



Instructor's Guide to Cognitive Reflection

Laboratory implementation, theoretical foundations,
experimental evidence, and variants

Behavioral and Experimental Economics Research Lab
Department of Economics



Revised edition | October 2026

<https://beer-lab.org/>

How to use this guide

This guide supports preparation and discussion of the BEER Lab Cognitive Reflection Test in economics and behavioral science courses. Its teaching objective is to distinguish a recorded answer, a score and an interpretation of reasoning.

The operational section supports rehearsal. The foundations explain units, error classification and familiarity. Worked cases support discussion of complete and incomplete records, risky choices and claims that the data cannot establish.

Both three- and seven-problem forms are covered. Situations preserve published numerical structures and keys, with independent teaching wording and a Spanish translation that is not claimed to be validated. Additional extended activities remain separate from CRT scores.

Author

This guide was prepared by Stanislao Maldonado, research professor in the Department of Economics at Tecnológico de Monterrey, Monterrey Campus. For questions or suggestions about its content, please contact the author at stanislao.maldonado@tec.mx.

For the instructor

Rehearse with synthetic data, confirm language and edition, and close the access-code panel before projection. Preserve form, language and key. Distinguish correct, intuitive error, other error, omitted and pending. State N and missing-data rules before interpreting differences.

Contents

How to use this guide	i
1 The activity and its administration	1
1.1 Institutional navigation and room selection	1
1.2 Local rehearsal and participant links	2
1.3 Instructor credentials and activity access code	2
1.4 The private report before disclosure	2
1.5 Lesson planning and pedagogical closing	4
1.6 Instructions, comprehension and response entry	4
1.7 Familiarity and the extended activity	8
1.8 Reading the report and checking downloads	8
2 What cognitive reflection scores can tell us	14
2.1 A response, a process and a construct	14
2.2 Why the tempting answer is useful	14
2.3 Three problems and seven problems	15
2.4 Units are part of the answer	15
2.5 Missing data and complete-battery counts	15
2.6 Familiarity, numeracy and language	16
2.7 What the original studies motivate	16
2.8 Descriptive uncertainty and small classes	16
3 Seven fully worked problem structures	18
3.1 Prices and a difference	18
3.2 Parallel production	18
3.3 Doubling and the previous day	18
3.4 Combining rates	19
3.5 Ranks and double counting	19
3.6 Trading profit across transactions	19
3.7 Successive percentages and the categorical answer	20
4 Worked cases for data and economic interpretation	21
5 References and provenance	28

List of Figures

1.1	Four rooms separate language and edition.	2
1.2	The basic Spanish room offers only its corresponding configuration.	3
1.3	Demo shows four configurations isolated to this family.	3
1.4	Neutral institutional welcome and private entry.	5
1.5	Instructions explain resources and response format.	6
1.6	Procedural feedback retains the selected choice.	6
1.7	The seventh problem retains a categorical response.	7
1.8	Excerpt showing the first two familiarity declarations; the full form asks about all seven problems.	9
1.9	Additional hypothetical decisions remain separate from CRT counts.	10
1.10	Closing confirms recording while feedback remains hidden.	11
1.11	Private extended feedback requires completion and current disclosure.	11
1.12	The report initially hides outcome aggregates.	12
1.13	Counts retain explicit maxima and valid-observation denominators.	13
1.14	Histogram frequencies sum to the complete-battery N.	13

The activity and its administration

Purpose	Teach cognitive reflection, intuitive errors and limits of inference.
Experimental unit	Person; individual activity without matching or waiting.
Basic edition	Three-problem CRT, preserving Frederick (2005) numerical structures.
Extended edition	Seven-problem CRT from Toplak and colleagues (2014), plus separate hypothetical decisions.
Decision	Numeric response in the stated unit; categorical seventh problem; explicit omission available.
Outcome	Correct counts 0-3 or 0-7; intuitive, other and omitted counts separate. Pending problems prevent complete counts. No payments.
Topics	cognitive reflection; intuitive responses; numerical reasoning; decisions under risk; measurement and familiarity
JEL	A22, C91, D81, D91
Participants	Any positive number; one person per interface.
Suggested duration	Allow 5-10 minutes for basic and 12-20 for extended, plus discussion. Pilot locally; no automatic limit.

<https://beer-lab.org/>

<https://teaching.beer-lab.org/>

https://crt.beer-lab.org/room/crt_en

<http://localhost:9174/demo>

1.1 Institutional navigation and room selection

Institutional access begins at <https://beer-lab.org/> and continues through its teaching section to <https://teaching.beer-lab.org/>. Once published, the family card identifies its title, topics and editions. Proposed rooms are crt_es and crt_en for the basic editions, and crt_extendido and crt_extended for the extended editions. A proposed public address does not establish that production deployment is already complete. Check the host and room before distributing instructions.

This family's instance should contain four configurations in Demo and four rooms in Rooms. Each room selects one configuration with the corresponding language and edition. A basic room must

not default to an extended configuration. If unrelated games appear, a selector is empty or the language is incorrect, stop distributing links and check the instance. This small operational check prevents substantial confusion between cohorts, forms and databases.

Rooms

Current rooms:

CRT-3 · Español · básica
CRT-3 · English · basic
CRT-7 · Español · extendida
CRT-7 · English · extended

Figure 1.1: Four rooms separate language and edition.

1.2 Local rehearsal and participant links

The delivery preview uses <http://localhost:9174/demo> and works only on the computer running that instance. The intended host is <https://crt.beer-lab.org/>; access from other devices requires the corresponding deployment. A room address is a shared entry route, an individual link identifies one position within a session, and a report is a private instructor page. These addresses are not interchangeable. Giving two people the same individual link mixes their record even if storage itself works correctly.

Prepare a rehearsal with demonstration responses separately from the classroom session. Check language, problem count, figures and closing from a participant position. Then create a new session for the expected attendance and record its identifier. Any positive number of participants is admissible. Using three positions for a rehearsal helps check a complete record, a partial record and an untouched position; it does not establish an experimental group size.

1.3 Instructor credentials and activity access code

Use the institutional administrator credentials and keep them outside materials distributed to students. Each session generates an activity access code, available inside an expandable report panel. This code differs from the administrator password and technical session identifier. Share it with the intended cohort and close the panel before projection. The code restricts operational entry; it does not make individual links appropriate for public circulation.

The welcome requires the code and acceptance of general laboratory rules. An incorrect code produces an explanation and allows correction. After successful entry, the application clears the field and records access. Someone who has not entered remains pending rather than having zero correct answers. Preserve this distinction when checking progress or downloading responses. The activity code must not appear in response exports.

1.4 The private report before disclosure

The report shows progress while keeping outcome statistics hidden until the instructor reveals them. First confirm session, language, edition and participant count. Close the access-code panel

Room: CRT-3 · Español · básica

Create a new session

Session Config

CRT-3 · Español · básica ▾

Number of participants

Decisiones individuales: se admite cualquier número positivo de participantes.

Create

[Configure session](#) ⚙️

App sequence

crt_assessment (7 rounds)

CRT-3 / CRT-7: correct, prespecified intuitive and other responses; prior exposure; separate extended decisions.

0 participants present

Start links

These URLs will stay constant for new sessions, even if the database is recreated.

Room-wide URL

Here is the room-wide URL anyone can use. Don't use this link when testing in multiple tabs on the same browser, because all tabs will be assigned to the same participant, using a cookie.

http://localhost:9174/room/crt_es

Figure 1.2: The basic Spanish room offers only its corresponding configuration.

Demo

CRT-3 · Español · básica
CRT-3 · English · basic
CRT-7 · Español · extendida
CRT-7 · English · extended

**Reflexión
cognitiva /
Cognitive
Reflection Test
— BEER Lab**

Selecione una variante de este juego. Select a version of this game.

Figure 1.3: Demo shows four configurations isolated to this family.

and consider whether enough observations exist to discuss aggregates without identifying people. In a small class, even an aggregate can disclose individual information when combined with what participants know about one another.

Disclosure should follow the lesson plan and current progress. Avoid treating an early capture as the cohort's final result. Exporting data is distinct from projecting results: private review can be necessary before showing anything to the class. Authentication and dedicated controls separate these operations. Never distribute the report address as a participant entry link.

1.5 Lesson planning and pedagogical closing

For a 75-minute class, allow a short procedural introduction, sufficient time for the selected form and at least twenty minutes for interpretation. The basic edition leaves more room for denominator exercises. The extended edition requires local piloting and can be assigned as a supervised preparatory activity, reserving classroom time for discussion. Timing estimates are planning choices rather than normative parameters or automatic program limits.

Close by requesting one descriptive statement supported by the report and one claim the design cannot support. Describing a mean with its N differs from claiming that a language causes lower ability. Assessment of student learning can focus on defending denominators, identifying exclusions and communicating uncertainty. It need not reward test performance. Save the export, manifest and incident notes before ending the administrative session.

Checkpoint

Before starting, check session, form, language and one position per person. Before projection, check that the access-code panel is closed, each measure has its own N, disclosure is intentional and personal identifiers are absent.

1.6 Instructions, comprehension and response entry

The observational unit is a person. There are no partners, social roles, synchronization or attendance multiples. Prepare one participant position per person and distribute distinct individual links. A room is a common entry route, an individual link points to one session position, and a report is a private instructor page. These addresses have different purposes. Two students using one individual link can mix their record even when storage functions correctly. Keep a new session for the actual class and save its identifier so rehearsal responses are never included in classroom statistics.

Instructions require independent work without calculators, online searches or outside assistance. There is no automatic countdown, performance payment or academic grade based on correct answers. Assessment of subsequent discussion can be appropriate, but announce that solving the CRT itself does not determine a grade. Avoid suggesting that the obvious answer is deceptive before administration: such guidance changes the task. Clarify navigation and input formats without teaching a problem solution. If external help occurs, document it rather than assuming that a correctly recorded answer establishes unaided performance.

The comprehension question checks procedure rather than mathematical skill. An incorrect choice about external tools receives a specific explanation while retaining the selected option for correction. Attempts are not part of the CRT score. Each substantive problem then appears on its own page in source order. This differs from a paper questionnaire and must be preserved in later comparisons. Seven technical periods support both forms. The basic edition activates the first three problems,

skips inactive periods and reaches the post-task familiarity declaration without artificial waiting.

The first six problems use numeric entry with an explicit unit; basic uses only the first three. Decimal comma and decimal point are accepted, up to six decimal places, without typed units or thousands separators. The price problem requests cents: entering 5 means five cents, whereas 0.05 means five hundredths of a cent. The software does not silently convert an incorrect unit into a correct answer. Format feedback explains how to enter a number without revealing the solution. Fractions such as $1/2$ and text such as five are rejected, retaining the original input for correction.

The seventh problem keeps the published categorical response structure for the investment: break even, gain or loss. It is not converted into a numeric final-value question. Every problem includes an explicit omission checkbox. An answer and omission submitted together produce a conflict message; the software does not silently prefer one. Confirmation records the submission and advances without backwards navigation. Recorded elapsed seconds include pauses and error corrections. They are operational timing information rather than a speed score or evidence that a particular cognitive process occurred.



Individual activity

The Behavioral and Experimental Economics Research Lab in the Department of Economics at Tecnológico de Monterrey is a place to learn about economic decision making through controlled activities.

Participation rules

- Work individually and privately.
- Remain silent and do not discuss your responses with others.
- Follow the instructions on each screen.
- Do not refresh or close the page while responding.
- If you need help, raise your hand or contact staff privately.

Responses are private. The instructor may show aggregate results after choosing to reveal them. Individual profiles will not be projected.

Access code provided by the instructor

☐ I have read the rules and wish to start.

Start

Figure 1.4: Neutral institutional welcome and private entry.

BEER Lab

Instructions

Solve individually, without a calculator, online searches or outside help. Enter only a number in the stated unit, using either decimal comma or point and no thousands separators. The last problem in the longer form uses choices. You may explicitly omit a problem. Submission confirms the answer and advances without going back. There is no automatic time limit, payment or academic grade.

This is an individual activity. There are no partners, endowments or payments linked to responses. You may indicate that you prefer not to answer an item.

Responses are private. The instructor may show aggregate results after choosing to reveal them. Individual profiles will not be projected.

7 problems.

Continue

Figure 1.5: Instructions explain resources and response format.

BEER Lab

Before you start

How should you solve the problems?

- ☐ Individually, without external tools or assistance
- ☒ By searching online or discussing answers with other people

Your answer: By searching online or discussing answers with other people. The correct answer is Individually, without external tools or assistance. Solve the problems individually, without a calculator, online searches or external assistance. You may use explicit omission if you prefer not to answer.

Continue

Figure 1.6: Procedural feedback retains the selected choice.

BEER Lab

Problem 7 of 7

Solve the problem

Simon invests 8000 dollars in stocks. The investment first falls in value by 50%. Its remaining value then rises by 75%. Compared with his original 8000 dollars, what is his final position?

choose one option

Response

☐ Neither gain nor loss

☐ Has gained money

☒ Has lost money

☐ I prefer to omit this problem

Continue

Figure 1.7: The seventh problem retains a categorical response.

1.7 Familiarity and the extended activity

After solving the problems, participants indicate whether they had previously seen each problem or something very similar. Options distinguish not seen, seen, uncertain and prefer not to answer. The last option is a submitted declaration rather than pending data. Asking afterwards avoids announcing known problems beforehand, but cannot verify whether someone remembered a solution. Exposure is self-reported and does not subtract correct answers. An analysis restricted to unfamiliar problems requires its own prespecified estimand and denominator; it must not silently replace the defined CRT count.

The extended edition then presents three hypothetical decisions. The first compares a sure gain of 100 ECU with a fifty-fifty lottery paying 300 or zero. The second compares a sure loss of 100 with a fifty-fifty lottery losing 300 or zero. The third compares winning with one red ball in ten against eight in a hundred. These are independently authored teaching examples; none changes the seven-problem correct count. Risk choices reflect preferences as well as possible understanding differences and are not automatically marked rational or irrational. ECU is an accounting unit, and no payment is promised in this activity.

The final screen confirms that responses were recorded. Basic editions do not provide an individual profile. Extended editions provide four private counts only when the whole activity is submitted and the instructor currently reveals results. Hiding results withdraws that feedback at the next check. Completing all seven problems permits CRT scoring even if familiarity or additional decisions remain unfinished. The report therefore distinguishes scoreable batteries and finished activities. A difference between these progress counts is expected and should be explained before interpreting attendance or exclusions.

1.8 Reading the report and checking downloads

Before disclosure, the report shows attendance and progress while keeping outcome aggregates hidden. After disclosure it presents four counts with maxima, valid N, exclusions, mean, median, sample standard deviation and integer histograms: correct, prespecified intuitive errors, other errors and omissions. In a complete battery these counts sum to three or seven. An intuitive error is an observable response category, not proof of unreflective thinking. A correct answer likewise does not demonstrate that someone suppressed an initial incorrect response. The report does not project answer keys or personal identifiers.

The item table retains its own submission denominator and pending count. Familiarity rows separate seen, unseen, uncertain, not disclosed and pending. In the extended report, additional decisions are grouped by CRT correct count, with a separate valid-choice denominator for each field. Omitted decisions are excluded from that field's denominator. A cell showing zero over zero means there were no valid choices, not that nobody prefers a lottery. Close the access-code panel before projection and consider whether very small cells could reveal individual responses when combined with what classmates already know.

The custom CSV contains 29 columns. It includes one row per active problem and participant position, including pending records; four score rows; and three additional decision rows in the extended form. This gives seven data rows per basic position and fourteen per extended position, excluding the header. Three positions therefore produce 21 or 42 rows. These are coverage checks rather than the sample size for a mean. Score rows are not additional people. The native wide export instead stores one row per person and periods in columns, so the two structures should not be joined by row position.

BEER Lab

Previous experience

Before this activity, had you seen each problem or a very similar version? Select one response per problem. This does not change your correct count.

Problem 1

☐ I prefer not to answer

☒ I had not seen it

☐ I had seen it

☐ I am not sure

Problem 2

☐ I prefer not to answer

☐ I had not seen it

☒ I had seen it

☐ I am not sure

Figure 1.8: Excerpt showing the first two familiarity declarations; the full form asks about all seven problems.

BEER Lab

Additional decisions

These hypothetical decisions are recorded separately and do not change earlier correct counts. ECU means Experimental Currency Units; here it is only an accounting unit and no payment will be made.

Choose between receiving 100 ECU for sure and a lottery with a 50% chance of receiving 300 ECU and a 50% chance of receiving 0 ECU.

☐ Sure amount

☒ Lottery

☐ I prefer not to answer

Choose between losing 100 ECU for sure and a lottery with a 50% chance of losing 300 ECU and a 50% chance of losing 0 ECU.

☐ Sure amount

☐ Lottery

☒ I prefer not to answer

One ball is drawn at random from your chosen urn; a red ball wins. A has 1 red and 9 white balls; B has 8 red and 92 white balls. Which urn offers a higher chance of winning?

☒ Urn A

☐ Urn B

☐ I prefer not to answer

Continue

Figure 1.9: Additional hypothetical decisions remain separate from CRT counts.

BEER Lab

Activity complete

Your responses have been recorded. Keep this screen open and wait for the instructor's directions.

If private feedback is enabled, it will appear here after the instructor reveals the results.

Figure 1.10: Closing confirms recording while feedback remains hidden.

BEER Lab

Activity complete

Your responses have been recorded. Keep this screen open and wait for the instructor's directions.

Your descriptive result in this activity

CRT correct answers: 5 / 7

Incorrect intuitive responses: 1 / 7

Other incorrect responses: 0 / 7

Explicit omissions: 1 / 7

Counts describe your activity. An intuitive response is a prespecified incorrect alternative, not an observation of how you thought. Counts are not diagnoses or a complete measure of intelligence.

Figure 1.11: Private extended feedback requires completion and current disclosure.

BEER Lab

Cognitive reflection instructor report

Form: CRT-7 · Extended · en

Language adaptation: BEER wording adaptation; Spanish teaching translation, with no claimed psychometric validation

[▶ Class access code](#)

Participants

3

Activities finished

1

Batteries finished

1

Scoreable batteries

1[Reveal aggregates and enable configured feedback](#)[Download this session's responses and scores](#)

Aggregates remain hidden. You can review progress without revealing profiles.

Instrument and provenance

Numerical designs and keys: Frederick (2005), DOI 10.1257/089533005775196732; Toplak, West and Stanovich (2014), DOI 10.1080/13546783.2013.844729. BEER teaching wording and Spanish translation; no validated psychometric translation is claimed.

No norms, percentiles or classifications of people. Recorded time includes pauses and corrections; it does not establish a mental process.

Figure 1.12: The report initially hides outcome aggregates.

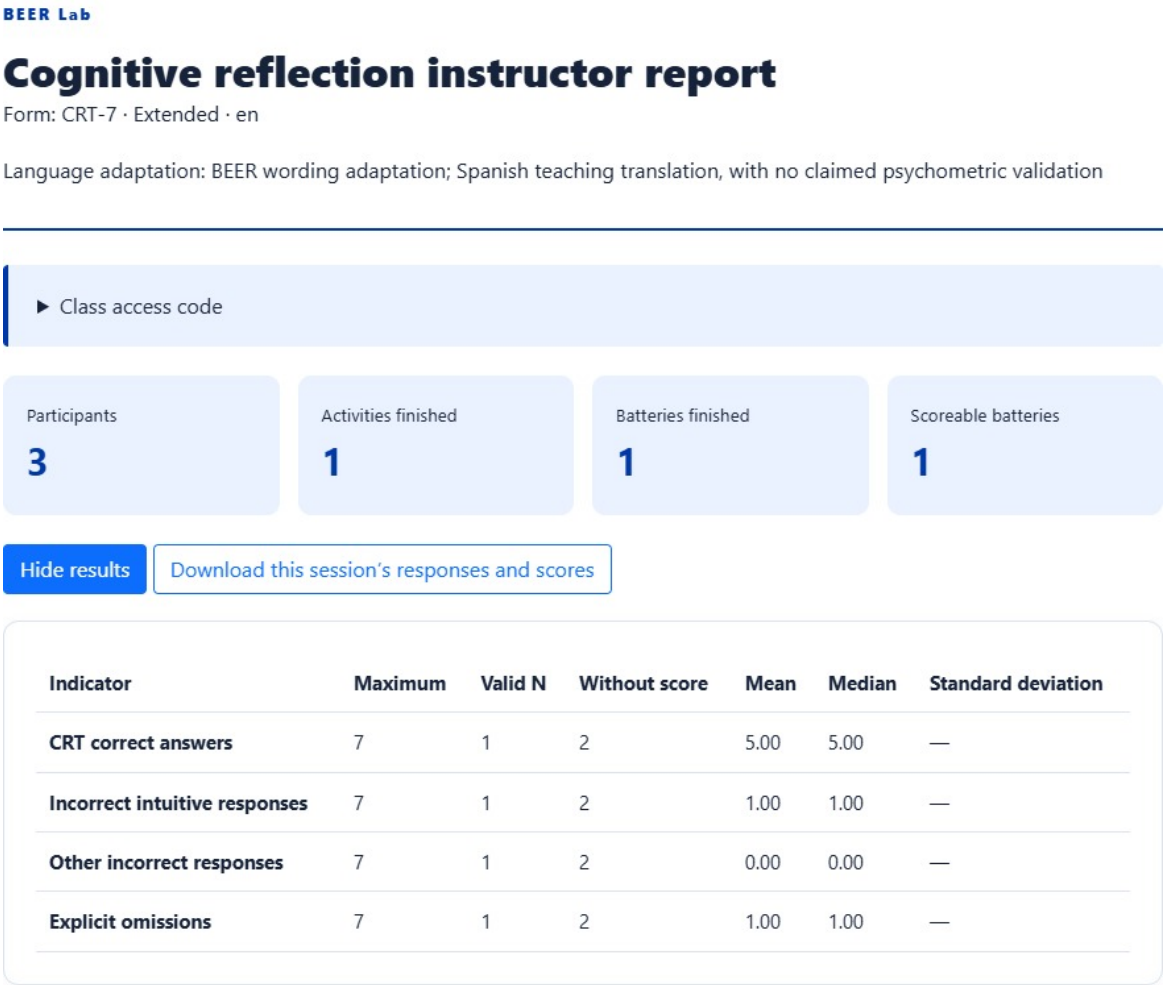


Figure 1.13: Counts retain explicit maxima and valid-observation denominators.

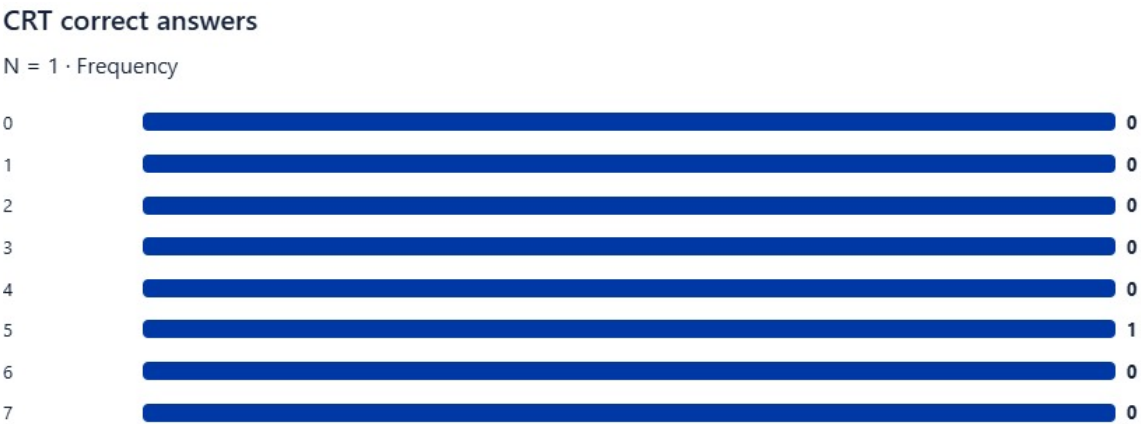


Figure 1.14: Histogram frequencies sum to the complete-battery N.

What cognitive reflection scores can tell us

2.1 A response, a process and a construct

Cognitive reflection concerns reconsidering an initially attractive answer when the problem calls for further checking. The implementation observes a final submitted answer, not the sequence of thoughts that produced it. This distinction matters immediately. A participant can answer correctly through deliberate calculation, recognition of a familiar solution or fluent mathematical intuition. An incorrect answer can follow sustained effort. The three labels used for substantive answers describe agreement with a key and a prespecified distractor, not measured levels of effort. Start discussion by asking students to separate what the screen recorded from the psychological explanation they are tempted to attach to it.

A construct is an interpretation that connects behavior across tasks and conditions. A count is a reproducible summary of observations from specified tasks. Reproducibility of the count is necessary for analysis but does not make every construct interpretation valid. To move from one classroom score to a statement about reasoning generally, one needs evidence about reliability, validity, language and administration. This guide uses the CRT to teach that inferential chain. It does not supply norms or a threshold separating reflective and unreflective people. Even a perfectly implemented key cannot establish an interpretation that the design did not measure.

Theoretical foundation

The observed variable is a submitted response in a stated unit. Its classification is correct, prespecified intuitive error, other error or omission. A cognitive process is an explanation of that observation. Keep classification and explanation separate when presenting evidence.

2.2 Why the tempting answer is useful

Some numerical problems invite an answer that uses visible numbers while overlooking a relationship. In the price problem, subtracting the stated difference from the total leaves an amount that must be divided between two objects. In the growth problem, halving the day number confuses a linear representation with doubling. A specified intuitive alternative therefore helps organize error analysis more precisely than a single correct-versus-wrong indicator. It identifies a particular result of a possible shortcut. Other incorrect answers remain informative, especially when unit mistakes or alternative calculations are plausible.

The presence of an intuitive answer does not imply that intuition is generally harmful. Many

familiar decisions are made accurately without explicit calculation, and deliberation can also fail when a person lacks relevant knowledge. Classroom discussion should focus on whether an answer satisfies the stated conditions. Ask what calculation would falsify it. This makes checking a transferable practice rather than a story about two kinds of people. It also avoids treating a wrong answer as evidence of laziness, a claim the final-response data cannot establish.

2.3 Three problems and seven problems

The basic form preserves the three numerical structures introduced by Frederick: prices, production rates and doubling. The extended form adds joint consumption rates, class ranks, trading profit and successive percentage changes, following Toplak, West and Stanovich. Its seventh response is categorical. Each form is administered once. The longer form contains the short form's content, so administering both to the same person does not produce two independent observations. Familiarity and feedback can change the later response even when underlying mathematical knowledge remains stable.

Raw correct counts have ranges zero to three and zero to seven. Dividing by the maximum gives a proportion correct on the administered problems, which may be useful descriptively. It does not equate the forms. Different content, difficulty, opportunities for error and familiarity affect that proportion. If one cohort completes the short form and another completes the long form, report them separately. Within the long form, the first three and added four can be examined as explicitly defined subsets in an external analysis, but the delivered report does not pretend that those subsets are independent measurements of different mental faculties.

2.4 Units are part of the answer

A number only becomes an answer when it is associated with a unit and question. Five cents and five dollars are different quantities. The interface displays the requested unit next to the response control and preserves it in the custom export. Exact decimal normalization makes 5,00 and 5.00 equal when either represents five in that unit. It does not transform 0.05 into five cents merely because a respondent may have thought in dollars. Such a transformation would require an unobserved assumption about intent and would change the prespecified scoring rule.

For a teaching exercise, discuss plausible unit mistakes after scoring. If a student identifies one, retain the recorded response and document the incident rather than retroactively guessing intent. A sensitivity analysis can transparently explore another coding rule, but the resulting score must have a separate name and version. Never overwrite the original answer or present a post hoc correction as the original keyed result. Reproducibility requires that another analyst can derive the same classifications from the stored answer, unit, form and scoring version.

2.5 Missing data and complete-battery counts

A submitted omission is a deliberate use of an available response control. It contributes zero correct answers and increments the omission count. A pending problem has not been submitted, so its eventual answer is unknown. Pending records do not become wrong answers. The four complete-battery counts are withheld whenever any active problem remains pending. Once all problems are submitted, the four counts sum to the form length. The observed number of correct answers in a partial record can support progress checks but is not silently promoted to a final score.

An omission and an error can have the same contribution to the correct count while representing different recorded behavior. Similarly, a person may have a scoreable battery but an unfinished activity because post-task declarations remain pending. Distinguish the denominator for a CRT mean, the denominator for an item percentage and the denominator for an additional choice. These sets can differ without a calculation error. Write the inclusion rule before computing statistics. Otherwise the apparent convenience of one common N can conceal selective completion and create misleading comparisons.

2.6 Familiarity, numeracy and language

Repeated public exposure is a practical issue for these well-known problems. The post-task declaration records whether each problem seemed familiar, including uncertainty and non-disclosure. Seen before does not mean the solution was remembered; not seen does not prove novelty. Treat these categories as self-reports and avoid using them to accuse a participant of cheating. Their value is to make one important administration condition visible. They can motivate a prespecified subgroup analysis, but excluding familiar responses after looking at results can change the research question and invite selective reporting.

Numeracy refers to skills for working with quantities, rates and probabilities. The CRT requires such knowledge alongside any tendency to check an answer. Language comprehension and familiarity with task conventions also matter. A translation that reads fluently is not thereby psychometrically equivalent. This implementation declares its Spanish wording as a teaching translation and documents its numerical fidelity. Comparing Spanish and English means without balancing populations, exposure and conditions cannot isolate a causal language effect. Encourage students to name at least two alternative explanations before interpreting such a difference.

2.7 What the original studies motivate

Frederick introduced the short task in work connecting cognitive performance to economic choices involving time and risk. Toplak and colleagues examined a longer set in relation to rational-thinking tasks and other individual differences. These articles motivate classroom questions about associations between performance and decisions. They do not make the hypothetical examples added here a validated battery, nor do they turn a class correlation into a causal effect. Read the original papers for their samples, measures and analysis rather than borrowing their conclusions as predictions that every classroom must reproduce.

The implementation preserves numerical content and keys while independently phrasing the situations. It presents one problem per page, explicit omissions, fixed order and no automatic time limit. These choices support an auditable teaching workflow but are administration differences. Record them when comparing results or planning research. The source articles and their publisher files are not bundled as reusable assets. Attribution of a mathematical design is distinct from a claim that the entire publisher document carries an open license. The guide's independently written explanations and demonstration screens document this specific product.

2.8 Descriptive uncertainty and small classes

A sample mean summarizes the observed complete batteries. It does not reveal the population mean exactly. In a small class, moving one person into or out of the complete set can noticeably change the average. The sample standard deviation describes variation among included scores;

it is unavailable with fewer than two values. A missing deviation in that situation is not zero dispersion. A histogram can reveal a concentration at endpoints or several separate clusters that a mean alone hides. Its categories are integer counts, not psychological types.

Do not read a percentage from a small cell without its denominator. One lottery choice among one valid respondent and ten choices among ten both yield one hundred percent but provide different information. A zero denominator means no proportion can be estimated from that cell. In discussion, request the numerator, denominator, inclusion rule and context before an interpretation. This practice transfers directly to applied economics, where changing sample composition can matter as much as changing the formula. Classroom convenience is not a reason to conceal the distinction.

Attention

Avoid ranking classmates or interpreting a low count as low intelligence. A task response is conditioned by content, exposure, language and circumstances. Explain mistakes through the problem's constraints and discuss the limits of inference before showing subgroup patterns.

Seven fully worked problem structures

3.1 Prices and a difference

Let x be the ball's price in dollars. The bat costs x plus 1.00. Adding the two prices gives $2x$ plus 1.00 equal to 1.10, so $2x$ equals 0.10 and x equals 0.05 dollars. The requested unit is cents, making the keyed entry 5. Check by reconstruction: a five-cent ball and a dollar-five bat total a dollar-ten and differ by one dollar. The attractive entry 10 would imply a dollar-ten bat and a total of a dollar-twenty if the stated difference were maintained.

The teaching point is representation of simultaneous constraints. Simply subtracting one dollar from the total does not yet isolate the ball because the ball's price appears in both the total and the definition of the bat. Ask students to explain the extra division by two in words before using algebra. An answer of 0.05 in the cent field remains another incorrect response under the declared unit rule. Discussing that case can teach measurement without asserting what the respondent intended.

3.2 Parallel production

Five machines producing five widgets in five minutes imply one widget per machine in five minutes, provided machines work independently at the same constant rate. A machine's rate is one fifth of a widget per minute. A hundred machines therefore produce twenty widgets per minute, so a hundred widgets require five minutes. The key is 5 minutes and the prespecified intuitive answer is 100. Both the number of machines and desired output rise by the same factor, leaving production time unchanged under these assumptions.

Use a table with columns machines, output and minutes if students conflate rate and total quantity. Five machines for five minutes supply twenty-five machine-minutes; one hundred machines for five minutes supply five hundred machine-minutes. At five machine-minutes per widget, the second total yields one hundred widgets. This does not describe every actual factory: congestion, setup costs and shared bottlenecks would change the production technology. The problem states a simple model. Identify its assumptions before using a real-world counterexample as if it refuted the calculation.

3.3 Doubling and the previous day

The patch covers the whole lake on day 48 and doubles each day. Its area one day earlier must therefore be half the lake, giving day 47. Dividing 48 by two gives 24, the prespecified intuitive answer, but halves elapsed time rather than area. Under constant doubling, moving backwards by

one day divides area by two. Moving backwards twenty-four days would divide it by two raised to the twenty-fourth power. Those are very different operations.

A useful explanation begins at the endpoint instead of attempting to reconstruct the first day's patch. Let the final area be one lake. The sequence moving backwards is one, one half, one quarter and one eighth. Label these day 48, day 47, day 46 and day 45. Students can then see that the answer follows from a local ratio, without needing the lake's physical size. This provides a bridge to compound growth in economics while keeping a deterministic doubling exercise distinct from an empirical model of disease, markets or population.

3.4 Combining rates

John drinks one barrel in six days, giving a daily rate of one sixth of a barrel. Mary's rate is one twelfth. Together their daily rates sum to one quarter, so one barrel requires four days. The prespecified intuitive answer is nine days, the average of six and twelve. Averaging completion times does not combine simultaneous contributions to the same stock. The combined rate is faster than either individual rate, making a time longer than six days a useful warning that the proposed answer needs checking.

The assumptions matter: their rates remain constant and they can contribute simultaneously without affecting each other. The exercise is about adding rates with a common unit, not recommending a consumption behavior. Ask students to verify the answer by computing quantities after four days: John consumes two thirds of a barrel and Mary one third. The sum is one barrel. This reconstruction is often clearer than memorizing a reciprocal formula and transfers to productivity problems where parallel workers contribute to one total.

3.5 Ranks and double counting

Jerry is fifteenth from the top and fifteenth from the bottom. With no ties, fourteen students stand above him and fourteen below. Adding those twenty-eight others and Jerry gives twenty-nine students. Adding fifteen plus fifteen gives thirty, the prespecified intuitive answer, because Jerry appears in both rank counts. The general expression for a person with upper rank a and lower rank b is $a + b - 1$ when the ordering has no ties and includes everyone.

Draw one central position with fourteen marks on each side to make the overlap visible. This is a counting problem rather than a claim about the distribution of grades. If two students share a rank, the simple formula may no longer identify class size under common ranking conventions; that is why the teaching wording states no ties explicitly. The answer demonstrates removal of a duplicate observation, an issue that later appears when joining data tables and counting unique participants across multiple exported rows.

3.6 Trading profit across transactions

The cash flows are a payment of 60, receipt of 70, payment of 80 and receipt of 90 dollars. Adding signed flows gives minus 60 plus 70 minus 80 plus 90 equal to 20. Equivalently, the first completed trade earns 10 and the second earns 10. Total profit is 20 dollars. The prespecified intuitive answer is 10. It can arise from an incomplete accounting representation, but the response alone cannot establish which intermediate calculation the person used.

Write a running cash ledger starting at zero for an accounting illustration: minus 60, plus 10, minus 70, then plus 20. Negative intermediate balances represent financing needs in this stylized

calculation; they do not change net profit without an additional borrowing cost. A real trading project might include transport, interest or taxes, but those are absent from the stated problem. Introducing them after the fact creates a different problem. The methodological lesson is to preserve the specified outcome and assumptions while recognizing which economic complications would be necessary in an applied model.

3.7 Successive percentages and the categorical answer

Starting with 8000 dollars, a fifty-percent fall leaves 4000. A seventy-five-percent rise then applies to 4000, adding 3000 and producing 7000. The final amount is 1000 below the starting amount, so the correct category is lost money. The prespecified intuitive category is ahead. Adding minus 50 and plus 75 as if both percentages shared the original base gives an apparent positive 25 percent, but the base changed after the first movement. Keeping that base explicit resolves the problem.

The interface preserves categories rather than requesting 7000. A person can correctly identify a loss without calculating the exact final value, and changing the question to a numeric balance would change the response demand. After scoring, the instructor can use the exact calculation for explanation. A fifty-percent loss requires a hundred-percent rise from the reduced base to break even; seventy-five percent is insufficient. This is a classroom illustration of compounding, not financial advice or a forecast about stocks. No real investment decision is elicited.

Feedback script

Ask which constraint an answer fails, then reconstruct an answer that satisfies all constraints. Invite an explanation of units and denominators. Describe the recorded error without assigning a motive or labeling the person who made it.

Worked cases for data and economic interpretation

All records and numbers in this chapter are constructed teaching examples. They are not observations from a research sample or expected classroom benchmarks. Work through each case by naming the unit of observation, active form, response state, calculation and permissible interpretation. The objective is to make the inclusion rule inspectable before discussing a behavioral story.

What participants will see

Work in small discussion groups after administration. For each case, write a calculation, one supported descriptive statement and one statement the example cannot support. Do not exchange problem solutions while other participants are still responding.

Scoring and missing data

Case 1. Three distinct substantive outcomes

A complete basic record contains 5 cents, 100 minutes and 25 days. The first answer matches its key; the second matches the prespecified intuitive error; the third matches neither. The resulting counts are one correct, one intuitive, one other and zero omitted. They sum to three. The correct proportion on this form is one third. It is not one correct among two errors after excluding the error that happens to have a familiar label. Both errors remain part of the administered content.

The report can describe this response profile without deciding which problem received the most effort. A plausible verbal account of the second response is linear scaling of time, but the final number alone cannot verify that account. Ask what additional evidence, such as a suitably designed process protocol, would be required. Such a protocol would itself change administration and should not be retroactively claimed to have been present. Preserve the response classification and distinguish it from the hypothesis about reasoning.

Case 2. Omission contributes zero, pending withholds a score

Participant A submits correct responses on the first two basic problems and explicitly omits the third. A has two correct, zero intuitive, zero other and one omitted. Participant B submits the same two correct responses but never sends the third screen. B has two observed correct answers and one pending problem, with no complete CRT count. A and B therefore do not both receive a final score of two. Their observed progress overlaps while their completion states differ.

If these are the only two positions, the report's mean correct count is two with valid N equal to one and one position excluded. It is not a mean of two over two completed people. A possible later response from B could change the average. Report the snapshot time and preserve the pending record. Completing A's omission was an observed action; completing B's unvisited problem would require information that does not yet exist.

Case 3. Reading an item denominator

Ten positions exist. On problem 2, four submitted answers are correct, two are intuitive errors, one is another error, one is an omission and two remain pending. There are eight submissions. The displayed item percentage is four divided by eight, or fifty percent. Dividing by ten would estimate correct submissions among all assigned positions, a different quantity. Excluding the omission would give four divided by seven, which is not the declared report rule.

Neither the item percentage nor the number of submissions establishes the complete-battery N. Some of these eight people may leave another problem pending. Before interpreting the item as easy or difficult, check its exposure pattern, unit and administration conditions. The table supports a description of this session's submissions. It does not identify a stable population item parameter or a causal reason for the two pending records.

Case 4. A cohort mean with one absent score

Four complete basic batteries have correct counts 0, 1, 2 and 3. A fifth position is pending. The mean among complete batteries is 1.5 and the median is also 1.5. The sample standard deviation is the square root of five thirds, approximately 1.291. There are four valid observations, not five. Treating the pending record as zero would lower the mean to 1.2 and change dispersion, silently replacing an unknown outcome with a particular response profile.

The histogram has one person at every possible basic score. That pattern communicates information the mean alone hides. Do not call the four bars personality categories. If the fifth person later finishes with three correct, the mean becomes 1.8. This change can occur without any completed participant revising a response. The denominator changed because the snapshot became more complete. A teacher can use the example to connect data cleaning decisions with interpretation of evolving administrative dashboards.

Case 5. No completed batteries

Three positions have each submitted the first basic problem correctly and stopped. The item table can show three correct submissions for that problem, but all complete-battery means remain unavailable with valid N zero. A histogram containing zero counts in every score bin reflects an empty set of complete scores. It does not mean that three people scored zero. Likewise, an unavailable sample standard deviation is not evidence that everybody performed identically on the entire battery.

Ask students to propose a progress statement that remains accurate: three first-problem submissions were correct and no battery is complete. Contrast that with a claim that the class achieved a perfect CRT score. The latter incorrectly extrapolates one item to the whole form. This example makes a general lesson visible: data availability has a structure, and a software interface should preserve that structure instead of filling empty measures with reassuring numbers.

Case 6. Unit-sensitive decimal normalization

Two participants type 5,00 and 5.000 for the price problem. Both normalize to 5 in cents and are correct. Another types 0.05 in the same field. That number normalizes to 0.05 cents and is classified as another incorrect response. The stored raw value helps identify what was entered; the normalized value supports exact comparison. A malformed entry such as 1,2,3 is rejected before a substantive response is recorded, because it has no unambiguous representation under the input rule.

A later explanation that the third person intended dollars may be useful in teaching, but does not change the original keyed count. For an explicit sensitivity exercise, create a second coding scheme and report how many records it changes. Keep its result separate and explain the assumption. Automatically converting all small decimals would also recode plausible genuine errors and therefore cannot be justified merely by the desire to increase apparent accuracy.

Familiarity and economic decisions

Case 7. Familiarity is not remembered correctness

In a complete seven-problem record, a participant marks two problems seen, three unseen, one uncertain and one not disclosed. Those declarations sum to seven submitted exposure records. None is pending. Suppose both seen problems are correct, two unseen problems are correct and the remaining three are wrong. The original CRT count remains four. Subtracting the two familiar successes would create a new outcome, not correct the original one.

A novel-content analysis could report two successes among three declared-unseen problems, but its denominator and interpretation differ from the seven-item form. The uncertain and undisclosed responses cannot simply be treated as unseen. Furthermore, a seen problem may have been encountered without its solution. Ask why self-report is useful yet incomplete. The answer should mention the available information, potential recall error and the absence of direct observation of what solution, if any, the person remembered.

Case 8. Expected value in the gain decision

The sure alternative gives 100 ECU. The lottery gives 300 with probability one half and zero otherwise, for an expected value of 150 ECU. Choosing the sure amount therefore forgoes 50 ECU in expected value under these stated payoffs. That does not make it an incorrect preference: expected utility can favor certainty for a person whose utility is sufficiently concave over the relevant outcomes. No utility function or real monetary stakes are measured by this single hypothetical question.

To illustrate, compare linear utility with a simple concave square-root representation over nonnegative gains. Linear utility favors the lottery's expected 150 over 100. Square-root utility gives the sure amount utility 10 and the lottery approximately 8.66, favoring certainty under that illustrative specification. These calculations demonstrate dependence on assumptions, not an estimated utility model for a classmate. The report records the choice separately from CRT correctness and never awards an additional CRT point for taking the higher expected-value option.

Case 9. Losses do not create a mirrored correct answer

The sure loss is 100 ECU. The loss lottery has outcomes minus 300 and zero with equal probability, for expected value minus 150 ECU. A person minimizing expected loss would prefer the sure alternative. Someone selecting the lottery might respond to the possibility of avoiding any loss, to preferences or to misunderstanding. The recorded choice does not distinguish those explanations. The report therefore counts lottery selections rather than marking them as errors.

Suppose six valid participants choose the loss lottery four times. The observed lottery share is four sixths, about 66.7 percent. It is a descriptive statistic for six hypothetical choices. Comparing it with the gain share may motivate discussion of framing and reference points, but the two offered payoff sets differ in signs and expected values. They are not a validated causal framing experiment with randomized equivalent descriptions. Be precise about what changed before assigning a familiar behavioral label.

Case 10. Comparing urn probabilities

Urn A contains one red ball and nine white balls, making ten in total and a winning probability of one tenth. Urn B contains eight red and ninety-two white balls, making one hundred and probability eight hundredths. Ten percent exceeds eight percent, so A offers the higher stated chance of winning. Counting only red balls would select B because eight exceeds one, overlooking the denominator. This example has an objectively better probability under the specified objective.

The additional urn response is not part of CRT-7. A single selection of B may be compatible with denominator neglect, but cannot establish a stable bias or reveal the exact calculation used. A clear explanation compares proportions and then checks that sampling is uniform within each urn. If the sampling mechanism changed, the counts alone might not determine winning probabilities. The model assumption is therefore part of the answer, just as constant rates are part of the production problem.

Case 11. Each transfer field has its own N

Four participants with five CRT successes complete the additional activity. Gain choices are lottery, sure, omitted and lottery. Loss choices are omitted, lottery, sure and omitted. Urn choices are A,A,B and omitted. The gain lottery cell is two over three; the loss lottery cell is one over two; the A cell is two over three. Using four as the denominator for every cell would include explicit omissions as substantive preferences and conflict with the report's valid-choice rule.

Zero is a meaningful omission code in these fields, whereas an unfinished additional screen is pending. An omitted field can coexist with a finished activity because choosing not to answer is a submitted option. Analysts should retain that distinction even if both are excluded from a preference proportion. The grouping variable uses a complete CRT count, so a person with a pending CRT problem does not enter any of these score groups. The denominator has two stages: eligible complete score and valid response to the particular additional question.

Case 12. An apparent association is not an intervention effect

Imagine eight complete extended records. Among four people with at least five correct, three choose A in the urn comparison. Among four with fewer correct, one chooses A. The descriptive shares are 75 and 25 percent. The difference is 50 percentage points. The score groups were not randomly assigned, and the exercise did not manipulate reflection. Mathematical knowledge, exposure, attention or language could contribute to both outcomes.

A causal statement that increasing someone's CRT score would improve urn choice by 50 points is therefore unsupported. A defensible statement names the constructed sample, grouping rule and observed difference. If the analysis were planned for research, specify grouping before inspecting data and consider how a continuous score or an alternative threshold would affect the question. Do not search several thresholds and report only the largest contrast as though it had been predicted. The case teaches association and selection without fabricating evidence of a treatment effect.

Comparability, privacy and classroom decisions

Case 13. Equal percentages on unequal forms

One student has three correct out of three; another has seven out of seven. Both proportions are one, but their response opportunities differ. Two other students have one of three and three of seven, about 33.3 and 42.9 percent. Those percentages describe their administered forms; their difference is not a calibrated ability difference. The added items may have different difficulty and familiarity, and the short subset is embedded within the long form.

For a demonstration involving two forms, report separate distributions, maxima and counts. Explain whether assignment to forms was random, chosen by students or determined by class scheduling. Even random form assignment would identify an administration comparison for that population rather than automatically equate individual scores. A transparent percentage can be useful without pretending it resolves every comparability issue. Ask students what evidence would be needed for an equating exercise and why a shared label alone cannot supply it.

Case 14. Language and selection

Suppose an English-language group has a mean of five on CRT-7 and a Spanish-language group a mean of four. Both means have ten complete batteries. Equal N and form length remove two simple differences but do not establish comparable populations. The groups may differ in course, prior exposure, mathematical preparation, supervision or self-selection into language. The one-point difference is descriptive under those conditions.

A stronger investigation would define a target population, assignment mechanism, translation evidence and analysis plan before data collection. It would also preserve administration incidents and examine completion differences. An item can retain numbers and keys while changing difficulty through wording. This implementation openly identifies the Spanish version as a teaching translation so readers can judge that limitation. Do not erase the limitation by calling the interface bilingual. Bilingual completeness is a product property; measurement equivalence is an empirical claim requiring evidence.

Case 15. A small aggregate can identify someone

A report cell contains one participant with seven correct and one valid loss decision. Even without a name, projecting that cell may reveal the person's choice if classmates know who finished every problem. Aggregation reduces direct identification but does not guarantee anonymity in a small cohort. Review cell sizes, what participants can infer and the instructional need before revealing subgroup tables. The access-code panel should also remain closed during projection.

The software allows the instructor to hide outcomes and review progress separately. Use that control deliberately. A private export may support preparation of a less detailed teaching summary without projecting the full report. Avoid joining a participant code to a name in public materials.

For classroom assessment, grade explanation of the probability or denominator rule rather than a disclosed personal preference. These choices preserve the lesson while reducing unnecessary exposure of individual records.

Case 16. Repeating after feedback changes the question

A class first completes CRT-3, discusses all solutions and then repeats the same problems. An increase in correct answers is compatible with learning those solutions. It does not by itself establish a durable change in general reflection. The repeated measure also shares content, context and feedback with the first. Treating both attempts as independent observations would understate their dependence and change the apparent amount of information.

A learning study would need an explicit outcome definition, comparison strategy and appropriately selected transfer tasks. Retaining the same numerical structures can be useful for checking retention, but retention and general transfer are different objectives. In ordinary teaching, state the modest result: students can now explain the constraints and reconstruct the worked solutions. The delivered configurations avoid automatic repetition and do not advertise a before-after causal treatment. Save first-administration records separately if a later exercise intentionally changes the task.

Planning discussion, assessment and future use

A class sequence that protects the measurement

For a seventy-five-minute class, reserve a short procedural introduction, sufficient private response time for the chosen form and a substantial discussion period. The basic form leaves more space for data exercises; the extended form supports discussion of seven structures and separate economic decisions. Suggested response windows in the summary table are planning allowances, not normative completion times. Pilot on the actual devices and preserve an untimed administration unless the teaching design explicitly changes that condition. Do not rush slower participants merely to produce a complete dashboard.

After private responses close, begin with constraints rather than a scoreboard. Ask students to reconstruct one price or rate problem, then explain why the tempting response fails. Move to a missing-data case and request an explicit denominator. Finally compare expected value with preferences using the additional gain and loss examples. This sequence links reasoning, measurement and economics without interpreting every personal response. If students are still responding, postpone solution discussion or separate the activity physically so that feedback does not contaminate unfinished records.

Assessing learning without grading CRT performance

A useful assignment asks for a reproducible calculation, a correct inclusion rule and a bounded interpretation. Award credit for showing why a percentage has its stated denominator, explaining the difference between a count and a process, and identifying an alternative explanation for an association. A student who initially answered incorrectly can demonstrate strong learning by correcting a representation and explaining its assumptions. This is different from awarding grades according to the number of original keyed responses.

A concise rubric can give equal attention to arithmetic, measurement and inference. Arithmetic requires a checkable equation and unit. Measurement requires correct handling of omissions

and pending records. Inference requires stating what the design cannot establish. Invite written feedback in neutral terms, such as the response violates the price difference, rather than descriptions of personal ability. The objective is improved reasoning about evidence and models, not public sorting of students.

Reproducibility and changing the product

Before another administration, check the four room selectors, language, form and closing screen. A change to wording, unit, key, order or response format requires a new documented version and affected checks. Retain the earlier version with its data, because recomputing old records with a different key can silently change a published table. Store the export and manifest together and keep any identity correspondence in a separate restricted location. File hashes establish which content was used, not whether the content is psychometrically valid.

The bilingual guides have one shared figure directory in Git and standalone language copies for Dropbox. Each standalone copy contains the PDF, LaTeX source and referenced figures, with hashes checked after copying. Compilation auxiliaries and private rehearsal data are excluded. Local access is verified separately from the proposed public host. A Git commit prepares review and sharing; a push and production deployment are later actions. Keep those operational states distinct when instructing another teacher how to find the activity.

For the instructor

Close the lesson with two sentences from each group: one descriptive claim supported by a numerator and denominator, and one stronger claim that would require another design or more evidence. Preserve uncertainty as part of the lesson rather than treating it as an obstacle to discussion.

Checkpoint

Check session, language, edition and access. Review N per measure, choose disclosure and download only the intended cohort. Do not infer mental processes from one answer or classify people.

References and provenance

Frederick, Shane (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25-42. DOI: <https://doi.org/10.1257/089533005775196732>

Toplak, Maggie E., Richard F. West and Keith E. Stanovich (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2), 147-168. DOI: <https://doi.org/10.1080/13546783.2013.844729>

American Economic Association. Official JEL classification consulted for A22, C91, D81 and D91: <https://www.aeaweb.org/econlit/jelCodes.php?view=econlit>

Numerical designs and keys are attributed to these studies. Situation wording, teaching translation, cases and additional decisions belong to this BEER implementation. No validated translation is claimed, and publisher PDFs are not bundled as redistributable materials. Figures are actual local screens with demonstration responses, not observations from a research sample. Pedagogical purpose is not treated as a universal permission to reproduce any assessment.

The catalog key is crt and the proposed host is <https://crt.beer-lab.org/>. Rooms are crt_es, crt_en, crt_extendido and crt_extended. Institutional authorship and contact appear at the beginning of this guide. Bilingual metadata describe basic and extended forms, decisions and outcomes so the catalog can recover a substantive task description. Public deployment is distinct from verified local operation.