Measurement Before Mathematics: An Evidence-Centered Protocol for Sampling, Instruments, and Analysis in School Research

Reysan Dugay

ABSTRACT

Quantitative analysis is credible only when the numbers entering it represent the intended constructs and the sampled observations support the population claim. This conceptual article develops an evidence-centered protocol for school-based quantitative research from a practical research guide. The protocol links construct definition, sampling, instrument development, administration, data quality, and statistical analysis through a chain of validity evidence. It distinguishes target population from accessible population and analytic sample; separates reliability from validity; and treats missing data, response processes, scoring, and assumption checks as integral parts of the study rather than technical afterthoughts. A measurement-risk matrix and a staged instrument-review process are proposed for student researchers and advisers. The article argues that statistical sophistication cannot rescue weak measurement or biased selection and that transparent reporting should include uncertainty, effect magnitude, and data-quality decisions. The protocol is a methodological synthesis, not a report of collected participant data.

Keywords: measurement; sampling; research instruments; data quality; statistical analysis; educational research

1. Introduction

Educational researchers often reach the statistics chapter with a completed questionnaire and a spreadsheet, only to discover that the variables are ambiguous, the response categories do not match the construct, or the available sample cannot support the intended conclusion. At that point, statistical refinement has limited value. The decisive work occurred earlier: defining what should be observed, selecting who could provide evidence, designing how responses would be produced, and protecting data quality during collection.

The source guide covers quantitative designs, sampling, instrument construction, intervention, data collection, processing, tables, graphs, statistics, conclusions, and reporting. This article reorganizes those topics around an evidence chain. The central proposition is that measurement must precede mathematics: before choosing a test, the researcher must know what the scores mean, whose observations they represent, and what uncertainty and bias remain.

2. Synthesis Approach

Content from the submitted guide was classified into six evidence stages: construct definition, sampling, instrument design, administration, data preparation, and analysis/reporting. At each stage, the analysis identified a principal risk, the evidence needed to manage that risk, and the reporting statement required for transparency. The resulting protocol draws distinctions that are frequently blurred in novice work, particularly reliability versus validity, sample size versus representativeness, and statistical significance versus educational importance.

3. Stage One: Define the Construct and Intended Use

A construct is not measured merely because a questionnaire gives it a label. Researchers should define the domain, intended interpretation, unit of analysis, response period, and use of the resulting score. If 'research readiness' is the construct, the study must decide whether it refers to factual knowledge, procedural skill, self-efficacy, access to resources, or a composite. Each interpretation requires different evidence.

An evidence-centered specification can be written in four sentences: What claim will be made about a respondent? What observable behavior or response would support that claim? What task or item will elicit the behavior? How will responses be scored? This sequence prevents item writing from becoming an unstructured collection of plausible questions.

4. Stage Two: Connect Population, Sampling Frame, and Sample

4.1 Three populations that should not be confused

The target population is the group to which the researcher hopes to generalize. The accessible population is the portion that can realistically be reached. The analytic sample is the set of cases remaining after recruitment, consent, participation, and data-quality exclusions. Differences among these groups define the boundary of inference.

4.2 Sampling method and coverage

Probability sampling assigns known selection chances and can support population estimates when the frame covers the target and nonresponse is addressed. Stratified or cluster designs may be appropriate for schools, but clustering changes precision and analysis. Convenience sampling is often practical for classroom research, yet its limitations cannot be repaired by calling the sample random after data collection.

4.3 Sample size as precision, not legitimacy

Sample-size planning should follow the intended estimate or test, expected variability, clustering, attrition, and desired precision or power. A very large biased sample may estimate the wrong population precisely. Conversely, a smaller probability sample may better support a bounded population inference. Researchers should report how the size was planned and how many cases were invited, responded, excluded, and analyzed.

5. Stage Three: Develop and Evaluate the Instrument

Risk	Diagnostic question	Evidence or corrective action
Construct underrepresentation	Do items cover the full intended domain?	Blueprint, expert mapping, and coverage review
Construct-irrelevant variance	Do reading load, language, format, or access affect scores?	Cognitive interviews, accessibility review, and administration supports
Ambiguous response process	Do respondents interpret items as intended?	Think-aloud or cognitive pretesting
Unstable scores	Would scores be sufficiently consistent for the intended use?	Internal-consistency, interrater, or temporal evidence as appropriate
Biased scoring	Are rules applied consistently across respondents?	Scoring guide, rater training, double scoring, and audit
Poor category function	Do options distinguish meaningful levels?	Pilot distributions, item analysis, and category revision

Table 1. Measurement risks and evidence-centered responses.

5.1 Blueprint before items

An instrument blueprint allocates items across domains and cognitive demands. It documents which research question and construct facet each item serves. Redundant items may improve consistency, but repetition should not crowd out important content. Response formats should match the construct: frequency options for behavior, agreement options for judgments, performance tasks for skills, and records for events when accurate records exist.

5.2 Expert review and response-process evidence

Experts can evaluate relevance, clarity, and domain coverage, but expert agreement alone does not show how intended respondents interpret an item. Cognitive interviews or small think-aloud trials reveal confusing terms, hidden assumptions, and answer strategies. Translation requires conceptual and functional equivalence, not word-for-word substitution.

5.3 Reliability and validity are not synonyms

Reliability concerns consistency under specified conditions; validity concerns the justification for an interpretation and use of scores. A highly consistent scale may measure the wrong construct. A low reliability estimate may reflect heterogeneous content, few items, restricted score variation, or inconsistent administration. Researchers should choose the reliability evidence that matches the scoring process and avoid treating one coefficient as universal certification.

6. Stage Four: Standardize Administration and Protect Data Quality

Administration procedures influence responses. Recruitment language, timing, privacy, device access, proctoring, item order, and assistance should be documented. In school settings, power relationships require particular care: learners should understand voluntariness where participation is optional, and teachers should avoid implying that research responses affect grades or services.

Data management should specify identifiers, coding, storage, access, and correction rules before collection. Range checks, duplicate detection, skip-pattern checks, and an audit trail reduce avoidable errors. Changes to raw data should never be silent; the cleaned dataset should remain traceable to the original record.

7. Stage Five: Prepare Data Without Manufacturing Certainty

Missing values are information about the data-generating process, not blank cells to be ignored automatically. Researchers should report the amount and pattern of missingness, examine whether missingness differs across groups, and justify deletion or imputation. Reverse-coded items, composite scores, outlier rules, and exclusion criteria should be defined and checked reproducibly.

Tables and graphs should reveal rather than decorate. Axes, denominators, category definitions, and units must be explicit. Means may be misleading for severely skewed distributions; percentages may be misleading when denominators are small or exclusions differ. Data visualization is part of analysis because it exposes patterns and anomalies that summary statistics can conceal.

8. Stage Six: Match Analysis to Design and Measurement

Analysis begins with the research question, variable type, design, and dependence structure. Descriptive statistics should precede inferential tests. Categorical outcomes may require proportions or models suited to categories; counts may require count models; ordinal responses should not be treated as interval measurements without justification. Repeated observations and learners nested within classes or schools violate independence assumptions of simple procedures.

Assumption checks should be planned, not used as a ritual to select whichever test yields significance. Researchers should examine model fit, residual patterns, influential observations, and the plausibility of the functional form. Robust or alternative analyses may be used when justified, with decisions reported transparently.

A complete result includes an estimate, its uncertainty, and its educational meaning. A p-value does not measure effect size, practical importance, or the probability that a hypothesis is true. Confidence intervals communicate a range of values compatible with the model and data, while effect measures allow readers to judge magnitude in context.

9. A Staged Review Protocol

Before launch, a proposal team can conduct four gates. Gate 1 - construct: approve definitions, score interpretations, and the blueprint. Gate 2 - sample: approve the target population, frame, recruitment, size rationale, and nonresponse plan. Gate 3 - instrument: complete expert review, cognitive pretesting, pilot administration, scoring checks, and revision. Gate 4 - analysis: lock primary questions, variables, exclusions, missing-data rules, and planned analyses before examining outcomes. Changes remain possible but should be labeled exploratory and documented.

10. Reporting the Evidence Chain

A transparent report lets readers reconstruct how observations became conclusions. The methods section should describe recruitment flow, instrument origin and adaptation, item examples or availability, scoring, evidence for score use, administration conditions, data cleaning, missingness, and analysis choices. Results should distinguish planned and exploratory analyses. Limitations should identify the likely direction and consequence of major measurement and selection risks rather than merely list generic weaknesses.

11. Conclusion

In quantitative school research, credible numbers are produced through a chain of evidence. Construct definitions determine what should be observed; sampling determines whose observations can support a population statement; instruments and administration shape response meaning; data preparation controls traceability; and analysis translates evidence into calibrated claims. The proposed protocol makes these dependencies visible and reinforces a simple priority: establish the quality and relevance of measurement before increasing the complexity of mathematics.

References

DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). SAGE.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method (4th ed.). Wiley.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. Social Psychological and Personality Science, 8(4), 370-378. https://doi.org/10.1177/1948550617693063

Lohr, S. L. (2021). Sampling: Design and analysis (3rd ed.). Chapman and Hall/CRC.

Suganob, I. D. (2024). Practical Research 2 made easy: A simplified guide on quantitative research in education. Shirley Publishing House.