



Instructor's Guide to General Self-Efficacy (GSE)

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 accompanies the ten-statement General Self-Efficacy Scale in the BEER Lab teaching platform. It is intended for economics and social science instructors teaching capability beliefs and distinguishing them from observed performance. Basic editions administer the scale; extended editions add four original problems with prior performance forecasts. There are no payments or academic grading of personal questionnaire responses.

The operational route explains access, rooms, comprehension, omission, closing, reports and downloads. The conceptual route develops scoring, differences between language forms, missing-data rules and forecast errors. Twenty-two worked cases support classroom preparation without assuming previous experience with the instrument. Figures are genuine captures of local sessions with synthetic participants.

Each language preserves the content, anchors and order of its official source. Digital administration changes and the added problems are documented separately. A complete sum, an available-response mean and a forecast error have different units and denominators. This guide reconstructs their calculations and explains the conclusions that remain unsupported by a classroom observation. Use it to organize discussion of theory, measurement and research design rather than to assign personal labels.

Author

This guide was prepared by Stanislaw 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. Keep GSE, forecasts and performance separate. 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 What is administered and what students can learn	1
1.2 Institutional navigation and room selection	2
1.3 Creating a session and managing its access	2
1.4 Welcome, instructions and comprehension	6
1.5 Completion, participation and score availability	9
2 Capability beliefs and general measurement	10
2.1 What a self-report describes	10
2.2 Capability beliefs, performance and outcomes	10
2.3 Version, translation and item equivalence	11
2.4 The complete sum and its units	12
2.5 Missing data and the seven-response available mean	12
2.6 Evidence, limitations and classroom use	13
3 Performance forecasts and original problems	14
3.1 Why add a task-specific measure	14
3.2 Recording probabilities without confusing zero and omission	14
3.3 Solving the four problems	16
3.4 Binary squared error and the signed gap	20
3.5 What good forecasting would mean	21
4 Reports, data and classroom discussion	22
4.1 Disclosure and counters with clear units	22
4.2 Reading distributions and sensitivity rules	22
4.3 Task tables and descriptive comparisons	25
4.4 Downloading and reconstructing calculations	25
4.5 Private feedback and interpretation limits	26
4.6 Inference errors to anticipate	26
4.7 Class plan and concluding discussion	28
5 Worked problems and solutions	29
5.1 Scoring and alignment between sources	29
5.2 Missing data and denominators	30
5.3 Forecasts and observed errors	31
5.4 Operation, disclosure and inference	33
6 References and provenance	34

List of Figures

1.1	Select Classroom to reach the teaching activity catalog.	2
1.2	Review the catalog; the new card appears after deployment.	3
1.3	Select language and edition among this family's four rooms.	3
1.4	Check that the basic English room selects only its configuration.	4
1.5	Choose among four configurations; extended editions add forecasts and tasks. . . .	5
1.6	Assign one link per person and keep synthetic rehearsals separate.	5
1.7	The welcome explains participation rules without suggesting a desirable profile. . .	6
1.8	Explain scale, statement count, omission and the absence of payments.	7
1.9	Comprehension feedback retains the choice and explains its correction.	7
1.10	Answer one statement per page or use explicit omission.	8
1.11	Completion confirms recording; configured feedback requires instructor disclosure. .	9
3.1	The comprehension check distinguishes prior forecasts, later answers and no payments.	15
3.2	Read the problem and store a probability before selecting an answer.	16
3.3	The budget has a uniquely determined remainder of 60 ECU.	17
3.4	The saving compares identical goods and subtracts 68 from 72 ECU.	18
3.5	Second-year growth applies to 110 rather than the initial principal.	19
3.6	Inventory adds inflows and subtracts outflows from the starting stock.	20
4.1	Progress remains visible while score and task aggregates are hidden.	23
4.2	Each measure retains its own valid N and excluded records.	24
4.3	Histogram frequencies add to the number of complete GSE sums.	24
4.4	Task tables separate forecasts, answers, valid pairs and prediction errors.	25
4.5	Authorized feedback keeps scale, forecasts and task solutions separate.	27

The activity and its administration

Purpose	Explore general capability beliefs using GSE-10 and distinguish them from success forecasts and observed task performance.
Experimental unit	Person; individual activity without matching or waiting.
Basic edition	Complete ten-statement GSE in Spanish and English.
Extended edition	Same GSE plus four original economic arithmetic problems, locked prior forecasts and controlled private feedback.
Decision	One response per statement; extended editions add a prior probability from 0 to 100 and one answer per problem; explicit omission.
Outcome	Complete sum 10 to 40, complete and available-item means with at least seven answers; accuracy and forecast errors remain separate. No payments.
Topics	general self-efficacy; self-reports; capability beliefs; performance forecasts; forecast errors
JEL	A22, C83, C91, D84
Participants	Any positive count; one person per interface.
Suggested duration	Basic: 3-5 minutes; extended: 7-12; class: 50-75. Planning estimates, no automatic time limit.

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

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

https://selfefficacy.beer-lab.org/room/selfefficacy_en

<http://localhost:9170/demo>

1.1 What is administered and what students can learn

The General Self-Efficacy Scale, or GSE, records beliefs about the ability to cope with demanding and unfamiliar situations. This family teaches how a general self-report is constructed, what summing ten responses assumes, and why a capability belief differs from a performance test. The unit is the individual, without interaction or waiting. Its outcome describes answers to an identified form; it does not certify talent, explain all behavior or classify students. The instructor connects measurement with decisions while preserving those distinctions throughout the session.

The basic edition contains the ten official statements. Extended editions retain exactly that

questionnaire and then add four original economic arithmetic problems. Each requires a forecast of answering correctly before the answer itself. Forecasts, correctness and prediction errors remain separate from GSE. There are no payments, real endowments, demographic questions or academic grading. The extension illustrates the distinction between a general belief and an expectation for a particular task. It does not validate a new intelligence scale or reproduce a published causal experiment.

1.2 Institutional navigation and room selection

Begin at the BEER Lab website and select Classroom to reach the games catalog. The general self-efficacy card identifies the family, its topics and its editions. Proposed rooms are `selfefficacy_es` for basic Spanish, `selfefficacy_en` for basic English, `selfefficacy_extendido` for extended Spanish and `selfefficacy_extended` for extended English. Within this family's instance, Demo should contain four configurations and Rooms four rooms. Each room offers one configuration in the corresponding language and edition. If another family appears or a selector is empty, stop distributing links and request an instance check before starting the cohort.

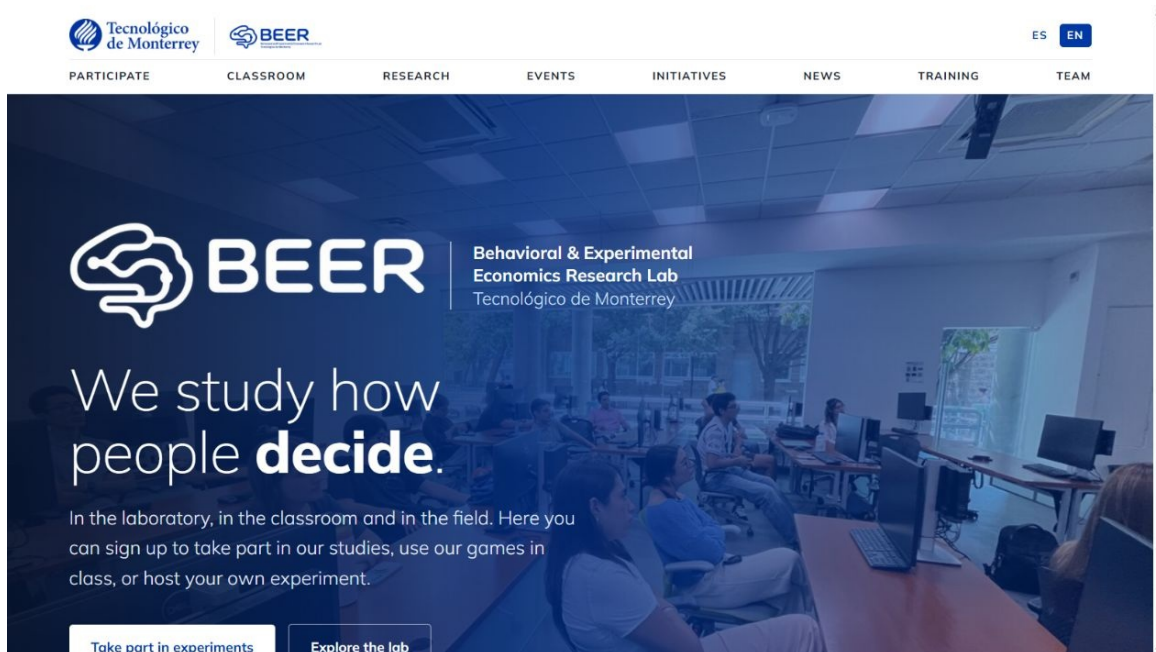


Figure 1.1: Select Classroom to reach the teaching activity catalog.

The verified delivery preview runs on localhost, port 9170, and can be accessed from the same computer. It is not an address for students using other devices. Addresses under `selfefficacy.beer-lab.org` are the intended deployment routes. A room is a shared entry point, an individual link refers to a position in a session, and a report is an authenticated instructor page. Confusing these addresses can send students to administration or cause two people to use the same position. Distribute the entry route appropriate to the teaching procedure and retain administrative links for staff.

1.3 Creating a session and managing its access

Sign into the instructor panel using the institutional credentials supplied by the laboratory. Do not share those credentials with students. Demo supports rehearsals with synthetic responses; use

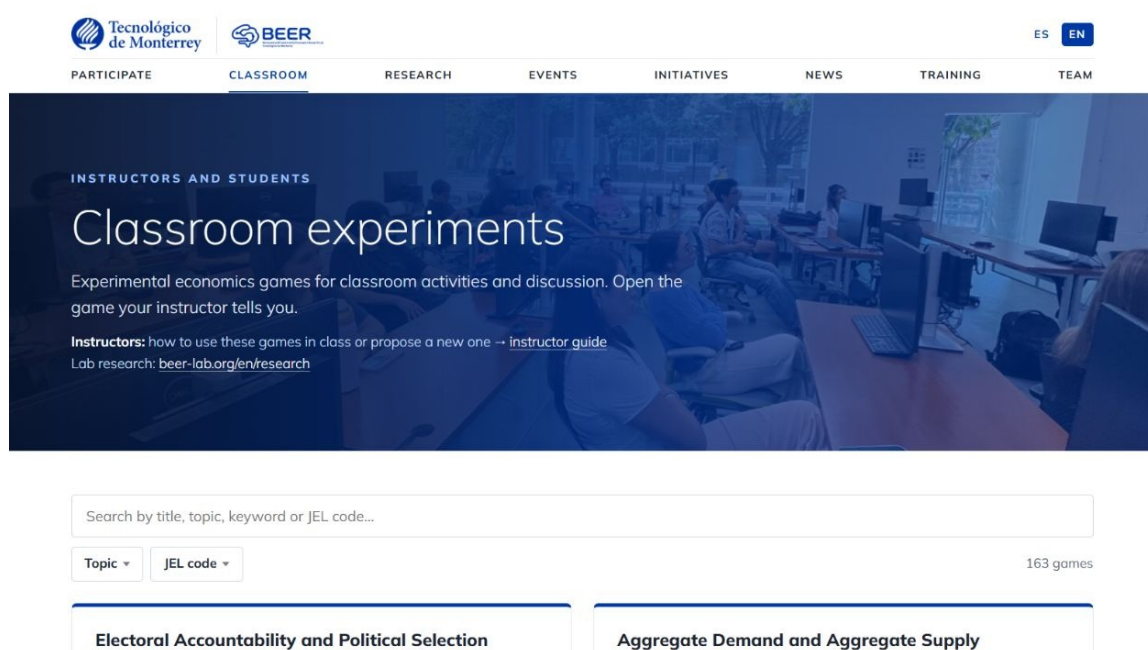


Figure 1.2: Review the catalog; the new card appears after deployment.

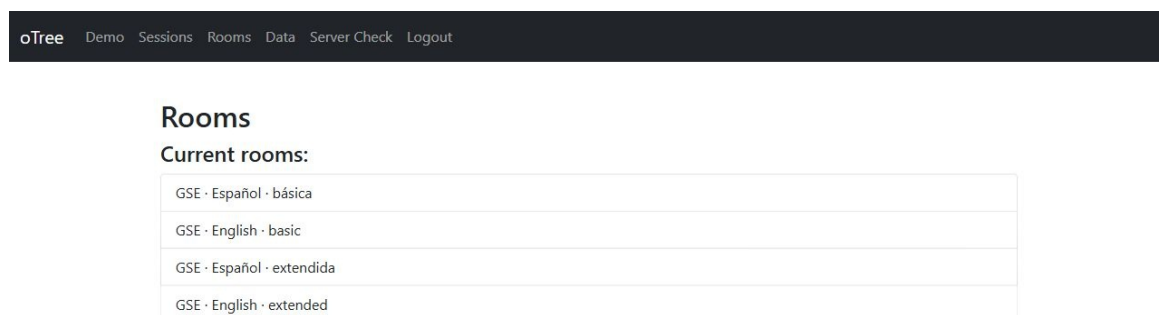


Figure 1.3: Select language and edition among this family's four rooms.

oTree Demo Sessions Rooms Data Server Check Logout

Room: GSE · English · basic

Create a new session

Session Config

GSE · 10 statements · basic ▾

Number of participants

Individual decisions: any positive participant count is allowed.

Create

[Configure session](#) ⚙

App sequence

selfefficacy_assessment (14 rounds) Official source-specific GSE-10; original arithmetic forecasts in extended editions.

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:9170/room/selfefficacy_en

Figure 1.4: Check that the basic English room selects only its configuration.

a separate session and stable rooms for the class. Any positive participant count is admissible because the task is individual. Three people can complete it without waiting for a fourth. Confirm the session size, selected configuration and cohort before distributing individual links. Keep rehearsal sessions separate from classroom data, and avoid interpreting an old demonstration as if it described the current group.

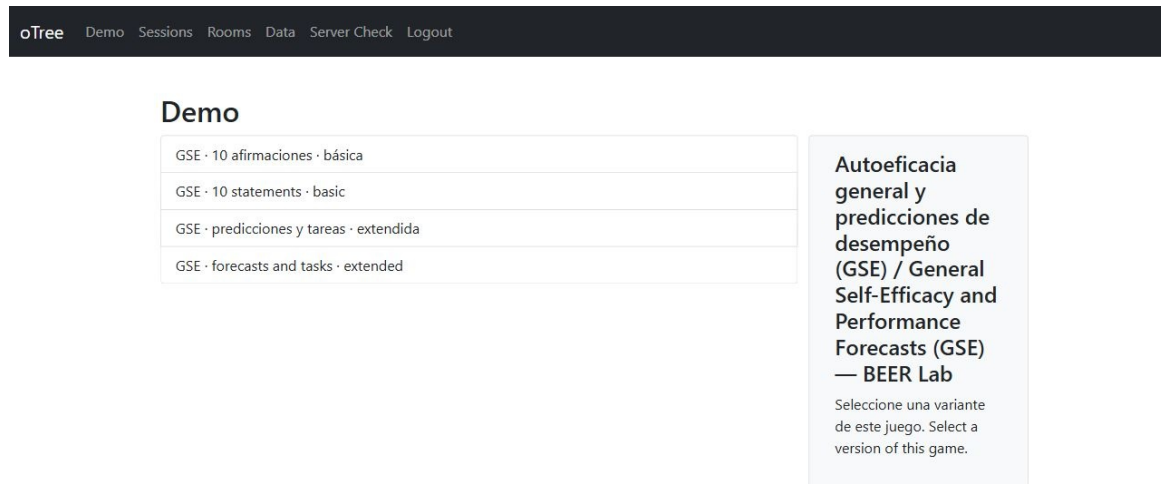


Figure 1.5: Choose among four configurations; extended editions add forecasts and tasks.

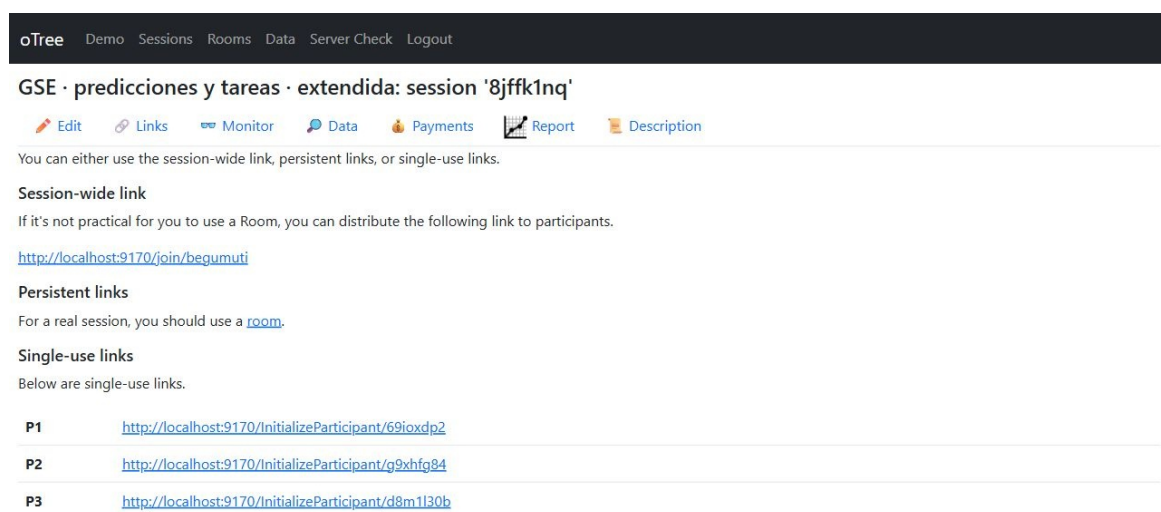


Figure 1.6: Assign one link per person and keep synthetic rehearsals separate.

Each session generates a questionnaire access code, available in a collapsed report panel. Share it with the cohort and close that panel before projection. This code differs from the administrative password and technical session code. An incorrect entry produces an explanation and allows correction; a successful entry is erased from its response field before proceeding. The gate restricts entry to the academic activity, but does not make screenshots or individual links suitable for unrestricted circulation. All demonstration positions shown in this guide are synthetic. The access procedure should remain an operational detail rather than a source of pressure to disclose personal responses.

1.4 Welcome, instructions and comprehension

The welcome introduces the laboratory, requests individual work and explains privacy. It does not announce a hypothesis about general self-efficacy or suggest a desirable profile before students answer. Instructions ask participants to describe how they typically are, explain that there are no correct questionnaire answers and present the scale from one to four. All four verbal anchors are retained from the official English form; they are not silently replaced with a different agreement scale. A separate option permits omission. Students should read privately, without being required to discuss autobiographical examples with the room or justify why they selected a particular response.



Individual questionnaire

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 1.7: The welcome explains participation rules without suggesting a desirable profile.

Before the first statement, a comprehension question asks what to do when a statement does not describe the participant well. The appropriate response is to choose the closest description or omit the item. Selecting what the instructor supposedly wants produces an explanation alongside the retained choice. This check teaches the procedure; it is neither a GSE item nor an honesty test. Allow correction without penalty. Answer questions about controls without recommending substantive responses. The purpose is a shared understanding of the task before generating the data that will later enter the discussion.

BEER Lab

Instructions

Indicate how true each statement is for you. Use the four response options. These statements describe your own beliefs; they are not knowledge questions.

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.

10 statements.

Continue

Figure 1.8: Explain scale, statement count, omission and the absence of payments.

BEER Lab

Before you start

If a statement does not describe you well, what should you do?

- ☐ Choose the option that best reflects how you are or indicate that you prefer not to answer
- ☒ Choose what you think the instructor wants to read

Your answer: Choose what you think the instructor wants to read. The correct answer is Choose the option that best reflects how you are or indicate that you prefer not to answer. Choose the first option: respond about yourself without trying to meet expectations. You may also omit an item using the corresponding option.

Continue

Figure 1.9: Comprehension feedback retains the choice and explains its correction.

BEER Lab

Statement 1 of 10

How true is this statement for you?

I can always manage to solve difficult problems if I try hard enough.

Your response

- ☐ 1 - Not at all true
- ☐ 2 - Hardly true
- ☐ 3 - Moderately true
- ☐ 4 - Exactly true
- ☐ I prefer not to answer

Confirm response

Figure 1.10: Answer one statement per page or use explicit omission.

1.5 Completion, participation and score availability

In the basic edition, the tenth statement leads to an explicit closing page. Seventeen technical periods support the shared application, but the other four do not contain basic questions or require waiting for other people. Completion confirms that responses were recorded and asks students to keep the screen open. Participants should never be left at an unexplained technical ending. Extended completion follows the four scenarios and may subsequently display private feedback when enabled by the instructor. Reaching the closing page does not guarantee that a GSE sum is available, because omission remains a permitted response.

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 1.11: Completion confirms recording; configured feedback requires instructor disclosure.

Distinguish how many people finished the procedure from how many supplied ten substantive answers. The former describes participation; the latter defines score availability. In an extended session, a participant can have a scoreable questionnaire while still answering scenarios. That GSE enters the score distribution but the activity is not counted as finished. This avoids treating a delay in a scenario as the loss of all preceding answers. It also lets an instructor discuss available data without inventing answers for participants who have not yet completed the activity or deciding that an omission represents low general self-efficacy.

Capability beliefs and general measurement

2.1 What a self-report describes

A personality measure summarizes regularities in how people describe themselves. Participants rate statements about habitual behavior, feelings or preferences, and a published rule combines those responses. The resulting score depends on those regularities, the response context, understanding of the wording and willingness to answer. An individual observation therefore does not exhaust what we know about a person. The exercise is useful for teaching the distinction between a theoretical characteristic, an observable response and a constructed measure. Those are connected concepts, but they are not interchangeable objects.

Traits can help formulate economic questions: do participants reporting greater conscientiousness maintain a plan more often, or do social dimensions relate to contributions in a game? The questionnaire does not directly measure a utility function. A classroom correlation also leaves causal mechanisms unresolved. Teaching connects concepts, measurement and evidence while examining alternative explanations.

The initial unit of analysis is the person. Items are observations contributing to that person's measurement; they are not additional participants. Ten answers from one person do not create a sample of ten independent people. This matters when reading histograms, discussing uncertainty and linking questionnaires with repeated games. A student making ten game decisions and answering one questionnaire provides one individual profile and ten related decisions. This family's report identifies valid observations separately for each dimension and retains that individual unit rather than treating each item as an independent person.

Theoretical foundation

The construct is not directly observed. We observe responses, apply a key and obtain a score. Interpretation requires a justification for each connection between those levels. The platform automates the scoring key, but it does not replace that reasoning.

2.2 Capability beliefs, performance and outcomes

Self-efficacy is a belief about being able to organize and execute the actions needed to meet a demand. Its object is one's capacity to act. An outcome expectation instead concerns what will happen if an action is performed. A student may believe that studying helps people pass while doubting their own ability to organize the study. Someone else may feel able to solve a task yet

expect an arbitrary assessment to prevent that performance from producing a good grade. Bandura develops this separation between efficacy and outcomes. Explain it before linking the questionnaire with an economic decision, because the same behavior can arise from different constraints.

GSE takes a general perspective, asking about coping with obstacles and difficulties across life situations. This scope allows research on a broad perceived resource but sacrifices specificity for an identified task. Feeling able to handle unexpected situations does not say how many compound-interest exercises someone will solve today. An expectation about one problem also depends on knowledge, information, familiarity with its format and conditions of completion. A positive association between self-report and performance would fit several explanations. A weak association would not by itself establish that the answers lack meaning. These measures have different objects and different scales.

Theoretical foundation

Distinguish three objects: a general capability belief, a forecast of answering an identified task correctly, and the observed answer. Extended editions record all three but never add them into one index. Theory proposes relationships among them; it does not require numerical equality.

Do not equate self-efficacy with self-esteem, optimism about every outcome or locus of control. Self-esteem evaluates personal worth; self-efficacy concerns perceived ability to act. Optimism can describe favorable expectations without specifying who produces the outcome. Locus of control concerns attribution of control over events. These distinctions remain useful even when observational measures correlate. Comparisons with other questionnaires should respect their definitions and sources. A verbal distinction alone does not establish which instrument is better, and a correlation does not collapse two concepts into the same construct. Ask students to identify what each question invites a respondent to report before comparing numbers.

2.3 Version, translation and item equivalence

The implementation reproduces the ten statements on the official pages hosted by Freie Universität Berlin. The English form is attributed to Schwarzer and Jerusalem, 1995. The Spanish page credits the adaptation to Bäßler, Schwarzer and Jerusalem, 1993, and cites Bäßler and Schwarzer, 1996. These are not new laboratory translations. Response anchors and each page's original order are preserved. The interface adds neutral directions, one statement per page and an explicit omission option. These changes support administration and progress tracking, but the digital implementation does not automatically inherit every psychometric property of previous administrations in different settings.

Spanish order differs from English order. By meaning, Spanish positions one through ten correspond to English positions 2, 1, 3, 4, 5, 7, 10, 6, 9 and 8. For example, the first Spanish statement concerns finding ways to obtain what one wants despite opposition; that content appears second in English. A complete sum does not change when the same ten contents are permuted. However, comparing answers by item number alone could compare different contents. The export preserves both a source-position identifier and a shared meaning identifier. Use the latter to align content across languages and the former to audit what participants actually saw in their administration.

The English documentation preferably recommends mixing the ten statements into a larger questionnaire with a compatible response format. This classroom implementation instead presents a standalone battery in fixed order, without adding other instruments. That is a declared administration adaptation, not a claim that fixed order is superior. Spanish retains the anchor “incorrecto,” which could be mistaken for an incorrect knowledge answer. Directions therefore

explain that it means the statement is untrue for the respondent. Silently replacing it with a strongly-disagree scale would create another form. Future research changes should document their source, wording and consequences for comparability, rather than inheriting a validation claim from the original label.

2.4 The complete sum and its units

Each substantive answer is an integer from one through four. Every statement is positively keyed; none is reversed. For a person providing all ten answers, their sum ranges from 10 to 40. Ten answers of three produce 30. The complete mean divides that sum by ten, giving 3.00 in the same example. Complete sum and complete mean contain exactly the same information and rank complete respondents identically. They are not separate dimensions or independent confirmations of a hypothesis. The report displays both to make scoring and the response scale easier to teach, not to multiply the number of measured constructs.

Adding responses treats them as comparable contributions to a summary. The numerical codes order verbal categories, but the psychological distance between one and two is not directly observed as a physical distance equal to that between three and four. Summation is a documented scoring convention supported by research on the instrument. That foundation does not make every interpersonal comparison cardinal. A sum of 40 does not indicate twice the capability of a sum of 20. This limitation helps instructors interpret the reported average without requiring every social measure to possess a naturally observable unit or abandoning careful use of established questionnaires.

The histogram uses six integer intervals: 10-14, 15-19, 20-24, 25-29, 30-34 and 35-40. Each complete sum enters once, so total frequency must match the number of available complete sums. These intervals are descriptive bins, not low and high self-efficacy categories. A concentration near 40 might suggest a ceiling in that cohort, commonly favorable responses or an influence of administration context. The distribution raises questions about variation and discrimination without assigning a diagnostic interpretation to a bar. Ask students to describe what the plot shows before proposing an explanation for its shape.

2.5 Missing data and the seven-response available mean

The primary sum requires ten substantive answers. Someone may submit all ten forms but omit one statement: the questionnaire is finished, yet a complete sum is unavailable. Another person may answer nine and remain on the tenth page: neither questionnaire completion nor a complete sum has been reached. Both are excluded from that score, without replacing the missing value by zero or the class average. Their records retain the distinction between an explicit omission and a pending page. The instructor can therefore distinguish scoring exclusion from interruption of the procedure and avoid treating every missing sum as withdrawal.

The authors' FAQ proposes computing a mean when no more than three statements are missing. This platform implements that rule as a separate sensitivity indicator: the mean of available values when at least seven substantive answers exist. Seven answers of two yield an available mean of 2.00 and a valid-item count of seven. The complete sum remains empty. Multiplying 2.00 by ten would produce 20, but that is not an observed sum of ten answers and the platform never presents it as one. Keeping both rules visible teaches how a missing-data decision changes coverage as well as interpretation.

The available mean can combine different sets of content. Seven observed answers do not necessarily constitute a validated seven-item short form; they are the statements that person happened to

answer. If omissions concentrate on less favorable descriptions, the available average could be biased. Showing the number of items per respondent and the N per measure reveals part of the issue without correcting it. Research should establish and justify a missing-data rule before analysis. A class should not choose whichever rule retrospectively produces the desired association. Instead, compare what the rules include and what assumptions would make either summary informative.

2.6 Evidence, limitations and classroom use

Scholz and colleagues examine GSE structure among 19,120 participants from 25 countries. Their analysis asks whether one general dimension represents responses across samples and finds support for that structure alongside cross-country differences requiring further investigation. This supports research on a general measure; it does not establish that sums from any two countries are directly interchangeable for every purpose. One-factor structure, equality of item relationships and equality of item starting points are different questions. Do not present dimensionality evidence as automatic proof of complete equivalence of means. A useful classroom discussion identifies which comparison an empirical study actually tested before extending its conclusion.

Reliability concerns measurement consistency under specified conditions; validity concerns justified interpretations and uses. An instrument may be internally consistent while remaining inappropriate for a decision outside its purpose. This guide does not turn GSE into an admission, hiring or clinical evaluation mechanism. The local cohort can be small, selected and responsive to context. There is no independent Mexican or digital validation of this interface. These limits define its classroom use while preserving opportunities to discuss theory, formulas and evidence precisely. They also help students recognize why software implementation quality and substantive measurement validity require different kinds of evidence.

The authors' FAQ does not endorse a universal cutoff for classifying people. This delivery imports no norms and does not split the class at its median. To teach how dichotomization changes a conclusion, use synthetic data and show the resulting loss of information. For the actual session, describe distributions and relationships cautiously. Appropriate feedback asks what was measured, what is missing and which new design could distinguish explanations. It does not require a student to defend a score or accept an identity assigned by software. The instructor's authority should support understanding the method rather than pressure participants to endorse an interpretation.

Attention

Spanish and English source positions have different orders. Align items using meaning identifiers. None is reversed. The seven-response mean is a separate analysis and is never converted into an imputed complete sum.

Performance forecasts and original problems

3.1 Why add a task-specific measure

The extension makes a conceptual distinction observable in a sequence. After GSE, the participant reads an arithmetic problem with four alternatives, records a probability of choosing the correct answer, and then answers. The forecast concerns a defined event: answering that problem correctly under the announced instructions. It does not ask about personal worth, confidence in everything or willingness to wager money. Observing a specific belief and its outcome allows discussion of why both differ from a general self-report. Order is fixed and nobody is randomly assigned to alternative conditions. The activity therefore does not estimate an effect of taking the questionnaire on subsequent performance.

The four problems are original BEER Lab teaching materials covering a budget remainder, bundle savings, compound growth and inventory balance. Their conditions are transparent enough for instructors to reconstruct every solution. They are not items copied from an intelligence test and have not been equated in difficulty. One student may recognize a solution quickly while another needs more time. There is no automatic time limit and speed does not enter scoring. Interpret performance on these four exercises under these conditions. Do not extend the outcome into a general cognitive ability score or a prediction of real economic decisions outside the activity.

3.2 Recording probabilities without confusing zero and omission

Forecasts accept integers from zero through one hundred. Zero expresses subjective certainty of not answering correctly; one hundred expresses certainty of answering correctly. These are not the questionnaire's four categories: a percentage probability is a different scale and describes a different object. Entering minus one explicitly omits the forecast; that technical value never enters a probability calculation. An empty field produces an explanation and can be corrected. The problem and alternatives are visible before the forecast is stored, so participants know which task the expectation concerns. They are not asked to anticipate an unknown problem's difficulty from its number or title alone.

Instructions require working without calculators, search engines or help from other people. This establishes a common condition for the teaching example; the platform does not monitor devices or guarantee compliance. Instructions also specify that no payments occur. Consequently, the activity makes no claim that an incentive mechanism elicits truthful beliefs. A declared probability may depend on understanding, effort or self-presentation. Evaluate its coherence without assuming that every entered number is a perfectly formed subjective probability. Classroom discussion can

BEER Lab

Four additional problems

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account: there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Four independent arithmetic problems follow. First read each problem and its four options and record your probability of selecting the correct answer, from 0 to 100. Then select an option. You cannot change a stored forecast. Do not use a calculator, search engine or help from others. You may omit either step. Solutions appear only at the end if enabled by the instructor. Amounts are hypothetical and there are no payments.

What happens after recording your forecast?

☐ I select an answer without changing my earlier forecast; there are no payments.

☒ I can revise my forecast after seeing the solution and receive the amount.

Your answer: I can revise my forecast after seeing the solution and receive the amount.. The correct answer is I select an answer without changing my earlier forecast; there are no payments.. The first option is correct: the forecast is stored before answering; no solution is shown during the task and there are no payments.

Continue

Figure 3.1: The comprehension check distinguishes prior forecasts, later answers and no payments.

explore what incentives or additional checks might change, but those proposed mechanisms are not part of the implemented design.

BEER Lab

Problem 1 / 4

First, your forecast

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account: there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Budget remainder

A hypothetical budget contains 120 ECU. Spending is 35 ECU and then 25 ECU. How many ECU remain?

1. 50 ECU
2. 60 ECU
3. 70 ECU
4. 85 ECU

What probability do you assign to selecting the correct answer to this problem?

Probability, from 0 to 100; enter -1 if you prefer not to answer

0 means you are certain you will not answer correctly; 100 means you are certain you will. Any integer between them is valid. Enter -1 to omit; 0 is a valid forecast.

Save forecast and continue

Figure 3.2: Read the problem and store a probability before selecting an answer.

Advancing stores the forecast separately from the editable form input. The following page shows the same conditions and allows an answer or omission without revising that stored forecast. No solution appears between problems. A corrected retrospective assessment therefore cannot masquerade as the earlier prediction through the ordinary flow. This restriction does not establish causality or eliminate learning across problems; it defines the observation sequence. If someone leaves after forecasting, the export preserves that forecast alongside a pending answer. It does not invent a failure, a zero-probability forecast or a completed pair to make the table easier to aggregate.

3.3 Solving the four problems

The budget begins at 120 ECU. Spending 35 and then 25 leaves 60: the calculation is 120 minus 35 minus 25. ECU means Experimental Currency Units and serves only as a hypothetical unit of account here. Nobody receives the remainder or converts it into money. The relevant quantity is a balance, not a return or experimental payment. This problem illustrates how a simple account can attract different subjective expectations. A correct answer alone does not reveal what probability was assigned before choosing it. Ask students to identify the initial stock and two outflows before

discussing any relationship with the questionnaire.

BEER Lab

Problem 1 / 4

Next, your answer

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account; there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Budget remainder

A hypothetical budget contains 120 ECU. Spending is 35 ECU and then 25 ECU. How many ECU remain?

Your forecast has already been stored.

Select one of the four options or indicate that you prefer not to answer.

Your response

☐ 50 ECU

☐ 60 ECU

☐ 70 ECU

☐ 85 ECU

☐ I prefer not to answer

Confirm response

Figure 3.3: The budget has a uniquely determined remainder of 60 ECU.

The purchasing comparison holds the four notebooks constant. Buying each for 18 ECU costs 72; the bundle costs 68; the difference is four ECU. There are no quality, quantity or risk differences to evaluate. Correctness means identifying that saving, not preferring a purchase or displaying a taste for discounts. If a participant forecasts zero and then chooses four, the forecast performs poorly on that trial even though the arithmetic answer is correct. This makes a useful distinction between evaluating a forecast and evaluating a solution. A low prediction error and a correct answer describe related but different aspects of an observation.

Compound growth begins with 100 ECU and adds ten percent of the current balance at each year end. The first balance is 110; the second addition is eleven; the final balance is 121. The alternative 120 applies two additions of ten to initial capital rather than following the announced rule. This provides an interpretable arithmetic error without establishing that the respondent has lower self-efficacy or a particular time preference. Belief and knowledge are recorded using different measures, and analysis preserves that distinction. The exercise describes a hypothetical account, not an investment recommendation or a choice among risky financial products.

BEER Lab

Problem 2 / 4

Next, your answer

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account; there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Comparing two purchases

Four notebooks cost 18 ECU each when bought separately. A bundle of the same four notebooks costs 68 ECU. How much does choosing the bundle save?

Your forecast has already been stored.

Select one of the four options or indicate that you prefer not to answer.

Your response

☐ 4 ECU

☐ 8 ECU

☐ 12 ECU

☐ 16 ECU

☐ I prefer not to answer

Confirm response

Figure 3.4: The saving compares identical goods and subtracts 68 from 72 ECU.

BEER Lab

Problem 3 / 4

Next, your answer

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account; there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Growth of a balance

A hypothetical account starts with 100 ECU. At the end of each year, 10% of its current balance is added. There are no fees, deposits or withdrawals. What is the balance after two years?

Your forecast has already been stored.

Select one of the four options or indicate that you prefer not to answer.

Your response

☐ 110 ECU

☐ 120 ECU

☐ 121 ECU

☐ 122 ECU

☐ I prefer not to answer

Confirm response

Figure 3.5: Second-year growth applies to 110 rather than the initial principal.

Inventory begins at 24 units, adds six and subtracts nine, leaving 21. Physical units are neither ECU nor money. Excluding other inflows and outflows makes the solution unique. During discussion, identify the initial stock and each signed flow before calculating. An error might reflect reading, arithmetic or distraction; the record cannot identify which. Solutions are disclosed only at the end and with instructor authorization. The guide can therefore explain the procedures after administration without supplying hints while people are answering. A correct calculation demonstrates completion of that particular problem, not competence across every setting where inventories matter.

BEER Lab

Problem 4 / 4

Next, your answer

Problems use hypothetical amounts. ECU means Experimental Currency Units and is only a unit of account: there is no endowment or conversion into money. Work without a calculator or outside help; there is no time limit.

Inventory balance

A store has 24 units. It receives 6 new units and sells 9. There are no other inflows or outflows. How many units remain?

Your forecast has already been stored.

Select one of the four options or indicate that you prefer not to answer.

Your response

☐ 15 units

☐ 18 units

☐ 20 units

☐ 21 units

☐ I prefer not to answer

Confirm response

Figure 3.6: Inventory adds inflows and subtracts outflows from the starting stock.

3.4 Binary squared error and the signed gap

To evaluate a valid forecast, divide its percentage by one hundred and call the resulting probability p . Define y as one for a correct substantive answer and zero for an incorrect substantive answer. The binary Brier error used here is the square of p minus y , ranging from zero to one. If p is 0.8

and the answer is correct, the error is 0.04. With the same forecast and an incorrect answer it is 0.64. Lower error represents a better forecast of the observed event. The number of answer alternatives does not change the formula because the forecasted event remains answering correctly versus not answering correctly.

Some conventions add squared errors across two categories and double this binary value. This implementation uses only the square of p minus y , as its report specifies. Establish the convention before comparing against an external file. The signed gap instead subtracts one hundred times y from the declared percentage. A forecast of 80 followed by success produces minus twenty percentage points; followed by failure it produces plus eighty. Mean gap summarizes error direction among observed pairs, but can hide large errors with opposite signs. It should not be read as the same statistic as squared error or as a personal overconfidence classification.

Theoretical foundation

Forecasting 80% and succeeding gives Brier 0.04; forecasting 20% and failing also gives 0.04. Their gaps are -20 and +20 points. Mean gap is zero even though neither forecast equals its outcome. Use both measures and retain the valid-pair denominator.

3.5 What good forecasting would mean

Calibration relates declared probabilities to frequencies across comparable events. Among many events assigned probabilities near eighty percent, an accuracy rate near eighty percent would be relevant evidence. One success following an eighty-percent forecast does not alone prove or disprove calibration. Four tasks likewise cannot establish a stable personal property. The class can calculate errors and discuss the observations needed for calibration research, but should avoid assigning the term as an individual label automatically derived from this small battery. Distinguish the arithmetic of one realized loss from the empirical pattern that would support a conclusion about repeated forecasts in a defined population.

A pattern consistent with informed task-specific beliefs would be lower squared error when comparing equivalent tasks and groups under comparable conditions. There is no necessary prediction that participants with higher GSE sums must always answer more problems correctly. GSE concerns broad demands, while the exercises require particular knowledge. An association could reflect experience, selection or common circumstances. Identifying a mechanism would require additional conditions, relevant measurements and an identification strategy. This extension supports formulating those questions; observing two related columns does not supply the missing design. Ask what a proposed comparison would hold constant before describing it as a test of theory.

Reports, data and classroom discussion

4.1 Disclosure and counters with clear units

The report initially displays progress while aggregates remain hidden. Revealing them requires instructor authentication. Keep the access-code panel closed during projection. Four counters distinguish participants, completed activities, submitted questionnaires and GSE records with ten scoreable answers. A submitted questionnaire can include omissions. An extended activity can remain pending after its GSE is complete. These differences concern procedure and inclusion rules, not automatic evidence of insufficient effort. Before discussing outcomes, ask the class to describe each counter's unit and inclusion condition. This establishes which population a subsequent mean or proportion describes and prevents progress statistics from being treated as personality measures.

The first table shows complete sum, complete mean and the available-item mean requiring at least seven answers. Each row includes valid N and excluded count, along with mean, median and sample standard deviation when at least two observations exist. With one complete sum, an average can be displayed but sample dispersion cannot be estimated with a zero denominator. A dash indicates an unavailable calculation rather than a zero result. The median describes position without removing the need to inspect the distribution or explain exclusions. Mean and median are summaries of the respondents included by that row's rule, not of everyone who opened a link.

4.2 Reading distributions and sensitivity rules

Only complete sums enter the histogram. Its observation count matches the first table row, not all participants and not necessarily the available-mean N. Suppose one person provides ten answers of three and another seven answers of two. Complete sum has N one and mean 30; complete mean has N one and value three; available mean has N two and mean 2.50. This is more than changing units: the included people changed. Comparing these averages as if they were transformations of one aggregate would be misleading. Reconstructing this example helps students understand why the report repeats denominators beside apparently similar summary statistics.

Allowing seven answers increases coverage without establishing that missing responses are ignorable. Examine which statements are omitted and how missingness is distributed. Differences between scoring rules are not a definitive correction for bias. If a conclusion changes when requiring ten answers rather than permitting seven, preserve that sensitivity in the interpretation. For a short class, reconstruct two or three synthetic profiles, state the rules and ask what evidence could justify either approach in a research study. Do not claim that the available mean recreates the actual answers a respondent would have supplied to the missing statements.

BEER Lab

General self-efficacy instructor report (GSE)

General Self-Efficacy Scale (GSE-10) · Extended · en

Language version: Ralf Schwarzer & Matthias Jerusalem (1995), official English form.

[▶ Class access code](#)

Participants

3

Activities finished

1

Questionnaires finished

1GSE with ten scoreable
responses**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

GSE: Schwarzer & Jerusalem (1995); Spanish adaptation: Bäßler, Schwarzer & Jerusalem (1993), Bäßler & Schwarzer (1996).

Complete sum: 10–40, with no reversed items. Complete mean: sum/10. The available-item mean with at least seven responses is a separate sensitivity analysis; no imputed sum or cutoffs are used.

Figure 4.1: Progress remains visible while score and task aggregates are hidden.

BEER Lab

General self-efficacy instructor report (GSE)

General Self-Efficacy Scale (GSE-10) · Extended · en

Language version: Ralf Schwarzer & Matthias Jerusalem (1995), official English form.

► Class access code

Participants

3

Activities finished

1

Questionnaires finished

1

GSE with ten scoreable responses

1

Hide results

Download this session's responses and scores

Indicator	Valid N	Without score	Mean	Median	Standard deviation
Complete GSE sum (10–40)	1	2	40.00	40.00	—
Complete GSE mean (sum/10; 1–4)	1	2	4.00	4.00	—
Available-item mean ≥ 7 responses (sensitivity; 1–4)	1	2	4.00	4.00	—

Figure 4.2: Each measure retains its own valid N and excluded records.

Distribution of total scores

N = 1 · Frequency

10–14	0
15–19	0
20–24	0
25–29	0
30–34	0
35–40	1

Figure 4.3: Histogram frequencies add to the number of complete GSE sums.

4.3 Task tables and descriptive comparisons

Every problem reports how many participants supplied a forecast, how many answered and how many supplied both. Omissions and pending records are separately counted for each step. Overall mean forecast uses all valid forecasts, while accuracy uses all substantive answers. Rows identified as valid pairs recompute forecast and accuracy on the same set of observations. Only these pairs enter Brier and gap. Subtracting a mean forecast from an accuracy rate based on different people can create a misleading comparison. The report shows both sets of denominators so students can identify whether a difference reflects outcomes, missingness or a change in who contributes data.

Budget remainder		
Forecast N: 1 · Omitted: 0 · Pending: 2		
Answer N: 1 · Omitted: 0 · Pending: 2		
Indicator	Valid N	Mean
Mean forecast (%)	1	80.000
Accuracy (%)	1	100.000
Mean forecast (%) (Valid pair N)	1	80.000
Accuracy (%) (Valid pair N)	1	100.000
Binary Brier [0,1]	1	0.040
Gap (percentage points)	1	-20.000
Your response	N GSE	Mean GSE
Incorrect answer	0	—
Correct answer	1	40.000

Figure 4.4: Task tables separate forecasts, answers, valid pairs and prediction errors.

A second table describes complete GSE sums among correct and incorrect answers to each problem. It requires a substantive task answer and all ten GSE responses. Its N can therefore differ from answer N and valid forecast-pair N. Someone answering correctly but omitting a GSE item contributes to accuracy without contributing to that conditional mean. The two rows do not compare randomly assigned groups. A difference can motivate questions about experience or capability beliefs, but does not estimate the effect of increasing self-efficacy on knowledge or the effect of knowledge on self-efficacy. Identify the selection rule before interpreting either row as evidence about a mechanism.

4.4 Downloading and reconstructing calculations

The report download contains only the selected session and thirty-six columns. Each person contributes ten item rows and three measure rows; extended editions add four task rows. Three participants therefore produce 39 basic data rows or 51 extended data rows, excluding the header. Active pending rows remain present. The native wide export instead provides one row per participant with technical-period columns. Both formats describe the same records but have

different row counts. Instructors must identify format before counting people or decisions. A long file's repeated participant key is expected and should not be treated as evidence of duplicated participants or repeated administrations.

On an item row, original response, anchor label and status support a scoring audit. Meaning identifiers align languages without confusing source positions. On a task row, the percentage forecast has its own status and correctness exists only for substantive answers. If an answer exists but a forecast is missing, accuracy remains available and Brier remains empty. If the forecast is zero and an answer exists, zero remains a valid observation. Instrument version, content fingerprint, scoring version and task version identify the procedure that produced the file. These fields allow a reviewer to distinguish a changed interface or source form from an unexplained scoring discrepancy.

Do not aggregate derived-measure rows as additional people. One person occupies several rows because the file is long. Technical participant keys identify observations within the session but do not automatically link instruments run in separate sessions. Keep rehearsals separate from the actual cohort. Preserve individual answers outside projected material and use aggregates for discussion. Before external analysis, retain pending and omitted statuses and then choose a missing-data treatment deliberately. An automatic replacement with zero would destroy distinctions that the report preserves. The codebook describes each column so an analyst can reconstruct the same inclusion rules rather than infer them from an empty cell alone.

4.5 Private feedback and interpretation limits

The basic edition closes without displaying an individual profile. In extended editions, after every active response page has been submitted, instructor disclosure allows the personal screen to show three GSE summaries, valid-item counts and solutions to the four tasks. Results are not sent to other class members. Hiding results withdraws this feedback on the next polling update. After a local server restart, an already open page needs reloading to reconnect its live channel. Disclosure does not authorize projection of personal profiles. Treat control of the session view and responsible handling of individual observations as separate parts of running a classroom activity well.

Use feedback to reconstruct operations and measurement decisions. A compound-interest error can be discussed through the second-year balance without concluding that a student lacks self-belief. A correct answer following a low forecast invites questions about available information and the meaning of the declared number. Avoid requesting personal accounts of hardship or associating names with histogram bars. The teaching purpose is learning how data are constructed and interpreted, not persuading each participant that a score reveals their true identity. Invite disagreement about interpretation while preserving the exact arithmetic needed to verify what the software reports.

4.6 Inference errors to anticipate

A score is not an unchanging essence or a moral identity. It is also not a recommendation for selecting partners or assigning educational opportunities. The activity teaches measurement and decisions. When a student finds a result surprising, invite consideration of item content and response context. Do not interpret disagreement with a profile as resistance to a truth discovered by the platform. A measure has limitations even when its formula is correctly implemented. Taking those limitations seriously makes the exercise more rigorous and can improve the quality of the classroom conversation.

Reliability and validity are not synonyms. Consistent responses may measure something different from the intended construct, and a broad dimension can include diverse behaviors. Do not use

BEER Lab

Activity complete

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

Your descriptive results in this activity

Complete GSE sum (10–40): 40 (n=10/10) / 40

Complete GSE mean (sum/10; 1–4): 4.00 (n=10/10) / 4

Available-item mean ≥ 7 responses (sensitivity; 1–4): 4.00 (n=10/10) / 4

Budget remainder: Your answer: 60 ECU. Mean forecast (%): 80%. Binary Brier [0,1]: 0.040. Solution to this task: $120 - 35 - 25 = 60$ ECU.

Comparing two purchases: Your answer: 4 ECU. Mean forecast (%): 0%. Binary Brier [0,1]: 1.000. Solution to this task: $4 \times 18 = 72$; $72 - 68 = 4$ ECU saved.

Growth of a balance: Your answer: 121 ECU. Mean forecast (%): —. Binary Brier [0,1]: —. Solution to this task: Year one: $100 \times 1.10 = 110$. Year two: $110 \times 1.10 = 121$ ECU.

Inventory balance: Your answer: Omitted. Mean forecast (%): 90%. Binary Brier [0,1]: —. Solution to this task: $24 + 6 - 9 = 21$ units.

The complete sum requires ten responses; the available-item mean requires at least seven. Additional tasks and forecasts are distinct measures. No diagnoses, percentiles or grades are assigned.

Figure 4.5: Authorized feedback keeps scale, forecasts and task solutions separate.

a coefficient calculated in a small class as a definitive certificate for the instrument. A formal study should specify hypotheses, procedure, population, analysis and handling of missing values. The classroom session should retain a descriptive reading and turn limitations into questions that students can understand and assess. That approach allows methodological sophistication without requiring the instructor to pretend that one short activity has resolved every measurement problem.

4.7 Class plan and concluding discussion

For fifty to seventy-five minutes, allocate about five to access and rules; three to five for basic administration or seven to twelve for extended administration; ten to reconstructing sums and missingness; fifteen to forecasts and solutions; and the remainder to evidence and conclusions. These are planning estimates, not validated timings or imposed limits. If the group needs longer, shorten discussion or select fewer subsequent paper exercises without interrupting the digital administration. Preserve an equal opportunity to read. Speed and self-efficacy are different objects, and changing the available time halfway through the activity would complicate comparisons the class might otherwise make.

Feedback script

Begin with observations: how many complete sums exist, which rule changes when seven answers are allowed, and which forecasts have verifiable answers? Then connect beliefs with performance: why might they agree or differ? End by identifying another observation or design that could distinguish two explanations.

Useful preparation is solving this guide's synthetic examples before opening a session. During administration, address procedural questions without recommending agreement levels or confidence values. Before disclosure, close the access-code panel and confirm that the records belong to the intended cohort. After discussion, ask students to write a short argument separating measurement, outcome and inference. That product assesses methodological learning without using personal questionnaire answers as grades. An effective assessment rewards clear definitions, correct denominators and justified conclusions even when students propose different explanations for the same aggregate pattern.

Conclude by formulating a research question. One team might investigate whether specific practice reduces forecast errors; another whether different tasks require more specific efficacy measures; another how omissions relate to item content. Each proposal should specify an analysis unit, a comparison, an outcome and a missing-data strategy. The present session provides a shared experience for forming these questions, but new designs need their own specification and evaluation. Mentioning a possible intervention during feedback does not make it an implemented treatment. Distinguish the activity students completed from the study they would need to conduct to support their proposed explanation.

Checkpoint

Check session, language, edition and access. Review N per measure, choose disclosure and download only the intended cohort. Keep tasks and scale separate.

Worked problems and solutions

These cases use synthetic people and data. Solve them before opening a real session. First identify the measured object, then reconstruct the calculation, then state which conclusion remains unsupported. A complete solution explains its denominator rather than merely reproducing a number on the screen. The Spanish guide contains the same cases, covering scoring rules, missing states and the separate evaluation of performance forecasts. Use the explanations to prepare discussion rather than distribute correct task answers before administration.

What participants will see

First solve cases with synthetic data. Identify response, status, inclusion rule, numerator and denominator. For tasks, write prior forecast, outcome and error in separate columns before comparing with GSE.

5.1 Scoring and alignment between sources

1 A complete sum

The ten responses are 1, 2, 3, 4, 2, 3, 4, 1, 3 and 4. Their sum is 27 and complete mean is 2.70. Nothing is recoded. Available-item mean is also 2.70 because all ten values exist. Three derived rows do not represent three dimensions: two express exactly the same complete information, and the third applies a coverage rule that happens to coincide here. Ask what would change if one response were missing before comparing the exported columns.

2 Scale endpoints

Ten answers of one yield sum 10 and mean one. Ten answers of four yield sum 40 and mean four. These are arithmetic bounds rather than clinical thresholds. A sum of 20 does not indicate twice the capability of a sum of 10, because response coding has no natural zero of capability. This exercise verifies the software range and opens discussion of why ordered scores cannot automatically be read as ratios of psychological quantities. The numerical operation is precise even when its interpretation has limits.

3 No statement is reversed

A participant selects four on Spanish statement seven. Its contribution is four, just as at every other position. Applying five minus response would give one and incorrectly alter the sum. Such a

mistake can arise from reusing another questionnaire's reversal routine. The solution is to attach scoring keys to the identified GSE source and check an all-equal response profile. Do not guess a key from item position, a remembered form or the assumption that every psychometric questionnaire must contain negatively keyed statements.

4 Two positions, one content

Spanish position one and English position one do not share content: the first Spanish statement corresponds in meaning to English position two. Align using meaning identifier GSE-02 while retaining both source-position identifiers. Complete summation still uses ten contents in either language, but a correlation between two columns called item one would be a different comparison. Semantic alignment also does not remove the need to investigate measurement equivalence between populations. Distinguish locating equivalent wording from establishing that every inferential use transfers unchanged.

5.2 Missing data and denominators

5 Seven answers of two

Someone answers seven statements with two and leaves three pending. Available mean is fourteen divided by seven, or 2.00. Complete sum and mean do not exist. Fourteen is a partial total, not a complete GSE score. Twenty, obtained by multiplying the mean by ten, is not an observed sum either. The file records seven valid items and the sensitivity indicator. The questionnaire remains pending because three forms have not been submitted. Explain both scoring and procedural status when reporting this case.

6 Six answers of four

With six answers of four, a partial arithmetic mean is four, but the platform does not report available mean because the minimum of seven is unmet. All three measures remain unavailable and valid-item count is six. This is not a score of zero. The case shows how a mathematically possible calculation may fail the announced measurement rule. Changing the threshold after seeing the record would alter the procedure and must be declared as another analysis, rather than silently presented as standard scoring.

7 Nine favorable answers

Nine answers of four and one explicit omission produce an available mean of four and no complete sum. Even if all ten pages were submitted, one substantive value is missing. The platform counts a finished questionnaire and nine valid answers without replacing the omission with four. Completion and scoreability are distinct conditions. Do not attribute the missing sum to a server fault before checking exported response statuses. The reported available mean is conditional on observed content and does not recreate the unobserved answer.

8 Omitted versus pending

Two people have nine valid answers. One explicitly omitted the tenth; the other has not opened that page. Both can have an available mean, but only the first finished the questionnaire. Scoring

does not distinguish their means if their observed values are identical, while progress tracking distinguishes their procedural situations. Omitted and pending remain separate export states. Ask which analysis needs that distinction and which uses only observed values. Keeping both states makes the record more useful without forcing every analysis to treat them identically.

9 Three people and different N

One person has sum 30, another 40, and a third has only seven answers of two. Complete sums have N two and mean 35; complete means have N two and average 3.50. Available mean includes all three and averages 3.00. Only 30 and 40 enter the histogram. The difference between 3.50 and 3.00 reflects a different included population as well as the chosen rule. Do not describe the latter as merely converting the mean sum into another unit for the same people.

10 Missing content matters

Two people answer seven statements and both obtain available mean three. One omitted the last three contents, the other the first three. Their values coincide while their observed content differs. The solution is not to assign an identical supposed true score. Identify missing content and ask under what assumptions the means would be comparable. The seven-answer rule is a documented coverage convention, not automatic validation of every possible seven-statement combination. This is why valid-item count is informative but cannot fully describe missingness on its own.

11 Sample dispersion

Three complete sums are 20, 30 and 40. Mean and median are 30. Squared deviations sum to 200; dividing by three minus one gives sample variance 100 and sample standard deviation 10. With only one available sum that denominator cannot be used, so the report shows a dash. Keep the unit, sum points, and distinguish dispersion between respondents from uncertainty about construct validity. A small standard deviation does not establish validity. A missing standard deviation does not imply all participants gave the same answer.

5.3 Forecasts and observed errors

12 Eighty percent followed by success

A participant forecasts 80% and answers correctly. Probability p is 0.8 and outcome y is one. Binary Brier is the square of minus 0.2, or 0.04; signed gap is minus twenty percentage points. The forecast is less extreme than the realized outcome, but that isolated trial cannot establish systematic underestimation. An eighty-percent probability allows successes and failures. Calibration requires a relevant set of comparable observations. Evaluate the realized prediction loss precisely while keeping its arithmetic distinct from a claim about a stable personal tendency.

13 The same forecast followed by failure

Keep the 80% forecast but change the answer to incorrect. Outcome y is now zero, Brier becomes 0.64 and gap is plus eighty points. The contrast shows why the observed outcome matters for forecast evaluation. Do not retrospectively replace 80% because the respondent might have declared less after discovering the error. That revision uses later information and changes the measured

object. The platform retains the forecast stored before answering. A later belief update would require its own separately identified observation and timing.

14 Zero does not mean omission

A zero forecast followed by success produces Brier one and gap minus one hundred. If the forecast is missing and the answer correct, accuracy remains one but Brier and gap are unavailable. Minus one is the technical omission input, not a negative probability. Preserve status separately from value. Replacing absence with zero would invent a declared certainty of failure. This example is especially useful when importing the CSV into software that might otherwise fill empty numeric cells automatically during data cleaning.

15 Different people behind two means

A forecasts 90% and succeeds; B forecasts 60% and leaves the answer pending; C omits forecasting and fails. Forecast N is two with mean 75%. Answer N is two with accuracy 50%. Only A forms a valid pair: forecast 90%, accuracy 100%, Brier 0.01 and gap minus ten points. Subtracting 50 from 75 would give plus twenty-five points across different groups. That is not the gap implemented by the report. Matching denominators prevents attributing this difference to observed errors for one common group.

16 Gaps that cancel

Consider two trials: 80% with success and 20% with failure. Their gaps are minus twenty and plus twenty, averaging zero. Both Brier values are 0.04, and their mean is also 0.04. Zero mean gap does not mean perfect prediction on every trial. Nor can two observations certify stable calibration. Direction and magnitude of error answer different questions, and one summary can hide relevant patterns. A good explanation shows both calculations, states the sample size and avoids assigning a personal confidence label from their signs.

17 Compound growth and the response object

An account begins at 100 and grows ten percent of its current balance for two years. Correct balance is 121. If someone forecast 50% and selected 120, correctness is zero, Brier 0.25 and gap fifty points. This is not a saving decision or a choice between dates; it is an answer to a calculation rule. The example explains why the same economic context can produce knowledge, belief or preference data depending on the question actually asked. An economic label alone does not determine what a response measures.

18 Conditional GSE means

Two respondents answer correctly: one has complete GSE sum 40 and the other lacks a complete sum. A third answers incorrectly and has sum 30. Accuracy uses N three and equals two thirds. Mean GSE for correct answers uses N one and equals 40; for incorrect answers it uses N one and equals 30. Neither group automatically has a valid forecast. The ten-point difference is descriptive and selected by answer and score availability, not a causal effect of changing GSE. State both inclusion conditions before interpreting it.

5.4 Operation, disclosure and inference

19 Counting a long file

Three basic participants generate thirty item rows and nine derived rows, totaling 39. Extended editions add twelve task rows, totaling 51. The header is not an observation. Pending participants retain active rows, so each row must not be counted as a completed response. Use technical participant keys to recover the number of people and response status to count substantive answers. These operations answer different questions about the same download. The native wide file changes organization, not the number of people who participated.

20 A forecast cannot be revised afterward

Someone stores 50%, advances to answering and wants to change the forecast to 90%. The sequence offers no such revision. Evaluation still uses 50%, even if another value appeared in an old form input. The implementation preserves a separate locked value. Studying belief revision would require a second identified measurement and a specified new design. Do not overwrite the initial forecast to represent an update, because that would destroy the information order needed to interpret forecast error and distinguish prediction from hindsight.

21 Complete questionnaire, pending task

A participant answered all ten statements but submitted only three of four tasks. Their sum can enter the GSE aggregate before the extended activity ends. Private feedback remains blocked until activity completion and instructor disclosure both occur. Results and solutions then appear; hiding withdraws them on the next update. Availability for analysis and availability for feedback are different conditions. Do not distribute solutions to people still completing problems in the same activity. Progress counters help instructors plan discussion without fabricating missing responses or withholding already valid aggregate observations.

22 Designing the next study

Higher GSE is observed among correct respondents. Propose at least two explanations, such as prior experience and selection into the cohort, alongside a hypothesis about beliefs. Studying an effect of specific practice could require a randomly assigned comparison and outcomes defined before observation. That design is not implemented here and would not automatically establish a change in general capability. A good answer identifies the missing comparison, the relevant outcome and the conclusion that would still remain unsupported. Treat the classroom result as a basis for research design rather than a substitute for it.

References and provenance

Schwarzer, R., and Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, and M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs (pp. 35-37). Windsor: NFER-NELSON. Official English form and documentation: <https://userpage.fu-berlin.de/~health/engscal.htm>.

Bäßler, J., Schwarzer, R., and Jerusalem, M. (1993). Official Spanish adaptation: <https://userpage.fu-berlin.de/~health/spanscal.htm>. Validation reference provided by the authors: Bäßler, J., and Schwarzer, R. (1996). Evaluación de la autoeficacia: Adaptación española de la escala de autoeficacia general. Ansiedad y Estrés, 2(1), 1-8.

Schwarzer, R. (October 2005). Everything you wanted to know about the self-efficacy scale but were afraid to ask. Authors' FAQ: https://userpage.fu-berlin.de/~health/materials/faq_gse_12oct05.pdf. Includes permission for study use and reproduction with attribution, absence of a cutoff, and the available-item mean rule requiring at least seven answers.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84(2), 191-215. DOI: 10.1037/0033-295X.84.2.191. Supports the distinction between efficacy and outcome expectations; the four tasks in this guide are original teaching exercises.

Scholz, U., Gutiérrez-Doña, B., Sud, S., and Schwarzer, R. (2002). Is general self-efficacy a universal construct? Psychometric findings from 25 countries. European Journal of Psychological Assessment, 18(3), 242-251. DOI: 10.1027//1015-5759.18.3.242. Author-hosted source: https://userpage.fu-berlin.de/~health/self/gse-25countries_2002.pdf.

American Meteorological Society. Brier score. Glossary of Meteorology. <https://glossary.amsoc.org/wiki/brier-score/>. This implementation uses the binary squared-error convention between zero and one, without payments or a claim of incentivized elicitation.

American Economic Association. JEL Classification System. <https://www.aeaweb.org/econlit/jelCodes.php?view=jel>. Verified descriptors: A22, undergraduate teaching; C83, survey methods; C91, individual laboratory behavior; D84, expectations.

Sources and institutional navigation consulted on 17 October 2026. Source hashes and digital adaptations are documented in source_manifest.json. No independent Mexican or digital validation is claimed. Guide and original teaching tasks: Stanislao Maldonado, BEER Lab.