challenge	5
Evidence sufficiency	5
Teaching potential	5

Strengths

1. Support for a label, exact-action classification, approval, importance, satisfaction and candidate choice are kept distinct. A and D show institutional restraints operating, not merely voters tolerating violations.
2. The current source explanation distinguishes scale coherence from coverage and includes the narrower judicial association surviving the failed generalized congruence account. It does not infer that measured understandings were randomized.
3. The linked table supports both forward and reverse conditional rates: Vale choice among disapproving recognizers and nonrecognition among Vale choosers. The 12-slot instrument and Sable transfer demand original construction, context-sensitive legal reasoning and an education-effect design.

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

P2

Location

dataWorkshop.reasonPrompt and checks.data-infer

Evidence

The table permits $72/120=60\%$, $136/200=68\%$, and $32/136\approx 23.5\%$. The guide emphasizes the row rate and whole sample; the second item reverses the starting group to all choosers.

Why it matters

This is an excellent test of inverse conditioning, but an accurate selected fraction may not show the student can justify the conditioning group or explain why 32/40 answers a different political question.

Recommended repair

Add a blank 'among whom?' sentence and ask for a plain-language contrast between $P(\text{choice} \mid \text{nonrecognition})$ and $P(\text{nonrecognition} \mid \text{choice})$, without requiring notation. Require naming the appropriate denominator and explaining why the other rate cannot test the headline. Preserve the difficulty.

Item 2**Severity**

P2

Location

revision.decision.fields.choice: 12-question pilot

Evidence

The pilot must cover four institutional domains and six distinct targets, include an otherwise comparable pair costing two slots, and address cognitive interviews across two contexts.

Why it matters

This is the suite's most demanding instrument-design task. A response can look polished while silently measuring approval where it claims recognition, or exceeding the slot budget. A 20-word recording minimum cannot distinguish these errors.

Recommended repair

Use a numbered blank instrument grid with item wording, response options, target, institution and cost. Display a neutral slot counter and require a crosswalk plus defended omission. A human should check the response target and context meaning; do not automatically claim a validated scale.

Item 3**Severity**

P2

Location

revision.transfer.lateRecord ST-03 and revisionFields.test

Evidence

The same action is classified differently by sponsor, but perceived credibility, implementation belief, order and durability are unmeasured. The task then asks for both a remaining-rival test and a course-effect test.

Why it matters

A learner may overdiagnose motivated reasoning from the label change, or turn a randomized lesson into proof of changed real voting. The current prose warns against both, but these are the most important conceptual errors to actively assess.

Recommended repair

Have the learner choose two competing interpretations of the sponsor pattern in their own words and specify the response they predict when exact legal powers, perceived credibility and implementation belief are measured separately. Keep the delayed unfamiliar choice in the education design and explicitly review knowledge gain versus durable institutional commitment.

What it teaches

Construct validity and measurement coverage, institutional protections rather than party labels, forward/reverse conditional probabilities, meaning versus tradeoff pathways, own-party restraints, non-randomized attitude moderators and a bounded intervention design.

What it still does not teach

It does not validate the newly drafted battery, demonstrate cross-context measurement equivalence or prove that education changes actual electoral behavior. Full scale estimation and psychometric diagnostics are introduced conceptually rather than performed.

Does it stretch learners?

Very strongly. Drafting an original constrained instrument and distinguishing classification change from choice change under sponsor/policy pressure require synthesis. This is my first choice for depth of measurement learning, not necessarily the easiest or most popular module.

Scaffolding verdict

Doable for a persistent capable learner with sufficient time and teacher review. Definitions of liberal, non-liberal and executive are useful and not covert answers. Add an instrument grid and conditional-group organizer; avoid a prewritten battery or mechanism menu.

Quantitative verdict

The linked dataset is coherent and the reversal of the denominator is genuinely discriminating. Rates are 60% within recognized/disapproved, 68% overall Vale choice and about 24% nonrecognizers among choosers. Neither an association nor a hypothetical choice establishes the effect of a lesson. This could become the best assessment of construct-plus-statistical reasoning if the instrument is semantically reviewed.

Design content relation

The manga/pop-art college debate with courthouse, ballots and expressive learners is distinct and fits interpretive disagreement. Blank speech bubbles do not insert a preferred conclusion. Court imagery can make judicial protection appear self-evidently beneficial; the content's distinction between appointment and removable/dependent judges should remain central. Real Gen Z appeal was not measured.

Overall verdict

The deepest combined political-measurement and student-generated assessment task. Retain the intellectual demand, make omissions diagnosable and teach how instrument construction differs from instrument validation.

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

The strongest S2 measurement product: a genuine feasible original instrument, correct forward/reverse conditional rates and a sophisticated rival-aware transfer. Two item targets need repair rather than easier questions.

Strengths

Item 1

Location

decision.choice

Evidence

S2 supplies ten numbered single questions plus one two-slot candidate pair, distinct response scales, four institutional areas and a cognitive-interview plan, with an explicit coverage sacrifice and no validation claim.

Assessment

The 12-slot budget is honored. Unlike generic prose, this is an actual assessable student-generated instrument.

Item 2

Location

brief.dataReason

Evidence

S2 computes $72/120=60\%$, $(72+32+32)/200=68\%$, and $32/136\approx 23.5\%$, explicitly rejecting 80% as the reversed rate.

Assessment

Conditional bases and interpretation are correct; recognition and disapproval are not substituted for choice. No arithmetic error was found.

Item 3

Location

investigation.reason; transfer.revision and test

Evidence

S2 distinguishes A's tradeoff, D's restraint, C's unmeasured attention and B/Ari's sponsor-sensitive classification. It proposes measuring implementation belief/credibility and randomizing sponsor/order, then a course comparison with delayed unfamiliar choice.

Assessment

It does not assert the sponsor pattern proves internal bias or that a course already changes votes. The additional knowledge-versus-willingness distinction is substantive.

Weaknesses**Item 1****Severity**

P2

Location

decision.choice, item 4

Quote

Whether equal voting and legal rights are essential.

Assessment

This combines voting eligibility and equal legal treatment in one response. A respondent supporting one but not the other cannot express that distinction; the apparent four-domain coverage therefore sacrifices more precision than merely one item per domain.

Improvement

Ask S2 to defend this double-barreled compromise explicitly or spend a slot splitting it and explain which other target loses depth. Require a target-to-item crosswalk rather than adding an uncoded question.

Item 2**Severity**

P2

Location

decision.choice, item 8

Quote

Rate satisfaction with present government on five points.

Assessment

This is a government-performance satisfaction item, not necessarily satisfaction with how democracy works. It may be a legitimate distinct target, but without labeling that substitution it can silently conflate government popularity with satisfaction in the institutional system the source distinguishes from support.

Improvement

Have S2 name the intended satisfaction construct and rewrite the item accordingly, or retain government satisfaction while explaining what democratic-system satisfaction remains unmeasured. Keep approval, importance and choice separate.

Item 3**Severity**

P3

Location

transfer.test

Quote

A belief account predicts status differences shrink when verified powers and likelihood are equal; selective standards may persist.

Assessment

This is a good discriminating prediction, but a written assurance of equal powers does not guarantee that respondents believe equal implementation likelihood. The response appropriately proposes measuring expectation, yet does not describe checks for whether the intended belief manipulation succeeds.

Improvement

Require a comprehension/belief check and a bounded interpretation if the assignment fails to equalize perceived credibility. Preserve the delayed unfamiliar choice rather than rewarding only repeated legal facts.

Completed versus mastered

This product provides unusually good evidence of generative measurement reasoning, but completion hides a double-barreled item and a possible satisfaction-construct substitution. Correct MCs do not certify an original scale's target validity or equivalence.

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. Recognition, approval, self-reported importance and choice are separately measured rather than collapsing all support for Vale into authoritarian demand.
2. A and D provide tolerance and genuine own-party restraint; B and C prevent one mechanism from explaining every respondent.
3. The source explanation correctly separates randomized candidate attributes from nonrandomized democratic understandings and treats generalized congruence differently from a judicial-attribute

association.

4. The 12-slot measurement design has a meaningful coverage/coherence tradeoff, context-sensitive wording and a constrained elite-response hypothesis.

5. Sable transfer requires revision after a changed classification under a preferred sponsor, while retaining credibility and order as alternatives.

Problems

Item 1

Severity

P2

Location

revision.decision.fields.choice

Evidence

The 12-slot plan must cover four institutions, six response targets, cross-context interpretation and a matched pair, with each pair costing two slots.

Why it matters

This is a strong authentic design task, but an apparently complete essay can omit usable response options, exceed the budget or collapse recognition and approval.

Recommended repair

Provide a blank numbered 12-slot editor with target, exact wording, response options and cost columns. Require students to mark omissions and defend them. Do not generate their items or tell them which constructs to sacrifice.

Item 2

Severity

P2

Location

ED-05 and ST-03 sponsor effects

Evidence

The legal action is fixed while the sponsor changes, but credibility, anticipated implementation and ordering are unmeasured.

Why it matters

The very engaging pattern can tempt a moralized diagnosis of 'hypocrisy' or automatic motivated reasoning, replacing inquiry with a label.

Recommended repair

Add a teacher checkpoint requiring two distinguishable same-link accounts of the changed status judgment and predicted responses with recall/implementation-belief measurement and counterbalanced sponsor order. Keep the motive unresolved.

Item 3**Severity**

P3

Location

revision.transfer.revisionFields.test

Evidence

A credible education test requires both competing change mechanisms and delayed unfamiliar choice under pressure.

Why it matters

A simple pre/post knowledge test may be proposed even though the task demands willingness to protect a constraint versus rehearsed responses.

Recommended repair

Use a blank assignment/outcome/time table with baseline, immediate measurement and delayed new institution; require one non-equivalent mechanism prediction. No lesson effect should be supplied.

What it teaches

Democratic meanings and commitments; within-person tolerance versus demand; political tradeoffs and restraints; measurement validity and equivalence; sponsorship cues and causal alternatives; what education evidence would require.

What it still does not teach

It does not demonstrate that a course changes votes or that every sponsor-dependent answer is motivated reasoning. It also does not validate the student's 12-question instrument or its cross-country ranking.

Does it stretch learners?

Very strongly. The learner must explain restraint as well as tolerance, build a feasible measure, update a mechanism and design an intervention test without claiming the sources establish its result.

Scaffolding verdict

The glossary is unusually important and includes the nonpartisan meaning of liberal. Preserve interpretive uncertainty; scaffold the item budget and repeated-observation ledger.

Quantitative verdict

The linked table makes recognition insufficient for rejection. Conditional and whole-sample bases are clear; the required numerical explanation should separately name the 120-person subgroup and 200-person total.

Design content relation

Manga/pop-art debate, campaign portraits and courthouse create a persuasive young-adult discussion atmosphere without labelling a respondent good or bad. Empty speech bubbles invite thought rather than embed answers. Do not equate this aesthetic judgment with evidence of Gen Z preference.

Overall verdict

My first choice for a political-behavior class: vivid, source-conscious and genuinely demanding. It needs a stronger inspectable survey-design scaffold, not simpler politics.

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 explanation of recognition, tolerance, restraint and sponsor-dependent classification; the survey budget is counted correctly but its final design is not yet fully usable as written.

Strengths

Item 1

Location

research.sourceB and investigation.provisional

Evidence

The response preserves the narrower majoritarian judicial association while rejecting generalized congruence, and differentiates A, B, C and D without assigning C a motive.

Judgment

This is conceptually accurate and more demanding than equating Vale choice with authoritarian demand.

Item 2

Location

investigation.revisedClaim/reason

Evidence

It identifies B's changed classification under a party sponsor rather than calling the entire pattern an unchanged disapproval tradeoff. Credibility, implementation belief and sequence remain alternatives.

Judgment

The late record changes a specific link of the account rather than simply adding a caution sentence.

Item 3

Location

decision.defense

Evidence

Cedar's independent removal appeal and fixed terms are contrasted with Doran's policy-based removal and same-minister appeal; elected appointment alone is not treated as the violation.

Judgment

The political institution is reconstructed accurately, including protection of election losers and the party-board constraint.

Item 4**Location**

brief.dataReason

Evidence

72/120=60% choose Vale among recognizers who disapprove; 136/200=68% choose overall; 32/136≈24% of choosers are nonrecognizers.

Judgment

These conditional and total calculations are correct and support the recognition-is-not-rejection claim without a causal education assertion.

Item 5**Location**

transfer.revision

Evidence

It narrows the subsidy account after Ari's sponsor-dependent classification and retains Bea's cross-party rejection.

Judgment

The transfer explains restraint and tolerance together and keeps legal appointment separate from removal authority.

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

P2

Location

decision.choice 12-slot instrument

Evidence

Slots 9 and 11-12 are described as 'recall which legal power changed with a multiple-choice item' and 'choose one otherwise matched candidate pair differing only in final-ruling compliance'; no actual recall options or full candidate descriptions are supplied.

Judgment

The slot total is feasible and response targets are separated, but these are placeholders for usable questions, not a completed survey instrument. The first four items are also topic shorthand rather than fully written respondent-facing questions.

Support without answer

Provide the blank 12-slot editor and require exact wording, every option and both profile texts for the learner's own chosen plan. Mark unfinished items as draft rather than treating the field's word count as a complete design.

Item 2**Severity**

P2

Location

transfer.test education comparison

Evidence

The course versus equal-length unrelated lesson measures knowledge separately from support and includes delayed unfamiliar choice, but does not state distinct expected patterns for knowledge gain, willingness to protect constraints and rehearsed socially desirable answers.

Judgment

The design is promising but the requested two-mechanism discriminating predictions are incomplete. Knowledge and support measures alone do not identify why an apparent lesson effect occurs.

Support without answer

Ask the learner to choose two change mechanisms, predict their different immediate and delayed patterns, and name one measure that tests each without supplying the course result.

Item 3**Severity**

P3

Location

transfer.test sponsor comparison

Evidence

The response predicts a sponsor pattern shrinking when credible implementation information is equalized, but mainly specifies sponsor/order randomization.

Judgment

Equalizing credibility is itself an intervention whose assignment should be explicit; merely conditioning on reported expectations would not isolate it and can alter the comparison.

Support without answer

Ask what information is experimentally varied and what is merely measured, then require the planned factor or matched condition to be named.

Quantitative assessment

No arithmetic error. The conditional nonrecognizer proportion is a useful additional calculation and is not mislabeled a population estimate.

Priority fix

Do not award the instrument-design criterion full credit until actual question and profile wording exists; require explicit rival learning-mechanism predictions.

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	4
Source use	5
Political depth	5
Challenge	5
Evidence sufficiency	4
Teaching potential	4

Strengths

1. Recognition, approval, importance, satisfaction and choice are separated; a voter tolerating a violation is not automatically ignorant or demanding every violation.
2. The Claassen coverage/coherence distinction and absence of validated cross-context ranking are clear. The Wunsch source differentiates measured understandings from randomized candidate profiles and preserves the narrower judicial result.
3. Cedar/Doran separates judicial appointment from dismissal, appeals and compliance. An elected appointment is not intrinsically equated with executive impunity.
4. The 12-slot design task forces usable question writing, a profile cost, concept coverage and a tradeoff. It is substantially more than repeating a democracy definition.
5. The media transfer requires adapting the model to appointment, protected reporting and policy pressure; its late record distinguishes changed classification from stable classification followed by a changed choice.

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

P1

Location

Incorrect-choice feedback in checks.comparison option b; scope option c; transfer option b; ruling option a

Evidence

Four wrong options retain affirmative feedback. For example comparison b says dropped election/equal-rights items are irrelevant because they fail to track court items, yet feedback says 'This preserves the measurement contribution'. Transfer b proposes only a broad democracy question, yet feedback says 'This treats the course as a promising intervention to evaluate'. Ruling a guarantees a party will withdraw its nomination, yet its feedback says 'This uses the measurement contribution and the commitment-choice finding'.

Why it matters

The check keys reject these responses but the explanation appears to endorse their very inferential mistake. A real learner answering incorrectly may leave more confused even if the current AI trials happened to select all correct choices.

Recommended repair

Replace each of these four feedback strings with an explanation of the specific missing coverage, nonidentified motive, inadequate intervention measure or unobserved elite decision. Add a regression test submitting EVERY wrong option and checking that feedback does not praise or validate its unsupported claim. Preserve method-oriented pre-answer hints.

Item 2**Severity**

P2

Location

12-slot survey plan field

Evidence

The learner must write response-ready numbered items covering four institutions, six distinct targets, a matched pair, cross-context adaptation, interview tests and a coverage sacrifice in one field.

Why it matters

A confident paragraph about institutions is not a usable instrument. Slot arithmetic and response scales can disappear behind general prose.

Recommended repair

Add a blank numbered slot ledger with target, wording, options, context and cost columns. Leave all content to the learner. Require a total and a coverage map before instructor approval; do not give a ready-made survey.

Item 3**Severity**

P3

Location

Course-test transfer

Evidence

The prompt correctly asks delayed unfamiliar choice, but the fictional charter rights make the legal status relatively explicit.

Why it matters

Success can demonstrate applying supplied rules without showing whether a learner can distinguish criticism, lawful reform and suspension in a borderline case.

Recommended repair

Add an optional fresh borderline vignette about lawful editor criticism or ordinary parliamentary budget changes. Require a record-specific difference from executive suspension; retain the nonpartisan equal standard.

What it teaches

A broad democracy label does not reveal concrete institutional commitments or candidate motives. Measurement needs coverage and context; government appointment does not erase independent review, protected removal or compliance duties. Political choice and elite response are separate inferential chains.

What it still does not teach

It does not validate the student's new battery, estimate a course's voting effect, determine every citizen's motives or experimentally identify democratic understandings as causes. It gives a practicable introduction to survey design rather than full psychometrics.

Does it stretch learners?

Very much: producing an actual instrument and differentiating competing processes from linked records are demanding. Repair the wrong-answer feedback before a learner-facing pilot.

Scaffolding verdict

The method hints are not answer dumps, but the four inconsistent post-answer messages are a real instructional defect. A blank survey organizer preserves difficulty while making the deliverable legible.

Quantitative verdict

Strong conditional-base reversal: 72/120 differs from 72/200, while 32/136 directly tests the 'most choosers failed to recognize' headline. No observed rate is automatically a course cure.

Design content relation

The political manga/collage visual establishes youth dialogue and competing candidate images without supplying an answer. It fits the case, although its apparent characters should not be presented as actual respondents or real preference evidence.

Overall verdict

I liked its political-behavior puzzle and genuine survey work. Content is deep enough, but the four misleading wrong-choice explanations are a release-blocking instructional fix; afterward it is a strong facilitated pilot candidate.

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 usable survey effort, linked-response distinctions and unfamiliar transfer. Some initial rival framing and survey wording need greater precision; no major numerical or institutional error was found.

Strengths

Item 1

Location

decision.choice

Evidence

S4 actually supplies twelve slots, five institutional understanding items, distinct recognition/approval/importance/satisfaction/expectation targets and a two-slot matched candidate pair. It explains interview paraphrasing and published limits.

Judgment

Real deliverable rather than a generic survey recommendation; retains the elected-decision component despite a coherence temptation.

Item 2

Location

investigation.provisional/revisedClaim

Evidence

A fits housing-sensitive tolerance, D demonstrates restraint at a cost, B's sponsor-sensitive status requires a revised account, and C remains ambiguous without attention/expectation data.

Judgment

Political behavior is not overdiagnosed from one choice or a general label.

Item 3**Location**

decision.defense

Evidence

The learner identifies fixed tenure/independent removal appeal versus policy-based dismissal/same-minister appeal, specifies Unity's board power and Harbor's platform-only alternative, and requires actual board minutes/amendment beyond survey responses.

Judgment

Institutions and elite incentives are active mechanisms, not backdrop.

Item 4**Location**

transfer and brief.dataReason

Evidence

Ari's changed classification narrows a pure tradeoff account; credibility and ordering remain rivals. The course proposal uses assignment, delayed unfamiliar rights choices and separate knowledge/commitment measures. Calculations are 60%, 68% and about 23.5% with correct bases.

Judgment

Strong semantic transfer and quantitative reasoning; does not call hypothetical choice an observed vote.

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

P2

Location

initial investigation.rival

Evidence

The first mechanism concerns 'some' recognized-disapproving citizens; the rival concerns other citizens regarding override as legitimate majority rule.

Issue

These can describe different respondent types simultaneously rather than rival causes of one person's same choice. Later B and Ari revisions are considerably sharper same-person tests.

Support preserving demand

Ask which named person's outcome both initial accounts explain and what observation would separate them. Preserve heterogeneous population accounts, but do not score their mere coexistence as a discriminating rival.

Item 2**Severity**

P3

Location

decision.choice item 9 and response specifications

Evidence

The satisfaction item asks about 'the current system'; five-point scales have endpoints but not every intermediate response label or legal vignette wording is written out.

Issue

This is a plausible plan rather than a field-ready questionnaire. 'System' can be heard as government performance rather than democratic operation, and target ambiguity weakens the carefully separated concepts.

Support preserving demand

Have the learner write the exact satisfaction target and complete respondent-facing text/options, then use the cognitive interviews to test interpretations rather than assuming equivalence.

Item 3**Severity**

P3

Location

transfer.test course trial

Evidence

A powers lesson is assigned against an equally engaging neutral lesson, with legal knowledge and delayed choice evidence.

Issue

A difference may also arise from salience or priming, not merely increased understanding. The answer identifies rehearsal but could say more about this competing pathway.

Support preserving demand

Require one matched-salience comparison or intermediate measure and a predicted pattern differentiating knowledge transfer from momentary issue salience.

Instructor follow up

Choose one respondent and formulate two causes of that respondent's same choice - not two different voter types - then write the precise comparison that produces different predictions.

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	4
Source use	5
Political depth	5
Challenge	5
Evidence sufficiency	5
Teaching potential	4

Strengths

1. Recognition, approval, reported importance and candidate choice are separated. A and D show different weights on an institutional constraint, B adds stated approval and C remains unresolved rather than receiving a convenient motive.
2. The source account separates covariance from concept coverage and assigned candidate attributes from measured understandings. The narrower judicial relationship is not erased by a failed broad matching claim.
3. Cedar and Doran make political power, removal, appeal and binding compliance central. Elected appointment alone is not treated as a universal democracy test.
4. The 12-slot plan forces students to author measures and defend omissions. The Sable transfer changes institution and then changes an observed classification, requiring more than repeating a tradeoff explanation.

Problems

Item 1

Severity

P1

Location

checks.comparison.options[b].feedback; scope[c].feedback; transfer[b].feedback; ruling[a].feedback

Evidence

Incorrect comparison b says dropped low-correlation rights are irrelevant but receives 'This preserves the measurement contribution.' Incorrect scope c infers active demand from weak penalties but receives 'This captures the supported commitment account.' Incorrect transfer b uses only the democracy label yet receives a positive course-comparison response. Incorrect ruling a guarantees a party nomination change yet receives an affirming bounded-inference response.

Why it matters

Even when the engine flags these options wrong, the prose can validate the exact misconception it should diagnose. This is an instructional reliability defect, not a reason to lower challenge.

Recommended repair

Audit every answer-feedback pair. Replace these four with specific post-submission questions/explanations naming the mistaken substitution and the missing observation. Retain the keyed alternatives and pre-answer hint restraint. Add a test that submits every wrong option and verifies feedback does not affirm its claim.

Item 2**Severity**

P2

Location

revision.decision.fields.choice

Evidence

A single field requires a numbered 12-slot instrument with response options, four institutional areas, six measurement targets, a matched pair, context adaptation, cognitive interviews and a coverage sacrifice.

Why it matters

The task is an authentic stretch, but it is closer to a small survey-design assignment than a 20-word response. Its bookkeeping can obscure whether the student understands the measurement target.

Recommended repair

Provide a blank numbered instrument editor with automatic slot totaling and target tags selected by the learner. Require a separate justification of overlaps and omissions. Do not supply the actual questions or select the target mapping for them.

Item 3**Severity**

P2

Location

revision.transfer late ST-03 and initial research-to-transfer assessment

Evidence

The scaffold carefully directs the learner to sponsor changes, credibility, expected implementation and sequence. This is useful but makes the transfer still highly cued.

Why it matters

A correct answer here can be guided near transfer; it does not establish the learner independently notices those missing measures in a fresh political conflict.

Recommended repair

After feedback, add a short optional or instructor-assigned delayed case with raw item wording and no supplied explanation categories. Require the learner to choose the measures and justify one discriminating comparison before a cue is shown.

What it teaches

Operationalize a political concept and respect its institutional coverage; distinguish disagreement about meaning from tolerating a known cost; understand that rights and ordinary majority decisions coexist; separate survey responses from party actions and legal powers.

What it still does not teach

It cannot prove motivation from a sponsor response, validate cross-context equivalence by drafting items, establish a course effect, or forecast actual votes. A fluent 12-question list is not automatically a valid battery.

Does it stretch learners?

Very strongly in measure authoring and integrated actor reasoning. The scenario is not merely 'people say democracy but are hypocrites.' The positive restraint example is essential and should remain.

Scaffolding verdict

Concept definitions and raw linked files make it doable. Repair the post-answer feedback before classroom use; then scaffold response organization without answering the political interpretation for students.

Quantitative verdict

Strong conditional-base reversal: 72/120 and 32/136 answer different questions. Separate four-person and 200-person records are clearly labeled. Add an authored classification or rate calculation before choices if independent numeracy is an assessment claim.

Design content relation

The manga campaign posters and young readers discussing documents give institutional measurement a human identity. Empty speech bubbles could become an optional learner-authored art-reading feature, not explanatory answer text. The court monument should not substitute aesthetic reverence for a defended institutional rule.

Overall verdict

Richest measurement-authoring assignment, but the mismatched wrong-answer feedback is a must-fix. Once repaired and organized it has high teaching potential; it should not yet be certified merely because learners can complete the keyed checks.

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

Very strong distinction of political judgments and choices; the instrument plan is promising but not fully ready-to-field, and the initial two accounts are population components rather than strict causal competitors.

Strengths

Item 1

Location

research and investigation.provisional/reason

Example

The answer distinguishes coherent measurement from coverage, measured understandings from randomized candidate features, A's tradeoff, B's approval, D's costly restraint and C's unresolved attention/belief link.

Judgment

This prevents generic hypocrisy/ignorance claims and preserves a positive democratic-restraint pathway. ED-05 changes a specific status judgment, not merely the final choice.

Item 2

Location

decision.choice

Example

The answer numbers eight single items, one two-slot candidate pair and two further items, totaling 12. It covers courts, elections, public media and equal rights and explicitly sacrifices multiple indicators per component.

Judgment

Concept coverage and resource accounting are substantially fulfilled. Satisfaction, status recognition, approval, importance, legal knowledge and expected implementation have separate jobs.

Item 3

Location

decision.defense and transfer

Example

Cedar's removal/appeal/compliance powers differ from Doran's; Unity and Harbor have different possible elite actions. In Sable, sponsor-conditioned classification narrows a stable-recognition tradeoff account while Bea's restraint survives.

Judgment

Actual institutional powers, election losers and additional party-decision evidence are analyzed. The unfamiliar case receives a reasoned update rather than the original answer repeated.

Item 4

Location

brief.dataReason

Example

S5 calculates $72/120 = 60\%$ and $136/200 = 68\%$, then offers benefit-weight and expected-implementation processes with a proposed matched comparison.

Judgment

The bases are correct and choosing Vale is not confused with approving the breach. It does not infer a course cure or an actual vote effect.

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

P2

Location

investigation.initialClaim/rival

Example

The first process describes some recognizers making a tradeoff; the rival describes other respondents approving the action.

Judgment

These can coexist in a heterogeneous population and do not strictly compete as causes of the same person's choice. The later files correctly integrate them, but initial rival-generation quality is less strong than the completion flag implies.

Item 2**Severity**

P2

Location

decision.choice

Example

Question 11 has 'options supplied from the charter' rather than actually authored options; the candidate pair is specified structurally rather than as two ready-to-display profiles.

Judgment

This is a feasible research plan, not a complete usable instrument. Five-point essential scales are described, but their labels are also not fully enumerated. Recording is complete while a teacher should mark ready-to-field item construction as partial.

Item 3**Severity**

P3

Location

transfer.test

Example

Party-conditioned classification is distinguished from implementation beliefs by checking equal recall and beliefs.

Judgment

Those measures help, but they are themselves potentially affected by sponsor labels and imperfectly measured. The proposed randomized sponsor test identifies the sponsor contrast more readily than its psychological mediator. Do not infer a pure mechanism from equal reported covariates.

Support retaining difficulty

Ask the learner to pick one named response/outcome when comparing mechanisms, write every actual item/option/profile and identify one mechanism measure that remains fallible. Use a blank12-slot editor; do not supply the questions. The independent wrong-option feedback defect was not encountered by S5's all-correct path.

Mastery claim allowed

The text demonstrates advanced supplied-case distinction and bounded transfer, but not a validated survey battery or full mastery of strict rival discrimination.