CHAPTER 4

BUSINESS RESEARCH, STATISTICS, AND EVIDENCE-BASED DECISION-MAKING

Sixbert SANGWA
African Leadership University, Kigali, Rwanda | ssangwa@alueducation.com

PUBLICATION DETAILS

PARENT BOOK
The Business Enterprise: Core Functions, Integrated Management, and Emerging Transformations
https://doi.org/10.65655/ocp.m30

EDITOR AND PUBLISHER
Prof. Dr. Placide MUTABAZI
Open Christian Press, A publishing imprint of the Open Christian University and the Center for Faith and Work

CHAPTER DOI
https://doi.org/10.65655/ocp.m30.c113

COPYRIGHT AND LICENCE
© 2026 Sixbert SANGWA · CC BY 4.0

ABSTRACT

Business decisions routinely require action before uncertainty can be eliminated. The relevant scholarly problem is therefore not merely how to collect more data, but how to decide which evidence can support which inference, under what assumptions, and at what cost. This chapter develops a critical, practice-oriented account of business research, statistics, and evidence-based decision-making for enterprises operating across formal and informal institutions, with particular attention to Rwanda and comparable African economies. It integrates evidence-based management, research philosophy, measurement theory, statistical inference, causal identification, forecasting, decision theory, and responsible analytics. The secondary-source design combines seminal research, professional reporting guidance, official institutional methodology, and the National Institute of Statistics of Rwanda's 2024 Integrated Business Enterprise Survey. A reproducible analysis of published enterprise indicators illustrates percentage-point contrasts, denominator discipline, and the limits of descriptive comparisons. Constructed datasets provide independently checkable examples of dispersion, interval estimation, regression reasoning, and decision trees; constructed values are not passed off as field observations. An original evidence-to-decision workflow and provenance protocol show how questions, data, uncertainty, stakeholder interests, implementation, and evaluation must connect. The chapter argues that defensible enterprise intelligence is a governed inferential process rather than a dashboard, algorithm, or significant p-value. Its educational conclusion is that evidence quality depends jointly on relevance, measurement, identification, interpretability, and ethical consequences, while managerial responsibility remains irreducible to statistical calculation.

Keywords: business research; evidence-based management; secondary data; research design; measurement validity; descriptive statistics; statistical inference; causal inference; decision analysis; Rwanda.

LEARNING OUTCOMES AND CONCEPTUAL ROADMAP

On successful completion, the reader should be able to: (LO1) translate an enterprise decision into an answerable research question and explicit estimand; (LO2) evaluate descriptive, explanatory, exploratory, experimental, observational, and mixed-method designs for their inference limits; (LO3) identify, document, and critically appraise official, administrative, corporate, and scholarly secondary sources;

(LO4) calculate and interpret essential descriptive statistics, uncertainty intervals, and decision-analysis measures; (LO5) distinguish association, prediction, and causal explanation while diagnosing confounding and selection; (LO6) reproduce and qualify a Rwanda-based secondary-data analysis against international methodological comparators; (LO7) design a proportionate, ethically responsible analytics and reporting workflow; and (LO8) defend a cross-functional managerial recommendation with uncertainty, sensitivity, and accountability explicitly disclosed.

The argument progresses from the logic of evidence and research philosophy (Sections 4.1–4.2) to data architectures and methodological literacy (4.3–4.4), measurement and descriptive analysis (4.5–4.6), inference and causation (4.7–4.8), forecasting and choice (4.9), research synthesis (4.10), responsible digital analytics (4.11), and scholarly communication (4.12). Sections 4.13–4.15 integrate the Rwandan demonstration, functional dependencies, emerging research questions, and the chapter's teaching resources. Chapter 4 operationalizes the enterprise-system perspective of Chapter 1, the economic distinctions of Chapter 2, and the institutional constraints of Chapter 3. Detailed functional applications remain with the specialized chapters that follow.

4.0 INTRODUCTION: WHEN A BUSINESS NUMBER BECOMES A CLAIM

Consider a hypothetical Rwandan distributor whose customer dashboard reports that delivery complaints decreased by 25% after managers introduced a route-planning application. Operations recommends immediate expansion, finance anticipates savings, marketing advertises improved reliability, and personnel managers attribute the change to the software. Yet what was counted as a complaint? Were the same customers served before and after adoption? Did volumes decline, weather improve, or the reporting channel change? A reduction from twelve to nine complaints is indeed 25%, but the numerator alone reveals neither the exposure per shipment nor the counterfactual outcome without the software. The difference between a number, a defensible empirical description, and a causal decision is the problem of this chapter.

Research is disciplined inquiry intended to reduce specified uncertainty through transparent concepts, evidence, procedures, and reasoning. Business statistics comprise methods for describing variation and making conditional inferences from data. Decision analysis connects beliefs about uncertain outcomes with feasible alternatives, objectives, risk tolerance, and consequences. Evidence-based management requires managers to critically use available research, local organizational information, professional expertise, and the concerns of affected stakeholders rather than treating either intuition or quantitative output as infallible (Rousseau, 2006; Pfeffer & Sutton, 2006). Each discipline illuminates a different stage of practical judgment; none independently supplies a moral purpose or guarantees successful action.

A first distinction concerns types of claims. Descriptive claims state what was observed in a defined population and period; diagnostic claims identify patterns or plausible explanations; predictive claims forecast outcomes under a model; causal claims estimate what would change if an intervention were implemented; and normative claims defend what ought to be done. The five cannot be collapsed into an ascending ladder of certainty. A precise forecast may be poor evidence of a policy's effect, and a causal effect established in a narrow experiment may be irrelevant to a different firm's constraints. Managerial decisions require an explicit bridge from the desired claim to its evidentiary design (Rubin, 1974; Hernán & Robins, 2020).

4.0.1 Scope, documentary method, and evidentiary limits

This chapter is a critical, purposive synthesis of secondary scholarship and official documentary sources, not a registered systematic review, new survey, interview study, field experiment, or original causal evaluation. Foundational texts were selected for distinctive theoretical mechanisms: evidence-based management, measurement validity, sampling, inference, prediction, causality, and decisions under risk. Targeted source checks conducted for this manuscript on 24 October 2026 covered the journal and publisher records for major methodological sources; NISR's official Integrated Business Enterprise Survey 2024 publication and its 2023 predecessor; World Bank Enterprise Surveys methodology; the PRISMA reporting framework; the American Statistical Association's p-value statement; NIST's AI risk-management guidance; and the OECD's AI recommendation. The searches used identifiable author-title combinations and terms such as “Integrated Business Enterprise Survey 2024 Rwanda,” “World Bank Enterprise Surveys methodology,” “ASA p-value statement,” and “AI Risk Management Framework 1.0.” They did not constitute exhaustive searches across every bibliographic database, and no fabricated screening totals are presented.

The empirical Rwanda illustration uses percentages quoted on the NISR source page, with displayed arithmetic and a source date. It does not reproduce survey-weighted microdata estimates, sampling errors, or causal effects. Numerical delivery, estimation, and investment exercises are explicitly hypothetical, with all input assumptions visible and computations independently checked. Citations to conceptual literature support the claims for which that literature is relevant; official statistical statements retain their defined survey populations. Legal and ethical obligations are treated as verification requirements rather than unqualified claims about all countries. The evidentiary cut-off is 24 October 2026, and later dataset revisions may change reported values.

The methodological orientation is fallibilist and decision-sensitive. Models are deliberately simplified representations whose adequacy must be assessed relative to an intended use, not evaluated by the number of parameters they contain. Box (1976) described scientific inquiry as a recurring confrontation between theoretical representation and observation. Applied to business research, that orientation rejects two equally misleading positions: treating quantitative outputs as self-validating facts, and treating all evidence claims as so situated that critical comparison is impossible. There is room for both context-specific interpretation and disciplined empirical criticism, provided that assumptions, inference targets, and alternative explanations remain inspectable.

4.0.2 Research as an enterprise-wide capability

An enterprise repeatedly faces questions that belong to different disciplines but share a common evidentiary architecture. Marketing asks which customers value an offering; operations asks which bottleneck causes late delivery; finance asks whether a project compensates risk; human resources asks whether training improves performance; governance asks whether decisions are lawful and fairly documented. These questions differ in unit, outcome, intervention, and acceptable error. Consequently, a single undifferentiated “business intelligence” database cannot replace conceptual specification or professional judgment. Research competence is both a specialized capability and a cross-functional discipline of asking what the numbers mean.

The chapter's organizing proposition is conditional: better decision performance is more plausible when a decision problem is clearly specified, sources are fit for purpose, inferential claims stay within their designs, feasible alternatives are compared under uncertainty, and consequences are reviewed. This is a reasoned conceptual synthesis rather than an empirically validated universal law. In small enterprises,

sophisticated algorithms may be less valuable than accurate sales records and honest stock counts. In large platforms, abundant transactional data may intensify problems of selection, privacy, and feedback loops. Appropriate sophistication is therefore determined by the decision, the data-generating process, the cost of error, and accountability, not the size of the dataset.

4.1 THE LOGIC OF EVIDENCE-BASED MANAGEMENT

4.1.1 From problem recognition to an answerable question

The first analytical task is to distinguish a symptom from a decision. “Sales are weak” describes a perceived condition; “Would extending weekday opening hours increase monthly contribution after labor costs among the firm's current customers?” identifies an intervention, outcome, population, and horizon. A good research question specifies a focal unit (transaction, worker, firm, branch, district), a construct, the relevant comparison, and the decision threshold. Where a causal effect is sought, the question should name the treatment and plausible counterfactual; where a description is sought, it should define the denominator and reference interval. Poorly formulated problems encourage data fishing: the analyst computes whatever is easiest to measure and retrofits a managerial story afterwards.

Research objectives follow the decision rather than merely its vocabulary. A descriptive inquiry may estimate the share of deliveries meeting a promised date, an explanatory inquiry may examine why performance differs by route, and an evaluation may ask whether a proposed scheduling intervention improves on current practice. In a formal causal formulation, the estimand is the precisely defined quantity to be learned, such as the average difference in on-time delivery probability among eligible shipments under two feasible policies. If the estimand cannot be described in ordinary language, a technical model will not rescue the decision from ambiguity (Hernán & Robins, 2020).

4.1.2 Theoretical models, constructs, mechanisms, and assumptions

Theory supplies concepts and conjectured mechanisms, not a decorative literature review. An enterprise may hypothesize that better training increases task knowledge, which reduces errors, which raises customer retention. Each arrow is a proposition requiring a time order, a measurement strategy, and a test against alternative causes, such as changes in customer mix or technology. A conceptual model is useful only if it exposes these potentially contestable relationships. It should distinguish measured variables from latent constructs such as trust, capability, fairness, or satisfaction and identify where an observed indicator is an incomplete proxy (Cronbach & Meehl, 1955).

Theories are rival explanatory tools. Economic rational-choice models emphasize incentives and constraints; bounded rationality emphasizes limited attention and search; institutional approaches emphasize rules and norms; organizational learning examines the exploration–exploitation tension; and behavioral decision research demonstrates systematic departures from expected-utility descriptions (Simon, 1955; Kahneman & Tversky, 1979; March, 1991). No one perspective should be presumed correct for every decision. A data-driven explanation must ask which mechanisms each theory predicts, what observations could discriminate among them, and whether unmeasured institutional differences make a transfer implausible.

The distinction between evidence and a claim about evidence can be made operational through an inferential contract. Before examining an outcome, researchers should specify which decision will be made, what change would count as materially beneficial, which counterfactual or comparator is credible, what sources may disconfirm the preferred explanation, and which assumptions cannot be checked directly.

This discipline is more demanding than attaching a confidence level to a completed spreadsheet. It makes visible the point where observation ends and managerial interpretation begins. In evidence-based management, the research literature, internal organizational facts, professional judgment, and stakeholder concerns are complementary inputs whose evidentiary authority depends on the question; none deserves automatic priority across all decisions (Briner et al., 2009; Rousseau, 2006).

4.1.3 Evidence, expertise, interests, and judgment

Rousseau (2006) characterizes evidence-based management as a response to the distance between credible organizational research and ordinary professional practice. Its appeal is practical: challenging customary management fashions with evidence while recognizing that research is incomplete, context-dependent, and interpreted by fallible people. Expertise remains essential for deciding which outcomes matter, whether an intervention is operationally possible, and whether historical data reflect the intended future process. Stakeholder testimony matters because a numerical improvement in unit cost may conceal unpaid labor, accessibility barriers, or unsafe service conditions. Nonetheless, expertise is not permission to ignore disconfirming evidence, and stakeholder preferences do not alone demonstrate empirical causation.

Decision makers should distinguish evidence strength from action urgency. A high-consequence irreversible investment requires deeper independent validation and sensitivity analysis than a reversible change to an internal report. Yet waiting for definitive proof may also be costly when a firm faces insolvency, unsafe products, or a rapidly expiring opportunity. Evidence-based judgment thus means matching analytical rigor to the stakes while recording assumptions and a plan to learn. It does not mean eliminating discretion or announcing false certainty (Pfeffer & Sutton, 2006).

A decision criterion should be stated separately from the research finding. An intervention might increase expected financial contribution but be infeasible because of working-capital constraints, an inadequate safety margin, or the rights of employees affected by its introduction. This is not an objection to quantitative reasoning. It is an insistence that its objective function and constraints be declared rather than smuggled into the word “optimal.” Where evidence is incomplete, a reversible pilot, a staged investment, or a policy with explicit review points may dominate both premature full implementation and indefinite deferral. The cost of information, the cost of delay, and the distribution of possible losses are therefore elements of evidence-based judgment, not subsequent administrative details (Simon, 1955).

4.2 RESEARCH PARADIGMS AND DESIGNS

4.2.1 Ontology, epistemology, and the inference problem

Research paradigms make assumptions about what exists, what can be known, and what counts as a warranted explanation. Postpositivist research generally treats an external world as investigable through fallible measurements and revisable hypotheses; interpretivist inquiry emphasizes socially constituted meanings and context; critical realist reasoning distinguishes observed events from deeper mechanisms that may not be directly visible; and pragmatic inquiry gives priority to the appropriateness of methods for a clearly posed problem. These traditions ask different but not mutually exclusive questions. A customer's invoice amount may be externally recorded, while perceptions of trust or unfair treatment require interpretation in social context. Philosophical pluralism is productive where competing claims are made explicit; it is unhelpful when “mixed methods” becomes a label for unrelated data collections with no integrated inference.

Explanations should match the phenomenon. A cross-sectional survey may estimate prevalence and identify associations but cannot ordinarily establish temporal sequence or rule out reverse causation. Longitudinal records improve temporal ordering yet remain susceptible to time-varying confounding. Ethnographic or interview-based evidence can reveal mechanisms, meanings, and overlooked categories, although results cannot automatically be treated as population frequencies. Experiments can strengthen internal validity under appropriate random assignment, compliance, and measurement, but their effects may be local to the trial's participants and procedures (Shadish et al., 2002; Rubin, 1974).

4.2.2 Descriptive, exploratory, explanatory, and evaluative purposes

A descriptive design characterizes a defined situation: for example, the distribution of business size in a survey's sampling frame. An exploratory design investigates a poorly understood phenomenon and may refine constructs or generate hypotheses. An explanatory design tests competing accounts of why outcomes vary. Evaluation compares an intervention with credible counterfactuals or implementation standards. A mixed-methods design intentionally connects quantitative and qualitative strands at the design, analysis, or interpretation stages; adding quotations to a numeric table does not by itself create integrative evidence (Creswell & Plano Clark, 2018).

Research designs are not interchangeable merely because they use similar techniques. Case studies can provide detailed process evidence and analytic generalization when their boundaries, sources, and competing explanations are specified; they do not supply unbiased prevalence estimates without a sampling strategy. Cross-sectional comparisons can generate hypotheses but are especially vulnerable to cohort and selection effects. Administrative data may measure an entire registered universe while systematically missing unregistered enterprises. Randomized experiments reduce some forms of confounding on average but cannot repair invalid measures, attrition, spillovers, or unethical treatment assignments.

There is a deeper philosophical disagreement over whether all legitimate inquiry can be reduced to measurement of regularities. Postpositivist approaches seek disciplined tests of revisable propositions; interpretive research can investigate the meanings through which participants understand a contract, workplace norm, or institutional practice; critical realism asks whether an observed regularity is generated by underlying mechanisms that operate differently in open systems. Pragmatism is useful where a decision needs several kinds of information, but it cannot dispense with ontological and epistemic consistency. To combine a survey of reported trust with interview accounts of unfair treatment, for example, the researcher must explain whether both sources represent the same construct, different dimensions of a phenomenon, or potentially conflicting interpretations. Integration requires an explicit account of what each method can warrant, not merely the coexistence of numbers and narrative (King et al., 1994; Creswell & Plano Clark, 2018).

4.2.3 Experimental and quasi-experimental identification

Random assignment seeks to make treatment and control groups comparable in expectation before intervention, permitting a design-based causal contrast under assumptions including consistency, adequate compliance, no problematic interference, and valid outcome capture. A trial of an inventory reminder system might randomize eligible shops, but spillovers could occur if neighboring shops share suppliers and information. The causal target would need to specify whether the intervention acts on the shop, its network, or individual employees. A poorly implemented experiment may be less informative than a carefully documented observational design with a credible identifying shock (Shadish et al., 2002).

Where randomization is unavailable, researchers may consider interrupted time series, difference-in-differences, regression discontinuity, instrumental variables, or matched comparisons. Each relies on different identifying assumptions. For example, a difference-in-differences interpretation requires appropriate trends in the untreated potential outcomes, not merely two measurements before and after a policy. Regression discontinuity estimates are commonly local to a threshold, while a valid instrumental variable must alter treatment without independently affecting the outcome through excluded pathways. Techniques should be selected because their assumptions are plausible and scrutinizable, not because their names sound sophisticated (Hernán & Robins, 2020).

A useful safeguard is to describe an ideal target trial before evaluating an observational dataset. For a proposed business training programme, the protocol would specify eligible firms, time zero, alternative treatment strategies, assignment and follow-up, outcome definitions, estimand, and analysis plan. Observational records can then be assessed for whether their timing and comparison groups approximate those features. This approach draws attention to immortal-time, self-selection, anticipatory adoption, and differential follow-up biases that are easily hidden when analysts begin from available columns rather than a decision question. It does not make nonexperimental data randomized; it supplies a disciplined benchmark for explaining where their inferential credibility begins and ends (Hernán & Robins, 2016).

4.2.4 The design-to-claim discipline

The defensibility of evidence depends on four separate tests: construct validity asks whether the measure captures the intended idea; internal validity asks whether an observed contrast supports the proposed causal explanation; external validity asks to whom and when it transfers; and statistical conclusion validity asks whether numerical relationships have been estimated and interpreted appropriately. These tests interact but cannot substitute for one another. A large, precisely estimated difference can be invalid if it compares unlike units; an internally credible experiment can be irrelevant to informal firms excluded by its recruitment criteria (Shadish et al., 2002).

A further distinction separates explanatory and predictive models. Prediction asks how well a model estimates as-yet-unseen outcomes; explanation asks what mechanism and relationship generated them. A variable may substantially improve forecasts without being a manipulable cause, while a causal treatment may have limited value in predicting which customer defaults next month. Breiman (2001) drew attention to tensions between statistical data models and algorithmic learning; responsible business inquiry should use out-of-sample prediction tests for forecasting and causal identification arguments for claims about interventions, without conflating the two traditions.

4.3 SECONDARY DATA ARCHITECTURE

4.3.1 What constitutes secondary evidence?

Secondary data are information collected or compiled by parties other than the analyst for another purpose, or previously generated data reused for the current inquiry. Examples include official population and enterprise statistics, regulatory disclosures, audited financial statements, tax or licensing administrative records made accessible lawfully, company transaction histories, satellite products, trade datasets, and research publications. “Secondary” does not mean inferior: an authoritative national statistical agency may implement a more rigorous sampling design than a small ad hoc manager survey. But the data's original purpose limits what can be inferred. A tax record captures administratively recognized transactions, not necessarily all economic activity; a platform log describes actions visible to that platform, not the full consumer market.

The main provenance questions are who originated the data, for which period, under what definitions and collection procedures, with which exclusions or weights, through what revision history, and subject to what permissions. Dataset metadata should report the unit of observation and aggregation, variables and codes, missing-value conventions, source version, retrieval date, geography, instrument, and licensing terms. A public PDF and a downloadable workbook may appear to show the same statistics but differ because of revisions or denominator selection. It is methodologically inappropriate to average incompatible figures merely to create apparent agreement.

4.3.2 Major source families and their strengths

Official household and enterprise surveys can estimate population characteristics when frames, design weights, and field protocols are respected. Administrative registers may provide high-frequency or near-census records for covered activities but may reflect enforcement incentives, registration gaps, and definitional changes. Corporate reports can describe management's reported results but require scrutiny of accounting policies, incentives, audit scope, and unaudited metrics. International harmonized databases facilitate comparison but can conceal national methodological differences, lagged revisions, or modelled observations. Peer-reviewed meta-analyses can synthesize related studies, yet their conclusions depend on the quality, comparability, and publication practices of those studies.

The World Bank Enterprise Surveys illustrate the importance of survey universes. Its standard formal-sector survey targets eligible registered firms with five or more workers in specified private sectors and uses stratification by size, sector, and region. Those data should not be presented as the experiences of every Rwandan enterprise, particularly where informal and very small activities constitute much of the business landscape (World Bank, n.d.). In contrast, NISR's Integrated Business Enterprise Survey covers both formal and informal non-agricultural enterprises under its stated national definitions (National Institute of Statistics of Rwanda [NISR], 2026). The surveys answer related but nonidentical questions and their estimates should not be spliced as if the populations were interchangeable.

Secondary sources also operate at different levels of aggregation. Financial accounts record the reporting entity under accounting rules; a labor-force survey samples people or households; an establishment census locates operating sites; an enterprise survey may consolidate several sites under a common owner. Attempting to join these records without a stable crosswalk can create false duplicates or attribute district characteristics to individual firms. Ecological inference is especially hazardous: a district with high mobile-money penetration does not establish that the particular exporting firms located there use digital finance. Conversely, national averages can conceal large operational differences among firms. A rigorous source appraisal identifies the unit for every variable, the target population for every inference, and the institutional event that generated the record before aggregating or linking it.

4.3.3 Provenance, licensing, and governance of reuse

Findability and accessibility are not equivalent to unrestricted reuse. The FAIR principles promote data that are findable, accessible, interoperable, and reusable, while acknowledging that access may be controlled where privacy or legitimate restrictions apply (Wilkinson et al., 2016). A professionally governed secondary-data study records source identifiers and acquisition procedures, obtains authorization where necessary, restricts sensitive fields, and publishes only outputs permitted by source terms. Aggregated public statistics often permit citation and reproduction with attribution; detailed administrative or microdata extracts may require application, confidentiality agreements, secure environments, or disclosure review.

A minimal provenance register should preserve the indicator name and code, statistical population, period, publication/version, retrieval date, URL, transformation formula, analysis file checksum where available, and permitted uses. Source errors should not be silently corrected in the working dataset; transformations belong in a reproducible script or explicitly documented spreadsheet with raw inputs retained. If values are revised by the publisher, the analyst should preserve the original version used for the decision and state whether the conclusion changes under the revision. These practices connect research integrity with legal compliance and future auditability.

In practice, a data-provenance record should distinguish the source file from the analytical dataset, and both from a published indicator. If raw records R are subjected to exclusions E, transformations T, and an aggregation A, the reported indicator may be represented as $I = A[T(E(R))]$. Each operation needs its own versioned specification and unit test. A later correction to a source table should not silently overwrite the values on which an earlier managerial decision depended; the investigator should produce a change comparison and determine whether the original recommendation remains robust. These safeguards are particularly important when donor, administrative, or platform data can be revised without a clearly numbered public release. Findability and responsible reuse are necessary conditions for auditability, but neither establishes validity (Wilkinson et al., 2016).

4.4 PRIMARY RESEARCH AS METHODOLOGICAL LITERACY

4.4.1 Surveys, interviews, observation, and field experimentation

Although the present chapter conducts no participant research, professionals must understand how primary evidence is generated. Surveys can measure reported attitudes, transactions, or practices across appropriately sampled units, but question wording, response modes, nonresponse, and social desirability can distort findings. Interviews offer flexibility for exploring experiences and mechanisms; they require intentional sampling, reflexive interpretation, accurate records, and care not to generalize a few articulate respondents into population prevalence. Observation can disclose the difference between formal process manuals and actual workflows, but the observer's presence and recording choices may change behavior. Field experiments test feasible interventions and can estimate effects under explicit assignment and exposure conditions (Cochran, 1977; Shadish et al., 2002).

Questionnaires require pretesting for comprehension, order effects, language equivalence, and relevant response categories. In multilingual African settings, direct translation does not guarantee construct equivalence: a phrase describing job satisfaction, informal credit, or household enterprise may carry different social meanings. Sampling should begin with a defined population and frame; convenience samples may be useful for exploratory diagnosis but cannot support claims of statistically representative prevalence without strong additional assumptions. A credible consent process explains purpose, risks, storage, voluntariness, and the consequences of declining, especially where employees may reasonably fear supervisor retaliation.

4.4.2 Ethics, risk, and limits of consent

Participation requires appropriate ethical review where applicable, but formal consent alone does not guarantee fairness. Employees may be pressured to participate, customers may not understand secondary uses of transaction records, and experiments can shift costs to people outside the analytic sample. Research plans should assess necessity, proportionality, de-identification limits, retention, security, and routes for participant redress. Applicable Rwandan data-protection and institutional requirements must be checked against current authoritative texts before real data collection or cross-border transfer, as Chapter

3 emphasizes. Researchers must not invent institutional approval, participant quotations, interview counts, or experimental results.

In enterprises, internal "A/B testing" is often described as if it were methodologically or ethically exemption-worthy. Yet a randomized product display may affect access, pricing, or protected groups, and a staff-management experiment may influence wages, distress, or disciplinary decisions. A risk-based protocol should ask whether consent or another lawful basis exists, whether an alternative design would impose fewer burdens, whether an independent reviewer is needed, and whether statistical gains justify the possible harm. The fact that software makes testing cheap does not establish that all tests are legitimate.

There is a corresponding methodological reason to treat consent, confidentiality, and respectful recruitment as part of design quality. Coerced staff may overreport satisfaction, dissatisfied customers may avoid participation, and workers who fear surveillance may alter the behavior under observation. These responses contaminate the very mechanisms the research seeks to measure. Ethical and statistical validity are not identical, but poorly governed research can damage both. A protocol for sensitive business research should specify who may invite participants, who may access identifiable information, how refusals are protected from workplace consequences, and how findings will be aggregated so that individuals or small groups cannot be indirectly identified. Where lawful grounds for reuse differ from the grounds for new collection, the analyst must evaluate both rather than presuming that internal company possession is unrestricted research permission.

4.5 DATA QUALITY AND MEASUREMENT

4.5.1 Constructs, operational definitions, and measurement scales

A variable is an observed or derived attribute that can take different values; a construct is an analytically defined property, sometimes not directly observable. "Customer loyalty" might refer to repeat purchase, stated intention, relationship duration, or resistance to alternatives. These are not interchangeable operational definitions. Repeat purchasing may reflect satisfaction, contractual lock-in, geographical monopoly, or temporary price. Cronbach and Meehl (1955) located construct validation within a network of theoretical relations, while Campbell and Fiske (1959) emphasized convergence across methods and discrimination between distinct traits. For business researchers, validity is an argument supported by content, measurement process, and patterns of evidence, not a single coefficient conferred by a software menu.

Nominal measures classify cases, such as business sector or payment channel; ordinal measures rank categories without guaranteeing equal distances; interval measures express meaningful differences on a scale without a true ratio zero; and ratio measures such as shipment duration or revenue permit meaningful proportional comparison when defined consistently. An ordinal five-point agreement item is not automatically an interval measure merely because spreadsheet software calculates its mean. Multiple-item indices may be informative when the underlying dimensionality and interpretation are defended, but a high internal consistency coefficient alone cannot demonstrate unidimensionality, cultural equivalence, or construct validity. Indicators are useful to the extent that their measurement rules remain coherent with the decision problem.

4.5.2 Reliability, accuracy, completeness, and comparability

Reliability refers to measurement consistency under specified conditions; accuracy concerns closeness to a relevant reference; validity concerns whether an intended interpretation is warranted. A delivery timestamp captured automatically by a device may be repeatable yet wrong because the clock is misconfigured. A customer satisfaction scale can be internally consistent yet fail to represent customers excluded from digital channels. Data quality must also consider completeness, timeliness, uniqueness, logical consistency, linkage accuracy, and fitness for a particular use. These dimensions should be separately assessed rather than summed into a spurious universal score.

Measurement error may be classical, systematic, differential, or related to strategic behavior. A firm's sales ledger may omit returns, taxes, or deferred transactions; its employee headcount may count part-time workers as full-time equivalents; its time-to-delivery may stop the clock when dispatch occurs rather than when the buyer receives the product. Definitions influence decisions. If a new reporting system improves complaint detection, reported complaints may increase even as service improves. Managers should investigate how the observation process changes alongside the phenomenon, particularly when indicators become targets for rewards, contracts, or compliance.

Measurement invariance becomes crucial in cross-country and multilingual comparisons. A survey item can be translated accurately at the level of words yet function differently if respondents interpret response categories, managerial authority, or employment status according to different institutional norms. Comparisons of scale means are then potentially ambiguous: observed differences may reflect the latent construct, response styles, item functioning, or population composition. Researchers should test conceptual and linguistic equivalence, assess factor structure where relevant, and explain when strict metric comparability is not defensible. A comparatively large sample cannot compensate for an unexamined change in what the instrument measures (Campbell & Fiske, 1959; Cronbach & Meehl, 1955).

4.5.3 Missingness, selection, and representativeness

Missing data are not a purely technical inconvenience. Under missing completely at random (MCAR), missingness is unrelated to observed or unobserved data; under missing at random (MAR), the missingness mechanism can depend on observed information given in the model; and under missing not at random (MNAR), it may depend on unobserved values even after conditioning. These are model assumptions, not diagnoses that can be proven from a blank cell. Listwise deletion can waste information and introduce bias; simple mean imputation understates uncertainty. Multiple imputation and likelihood-based methods may be justified under appropriate assumptions, whereas MNAR concerns call for sensitivity analysis and sometimes new data collection (Little & Rubin, 2019).

Selection bias arises where inclusion in the observed dataset is related to variables influencing the estimated outcome. If a retailer analyzes only customers who have already completed digital registration, its findings may exclude older, poorer, or less connected customers. Survivorship bias arises when failed enterprises disappear from datasets and surviving firms appear more productive. Nonresponse bias persists even with a large numerical sample if nonrespondents differ systematically on relevant variables. Representativeness therefore depends on the population, coverage, design, response process, weighting, and identification of unobserved groups, not simply on sample size (Cochran, 1977).

A missingness audit should be decision-specific rather than a generic percentage threshold. If 2% of invoice values are missing, the concern is not necessarily small: the missing entries may be concentrated among the largest transactions or a newly opened branch. Under MAR, imputation models must incorporate variables plausibly related to both missingness and the substantive outcome; imputed values

and uncertainty cannot be treated as if directly observed. Under MNAR, transparent sensitivity analyses can vary assumptions about unobserved records. Researchers should also distinguish weighting intended to restore a sampling frame from weighting intended to address nonresponse. Neither is a magic correction when the selection process depends on unmeasured characteristics (Little & Rubin, 2019).

4.5.4 Practical data-quality audit

A defensible audit begins by naming the expected unit and unique identifier, verifying allowable ranges and currency units, identifying duplicates, measuring missingness by meaningful subgroup, checking chronology, and comparing key aggregates with independent control totals. Analysts should preserve raw files, script every substantive transformation, and investigate rather than silently remove improbable observations. Outliers may be data-entry errors, rare genuine transactions, fraud indicators, or the most important source of insight. Distributional visualizations and domain checks should precede deletion or winsorization. A change log should record every rule and its effect on the number of records and outcome estimates.

Table 4.1 | Data quality risk

Data-quality domain	Diagnostic question	Minimum documentary response	Consequence if ignored
Coverage	Which enterprises or customers can never enter the dataset?	Target population, frame, exclusion inventory	Nonrepresentative estimates
Measurement	Does the recorded field capture the intended construct?	Operational definition, validation evidence	Apparently precise but invalid claims
Missingness	Who has blank values and why?	Missingness rates by subgroup and sensitivity plan	Biased group comparisons
Linkage and duplication	Do keys represent unique and correct units?	Join specification, duplicate checks	Inflated volumes or false correlations
Revision and timing	Which vintage, period, and transformation apply?	Source release, date, version, code	Spurious change over time
Permissions and security	May these fields be reused and disclosed?	Licence, lawful basis, role permissions	Harm, noncompliance, loss of trust

Table 4.1 provides a structured audit of source and analysis risks and responses informed by measurement theory, survey methodology, and the FAIR data principles (Cochran, 1977; Cronbach & Meehl, 1955; Wilkinson et al., 2016). It is a proposed diagnostic instrument, not a validated index of data quality.

4.6 DESCRIPTIVE STATISTICS AND VISUALIZATION

4.6.1 Distributions, denominators, and rates

Descriptive statistics summarize observed values under explicit scope. Frequency counts identify numbers of cases; shares divide a count by a clearly defined total; rates divide events by a meaningful exposure such as shipments, worker-hours, or customer-months. A percentage-point difference is the arithmetic subtraction of two percentages, whereas a percentage change expresses the relative change with respect to a chosen baseline. For instance, a movement from 20% to 25% is five percentage points and a 25% relative

increase; calling both “a 5% increase” is incorrect. A business report should state the unit, population, time, source, and denominator alongside each headline percentage.

Central tendency describes typical values but is sensitive to distribution shape. The arithmetic mean is the total divided by the number of observations; the median is the central ordered value; and the mode is the most frequent value. The mean suits some additive decisions such as expected resource use, but skewed delivery times or incomes may be communicated more fairly by a median and upper quantiles. A subgroup comparison should distinguish changes in underlying performance from changes in mix. If a retailer acquires more rural customers, its overall delivery average may rise even if routes improve within each location. Such aggregation effects are a warning against uncritical dashboard interpretation.

4.6.2 Dispersion and a worked delivery-time example

Consider eight *illustrative, invented* delivery times, in days: 2, 3, 3, 4, 5, 7, 8, and 16. Their sum is 48 and the arithmetic mean is $\bar{x} = 48/8 = 6.0$ days. The middle ordered observations are 4 and 5, giving a median of 4.5 days; the mode is 3 days. Relative to the mean, the squared deviations sum to 144; the sample variance is $s^2 = 144/(8 - 1) = 20.5714$ days^2, and the sample standard deviation is $s \approx 4.5356$ days. The population variance for this complete eight-record illustrative set is $144/8 = 18$ days^2, with population standard deviation $\sqrt{18} \approx 4.2426$ days. Which denominator is appropriate depends on whether these eight times constitute the entire population of interest or a sample used to estimate a larger process.

The mean of 6.0 days is pulled upward by a single 16-day delivery. Removing that observation for a sensitivity illustration, not for a claim that it is erroneous, produces a seven-record mean of $32/7 \approx 4.5714$ days. A responsible analyst would investigate the exceptional delivery before excluding it and report the tail risk relevant to service guarantees. If the enterprise promises delivery within five days, five of the eight deliveries meet the threshold, so observed on-time performance is $5/8 = 62.5\%$. That fraction, too, is illustrative and would be unstable as a general estimate from only eight cases.

The distinction between mean performance and tail exposure can be shown without another statistical model. Suppose a service contract imposes a penalty for any delivery exceeding seven days. In the illustrative eight deliveries, two observations, eight and sixteen days, exceed that threshold; the observed penalty exposure is therefore $2/8 = 25\%$, despite a mean below seven days. If the firm changes its objective from minimizing average transit time to limiting the 90th percentile or late-delivery probability, a different operational intervention may be warranted. Such proportions from eight invented observations are not reliable population estimates; their purpose is to demonstrate that the choice of loss function and service standard must precede the choice of summary statistic.

4.6.3 Quantiles, indexes, and normalized comparison

Quantiles locate observations within an ordered distribution. The 90th percentile of waiting time may be more operationally meaningful than its mean where extreme delays create customer harm. Index numbers compare changes relative to a reference, usually normalized to 100, but their interpretation depends on the included items, weights, base year, and treatment of changing composition. In constructing productivity or price indexes, an analyst should document whether output quality, labor hours, and deflators have been adjusted. A nominal revenue increase need not indicate a higher volume sold or greater real value added, as Chapter 2 explains.

For comparisons across differently sized enterprises, normalization can improve fairness but can also conceal context. Customer complaints per 1,000 completed orders is usually more interpretable than raw complaints; accidents per worker-hour may be more meaningful than accidents per worker in sectors with

different hours. However, a denominator can itself be manipulated. A company that stops serving complex customers might improve its complaint rate without improving its service to those excluded. Effective indicator design requires an outcome definition, reference population, and a safeguard against strategic exclusion.

4.6.4 Charts as analytical arguments

Tables permit precise comparisons; graphics expose shape, trend, co-movement, and uncertainty. Histograms display distributions, box plots summarize spread and unusual values, line charts show ordered changes over time, and scatterplots help reveal nonlinear relationships and distinct clusters. The graphical choice should follow the variable and substantive question. Axes must have units, comparisons must use honest scales, and any uncertainty band should be identified by its mathematical meaning. Truncated axes can exaggerate small differences in bars; dual axes can imply a relationship determined by arbitrary scale choices; excessive colors can conceal the groups a reader needs to distinguish. Anscombe's (1973) constructed examples famously show how datasets with similar summary statistics can have markedly different graphical structures, making visual inspection part of the analytic method rather than decoration.

Responsible visualization also includes the people and activities omitted from a graph. A geographically aggregated chart of Rwanda's commercial activity should not imply that Kigali represents every district. A chart of registered large firms cannot establish the characteristics of informal microenterprises. Graphical accessibility requires readable labels, sufficient contrast, interpretations that do not rely solely on color, and, where electronic formats allow, descriptive alternative text. Every visual should be traceable to data, transformation, and source, with a note clearly separating official observations from hypothetical teaching quantities.

Simpson's paradox offers a further warning about visualization and aggregation. A company may improve completion rates within both urban and rural routes while its aggregate rate declines, if the share of difficult rural deliveries rises sharply. This is not a mathematical contradiction: a weighted mean changes when either subgroup performance or subgroup weights change. A sound dashboard displays both the overall rate and relevant strata, with weights and periods made explicit. In longitudinal graphics, changes in questionnaire wording, geographic coverage, or automated logging should be marked as measurement breaks rather than visually blended into a continuous trend. The best display makes plausible counterinterpretations easier to detect, not harder.

4.7 PROBABILITY AND STATISTICAL INFERENCE

4.7.1 Probability as a model of uncertainty

Probability assigns numerical structure to uncertainty. Under a frequentist interpretation, probabilities characterize repeatable event mechanisms or hypothetical repetitions; Bayesian inference expresses beliefs over hypotheses or unknown quantities conditional on a specified prior and likelihood. Neither tradition allows unsupported exactness. A probability of supplier disruption may be estimated from comparable historical contracts, expert elicitation, or a forecasting model, each with different uncertainties. Probability must also be separated from risk consequence: two disruptions with the same probability can have radically different effects on cash flow, safety, and institutional legitimacy (Kahneman & Tversky, 1979).

Conditional probability addresses how information changes a likelihood: $P(A|B) = P(A \cap B)/P(B)$, for $P(B) > 0$. Bayes' rule connects new evidence to prior odds: $P(H|E) = P(E|H)P(H)/P(E)$. Business managers should distinguish P(fraud indicator | fraud) from P(fraud | indicator). Where actual fraud is rare, a screening model with many false positives can incorrectly accuse a large number of legitimate customers despite high apparent sensitivity. Base rates and error costs therefore matter to both analytic accuracy and procedural fairness.

A numerical Bayesian screening example demonstrates why conditionals cannot be reversed. Suppose that, in an entirely hypothetical population, 1% of transactions are fraudulent, a screening rule detects 90% of fraud, and it incorrectly flags 5% of legitimate transactions. Of 10,000 transactions, 100 are fraudulent and 9,900 legitimate; 90 fraudulent transactions and 495 legitimate ones are flagged. Consequently $P(\text{fraud} | \text{flagged}) = 90/(90 + 495) \approx 15.38\%$, even though $P(\text{flagged} | \text{fraud}) = 90\%$. If all flags were treated as proof, most accused customers would be innocent under the assumed model. This calculation does not estimate fraud in Rwanda; it shows why prevalence, positive predictive value, review rights, and the costs of erroneous denial must be considered together. Bayesian updating remains conditional on the accuracy and transportability of the assumed base rate and test characteristics.

4.7.2 Sampling distributions and standard errors

When observations are obtained through an appropriate sampling design, an estimator's sampling distribution describes variation across hypothetical repetitions of that design. For independent and identically distributed observations with finite variance, the standard error of a sample mean can be estimated as $SE(\bar{x}) = s/\sqrt{n}$, subject to applicable assumptions. For complex surveys, simple random-sample formulas are typically inadequate: stratification, clustering, unequal inclusion probabilities, weighting, finite-population effects, and nonresponse adjustments may materially change uncertainty estimates (Cochran, 1977).

A confidence interval is produced by a procedure designed to cover a fixed population parameter at a specified long-run frequency under the model and sampling assumptions. A particular 95% frequentist interval does not mean the fixed parameter has a 95% probability of lying within it. A Bayesian credible interval has a different interpretation conditional on its prior and model. Neither interval automatically incorporates measurement bias, omitted variables, unobserved informality, or errors in the population frame. Managers must communicate what uncertainty has been quantified and what remains outside the calculation.

The distinction between sampling uncertainty and total epistemic uncertainty is foundational. Standard errors describe variability induced by a particular repeated sampling or modeling procedure, conditional on a sufficiently accurate sampling frame and statistical specification. They cannot capture an unobserved sector omitted from the frame, a consistently misclassified business type, or deliberate underreporting. For cluster-sampled enterprise data, treating employee-level observations from one employer as independent can produce artificially narrow intervals. Effective sample size depends on the number and structure of independent information units, not simply the count of rows in a database. This is why survey documentation, sampling weights, clustering, and an explicit uncertainty taxonomy are indispensable to substantive interpretation (Cochran, 1977).

4.7.3 Worked interval estimate under declared assumptions

Suppose a *hypothetical* simple random sample of $n = 100$ independently observed orders shows a mean fulfillment time of 4.8 days and sample standard deviation of 2.0 days. The estimated standard error is $2/\sqrt{100} = 0.20$ days. Using a normal critical value of 1.96 for pedagogical approximation, a 95% interval is

4.8 ± 1.96(0.20), namely 4.408 to 5.192 days. With a genuinely normally distributed underlying population and an estimated standard deviation, a t critical value with 99 degrees of freedom would be slightly larger; with serially correlated orders or cluster sampling the naive interval could be too narrow. The example teaches the mechanics, not the sampling properties of any actual firm.

A confidence interval is not a substitute for a decision threshold. If a delivery promise is five days, the illustrated interval crosses that benchmark, and the business may need further data or a conservative operational buffer. Even if an average were confidently under five, a heavy-tailed distribution could still leave an unacceptable proportion of customers substantially delayed. The appropriate target might be the probability of delivery within five days, not the mean. Research design must therefore follow the promised service outcome rather than the most convenient estimator.

4.7.4 Hypothesis tests, p-values, and practical relevance

A p-value is the probability, under a specified statistical model including a null hypothesis, of obtaining a test statistic at least as incompatible with that model as the statistic observed. It is not the probability that the null hypothesis is true, the probability that findings occurred “by chance,” or the magnitude of an effect. The American Statistical Association warns against basing scientific or business conclusions on crossing a single p-value threshold, against selective reporting, and against confusing statistical significance with practical importance (Wasserstein & Lazar, 2016). The controversy is not solved by replacing all p-values with one other mechanical criterion: interval estimates, substantive effect sizes, prior evidence, model assumptions, and decision consequences remain necessary (Amrhein et al., 2019).

Suppose a large online platform estimates that a design change raises order completion by 0.08 percentage points with a narrow interval. That result might be statistically distinguishable from zero but financially negligible or ethically harmful if it relies on manipulative defaults. Conversely, a smaller study could estimate a clinically or commercially important improvement imprecisely; a nonsignificant result would not prove the intervention has no effect. Multiple outcomes and repeated model selection can generate misleadingly small p-values unless the search process, preregistration where appropriate, and adjustment or confirmation strategies are disclosed. Statistical interpretation should specify what would count as consequential before inspecting favorable results (Ioannidis, 2005).

Two distinct statistical traditions are often treated as one. Fisher's significance-testing tradition uses the p-value as a measure of incompatibility between data and a hypothesized model, whereas Neyman–Pearson procedures emphasize predeclared decision rules, error rates, and alternatives. Routine business reporting often combines the language of both without acknowledging the different inferential purposes. Neither tradition turns a threshold into a posterior probability that an intervention works. For consequential choices, the estimated effect and its plausible range should be compared with a minimum substantively worthwhile benefit, the costs of implementation, and the relative harms of false positive and false negative decisions. Wasserstein et al. (2019) therefore argue against reducing evidence to whether a p-value falls on one side of 0.05.

4.8 RELATIONSHIPS, REGRESSION, AND CAUSATION

4.8.1 Correlation is not an intervention effect

Correlation measures the direction and strength of association according to a specified statistic, but does not itself establish causal direction. Pearson's r measures linear association for numerical variables; Spearman's rank correlation is useful when an ordinal or monotonic relationship is more appropriate. Both

can be dominated by nonlinearity, outliers, subgroup composition, or restricted range. A correlation between staff training expenditure and output may arise because training improves capability, successful companies can afford more training, or both reflect better management. Explaining the relationship requires a causal theory and evidence capable of distinguishing alternatives.

The distinction among association, prediction, and intervention is particularly important with automated analytics. A customer characteristic may be an excellent predictor of loan repayment yet impossible or impermissible to manipulate. A price reduction may causally increase sales even when price and quantity correlate positively in uncontrolled historical records because managers raise prices in high-demand periods. For business decisions the relevant question is often counterfactual: what would the same eligible enterprise, customer, or transaction have experienced under an alternative policy at the same time? Potential-outcomes reasoning formalizes this otherwise inaccessible comparison (Rubin, 1974).

4.8.2 Regression interpretation and omitted-variable bias

In an ordinary least squares model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, β_1 is the fitted change in expected Y associated with a one-unit change in X under the specified model. It is not automatically a causal effect; causal interpretation additionally requires an identification argument, valid measurement, and assumptions about the assignment of X, error structure, and model specification. With several covariates, coefficients are conditional associations holding included regressors fixed in the linear approximation. Heteroskedasticity, clustering, serial dependence, influential cases, multicollinearity, and functional-form errors affect the stability or uncertainty of estimates in different ways.

Omitted-variable bias emerges when an excluded determinant of Y is related to X. A hypothetical productivity regression omitting store location may attribute a dense urban customer base to an adopted application if higher-demand stores adopt technology first. A larger sample cannot remove a structural confounder automatically; it may merely yield a more precise estimate of the wrong contrast. Reverse causality and simultaneous determination further complicate management data. Claims about “drivers” based on ranking predictive features require care, because a feature's importance to a predictive algorithm is not equivalent to the causal effect of changing it.

The familiar omitted-variable-bias formula makes the underlying argument concrete. Suppose the true linear relationship is $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$, but the analyst regresses Y on X while omitting Z. Under the relevant linear projection assumptions, the coefficient on X equals $\beta_1 + \beta_2 \, \mathrm{Cov}(X,Z)/\mathrm{Var}(X)$. The bias depends jointly on Z's relationship to Y and its association with X; its direction cannot be inferred from an attractive regression coefficient alone. For example, if more capable managers both adopt software and increase sales, attributing the whole adoption–sales association to the software will tend to overstate the software contribution when the two correlations are positive. This algebra is diagnostic, not a claim that any observed Rwandan regression has a specific bias magnitude.

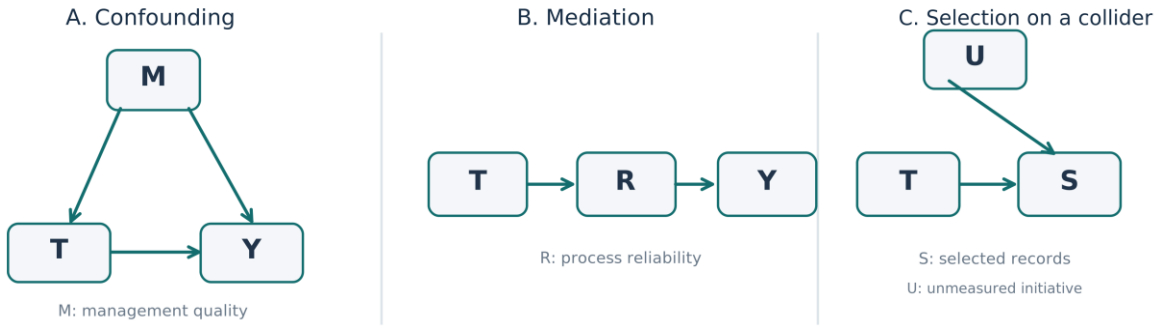
4.8.3 Causal diagrams and identification

A directed acyclic graph can document hypothesized causal paths. If prior management quality M affects both technology adoption T and sales Y, the open back-door path $T \leftarrow M \rightarrow Y$ confounds the observed relationship; measuring and properly adjusting for M may help if additional assumptions hold. However, adjusting for a variable caused by T can remove part of the effect of interest, and conditioning on a collider can introduce bias that was absent beforehand. A useful diagram is therefore not just arrows connecting attractive variables; it is a testable specification of ordering, omitted pathways, unit of analysis, and feasible interventions (Pearl, 2009; Hernán & Robins, 2020).

Causal inference also demands a credible definition of treatment. “Digital transformation” is too heterogeneous to be a single identifiable intervention unless the program's content, timing, adoption, and alternatives are specified. Spillovers and market equilibrium responses can undermine the assumption that one firm's outcome is unaffected by another's treatment. These concerns are especially important in cooperatives, industrial clusters, and supply-chain networks, where interventions often propagate through shared counterparties. External validity then requires considering the business's size, market power, technical skills, labor conditions, and institutional setting before transporting a result from another country.

A causal diagram should also distinguish confounders from mediators and colliders. If training improves competence and competence raises productivity, adjusting for competence may remove part of the total effect of training that a manager is trying to evaluate. If both training and unmeasured employee initiative influence whether workers remain in an evaluation sample, conditioning on continued employment may induce an association between treatment and initiative even where randomization originally made them independent. The diagram's value lies in identifying which variables should be measured and adjusted for a specific estimand, and which apparently helpful controls may create bias. No directed graph establishes the truth of its own causal arrows; chronology, substantive evidence, and diagnostic testing remain necessary (Pearl, 2009; Imbens & Rubin, 2015).

Figure 4.1 | Three causal roles in the evaluation of business technology adoption



Note. Original causal teaching diagram. The hypothesized confounder management capability (M) affects both adoption (T) and sales (Y); technology may act through process reliability (R); conditioning on a collider can create an association. Arrows are illustrative causal assumptions, not empirical findings.

4.8.4 Identification, model criticism, and triangulation

A credible study compares its key assumptions with plausible alternatives. For a difference-in-differences estimate, preintervention patterns and differential shocks deserve scrutiny; for an instrumental-variable design, the proposed exclusion restriction requires substantive justification; for regression discontinuity, manipulation around the threshold and local applicability must be evaluated. Fixed effects do not automatically resolve time-varying confounding; matching does not balance variables never measured; machine-learning controls do not create random assignment. Negative controls, placebo tests, alternate specifications, out-of-sample checks, and independent replications can strengthen a case when their purposes are understood. None independently turns an observational association into proof.

An ethically and intellectually serious analyst should state how the preferred conclusion could be wrong. Perhaps a time series contains a structural break, an intervention coincides with new regulation, or the measurement algorithm changes after deployment. Triangulating administrative, survey, and process evidence can help distinguish such explanations, provided their weaknesses are not correlated. Different

sources repeating the same original error are not independent corroboration. Evidence-based enterprise decisions should preserve the possibility of revision when new information contradicts the current account.

A compact constructed difference-in-differences example makes the identifying burden visible. Suppose an adopting branch's monthly on-time rate rises from 70% to 82%, while a comparable nonadopting branch rises from 75% to 80%. The difference-in-differences contrast is (82 − 70) − (80 − 75) = 7 percentage points. This is an arithmetic contrast, not a causal finding. It becomes a defensible treatment-effect estimate only if counterfactual trends would have been appropriately parallel, treatment timing is correctly defined, other differential shocks are addressed, and outcomes are measured consistently. A pretrend plot can reveal some violations, but failing to reject a pretrend difference does not prove that the unobserved postintervention counterfactual is correct. Staggered adoption further requires estimators appropriate to heterogeneous effects and timing rather than uncritical use of a single two-way fixed-effects coefficient (Hernán & Robins, 2020).

4.9 BUSINESS FORECASTING AND DECISION ANALYSIS

4.9.1 Forecasting is not explanation

Forecasting estimates future outcomes from information available at the forecast origin, conditional on assumptions and the stability of a data-generating process. A seasonal retail forecast may use past sales, holiday timing, weather, prices, promotions, and stockouts. Time-series analysis distinguishes trend, seasonality, cycles, irregular shocks, and level changes; naïve benchmarks can be surprisingly competitive when historical series are short or structural conditions change. The proper evaluation holds out future observations in chronological order and compares errors against a relevant simple alternative rather than testing a model on the same observations used to select it (Hyndman & Athanasopoulos, 2021).

Forecast accuracy must be tied to managerial consequences. Mean absolute error, $MAE = (1/n)\Sigma|y_t - \hat{y}_t|$, expresses average absolute deviation in original units; root mean squared error, $RMSE = \sqrt{[(1/n)\Sigma(y_t - \hat{y}_t)^2]}$, penalizes large errors more heavily. Percentage errors can become unstable where actual demand approaches zero, and pooled averages may hide serious failures in small but vital product lines. Prediction intervals communicate outcome uncertainty under an assumed model; they are not guarantees against regime changes, missing demand, abrupt policy interventions, climate shocks, or conflict. A forecasting process requires regular recalibration and clear separation between information available when a prediction was made and information revealed afterwards.

The evaluation horizon should match the operational horizon. A model optimized for next-day volume need not be useful for quarterly staffing, and a forecast produced after the order deadline has no decision value even when accurate. Consider a constructed three-month test with actual orders of 100, 120, and 80, and forecasts of 110, 100, and 90. Absolute errors are 10, 20, and 10 orders, so $MAE = 40/3 \approx 13.33$ orders; squared errors sum to 600, giving $RMSE = \sqrt{(600/3)} \approx 14.14$ orders. Those values evaluate only the declared test observations. A valid managerial assessment would compare them with a contemporaneously available naïve forecast, examine bias and service failures, and calculate whether lower error produces better inventory or staffing choices. Shmueli (2010) explains why apparent explanatory fit is not a substitute for this out-of-sample evaluation.

4.9.2 Scenario planning, sensitivity, and stress tests

Scenarios are internally coherent descriptions of possible conditions, not necessarily probability forecasts. A Rwandan export-oriented processor might examine a baseline supply environment, an adverse logistics disruption, and a favorable regional-demand expansion. Each should specify observable drivers, trigger indicators, effects on price, sourcing, lead times, cash, staffing, and feasible responses. Sensitivity analysis varies one or several model inputs to reveal which assumptions determine a recommendation; stress testing examines whether the enterprise can survive extreme but plausible adverse conditions. A scenario becomes useful when it changes a decision or contingency plan, rather than merely presenting three narrative labels.

When assumptions interact, isolated one-variable sensitivities can mislead. A currency depreciation may increase imported packaging costs while supporting export receipts; a transport interruption may simultaneously reduce sales and increase financing needs. Analysts should record dependencies among shocks, whether probabilities are estimated or elicited, and which risks cannot be meaningfully summarized in a single expected value. An enterprise with thin working capital may require a liquidity floor even when a project offers attractive expected profits. These constraints are part of rational decision design, not defects in analytical sophistication.

4.9.3 Expected value, utility, and the decision-tree worked example

A decision tree represents choices, uncertain events, conditional outcomes, and consequences over time. At each chance node, probabilities must sum to one for mutually exclusive, exhaustive alternatives; at each decision node, the analyst compares feasible choices given information then available. Expected monetary value (EMV) is the probability-weighted sum of outcomes, $EMV(a) = \Sigma_s\, p_s V(a,s)$, where a denotes an action, s an outcome state, p_s the state's probability, and V the monetary consequence. Expected value is appropriate for an explicitly risk-neutral decision maker under the stated payoff model; organizations facing insolvency risk, ethical constraints, or diminishing marginal value of funds may need expected utility, loss limits, or multiple criteria instead.

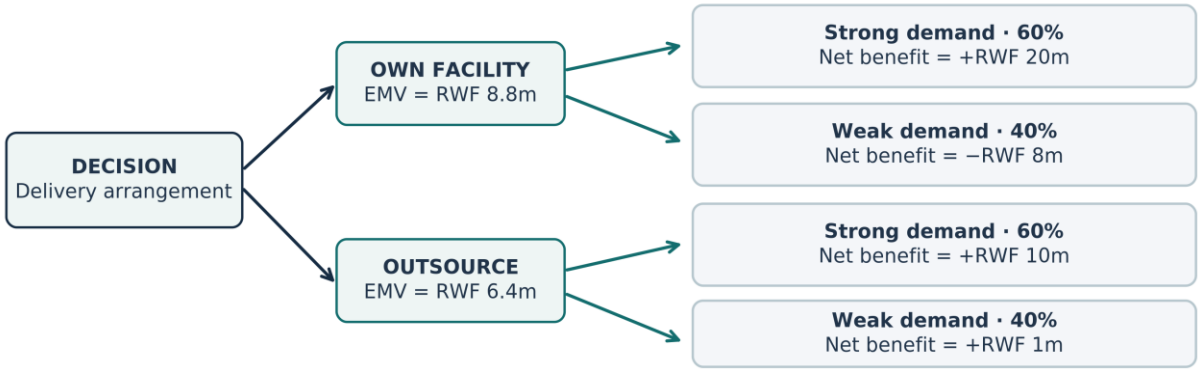
Consider a *hypothetical* Rwandan distributor deciding between a proprietary delivery facility and outsourcing. Management assigns a 0.60 probability to strong demand and 0.40 to weak demand, based solely on a teaching assumption. The facility yields a net present benefit of RWF 20 million under strong demand and a loss of RWF 8 million under weak demand; outsourcing yields benefits of RWF 10 million and RWF 1 million, respectively. EMV(facility) = 0.60(20) + 0.40(−8) = RWF 8.8 million; EMV(outsource) = 0.60(10) + 0.40(1) = RWF 6.4 million. A risk-neutral choice with no further constraints therefore prefers the facility by RWF 2.4 million. These are constructed payoffs, not market observations or observed probabilities.

If the demand state were known with certainty before the commitment, management would choose the facility in the strong state and outsourcing in the weak state. Expected value with perfect information is EVwPI = 0.60(20) + 0.40(1) = RWF 12.4 million, and expected value of perfect information is EVPI = 12.4 − 8.8 = RWF 3.6 million. EVPI is an upper bound on what a risk-neutral decision maker should pay for costless, timely, genuinely perfect information within this model; it is not the value of a real survey with errors and delay. The break-even strong-demand probability p solves $20p - 8(1 - p) = 10p + 1(1 - p)$, so $p = 9/19 \approx 47.37\%$. Below that assumed probability, outsourcing has the higher expected monetary value. The threshold illustrates how uncertain beliefs, rather than a spreadsheet's apparent precision, govern the decision.

Table 4.2 | Illustrative decision-tree payoffs (RWF million)

Investment option	Strong demand (p = 0.60)	Weak demand (p = 0.40)	Expected monetary value
Own delivery facility	20	−8	8.8
Outsource	10	1	6.4
Best decision with perfect state information	20	1	12.4

Figure 4.2 | Alternative delivery arrangements and expected monetary value



Note. Author-constructed scenario in RWF million. Payoffs, probabilities and expected values match Section 4.9.3 and Table 4.2; no real enterprise data are represented.

Table 4.2 contains invented decision payoffs and probabilities. The difference between the perfect-information benchmark and the best current decision is RWF 3.6 million; risk tolerance, financing constraints, contract provisions, and stakeholder impacts require separate evaluation.

Expected monetary value should not be mistaken for an instruction to choose the alternative with the most favorable average outcome under any circumstances. In the present illustration, the facility exposes the enterprise to an RWF 8 million loss in the weak-demand state, while outsourcing retains a positive RWF 1 million outcome in both states. If the facility's downside would violate a lender covenant or exhaust wages and supplier payments, a liquidity-constrained enterprise may rationally select outsourcing despite the facility's higher risk-neutral expected value. This is a distinction between optimization under a specified utility and feasibility structure and merely comparing means. A responsible decision record states the financial horizon, whether values are already net of capital expenditure, the distribution of losses, and how uncertainty affects solvency and stakeholder duties (Kahneman & Tversky, 1979).

4.9.4 Information value, learning, and decision governance

The value of sample information is lower than or equal to the value of perfect information under a coherent common model, but evaluating it requires the diagnostic accuracy of the proposed study, its cost, delay, and the options still available afterwards. A pilot project can create information through reversible action, although pilot participants may differ from the intended scale-up population. Decisions should log the analysis date, alternatives rejected, thresholds, probability sources, funding constraints, sign-off authority, and review trigger. Such a record allows subsequent managers to judge whether a disappointing outcome reflected poor reasoning or a low-probability event rather than rewriting history around results (Simon, 1955).

The value of realistically imperfect information can be calculated rather than assumed. Retain the hypothetical demand probabilities of 0.60 and 0.40, and suppose a timely market signal has 80% sensitivity for strong demand and 75% specificity for weak demand. The probability of a positive signal is 0.60(0.80) + 0.40(0.25) = 0.58; Bayes' rule gives P(strong | positive) = 0.48/0.58 ≈ 0.8276 and P(strong | negative) = 0.12/0.42 ≈ 0.2857. Applying the 47.37% decision threshold, management would own the facility after a positive signal and outsource after a negative signal. The unconditional expected outcome of this signal-based policy is [0.48(20) + 0.10(−8)] + [0.12(10) + 0.30(1)] = RWF 10.3 million. Its expected value of sample information is therefore 10.3 − 8.8 = RWF 1.5 million, before the cost of testing and delay. The calculations assume fixed state-dependent payoffs, accurate signal properties, and availability of both choices after the signal. EVSI remains below the RWF 3.6 million perfect-information ceiling, as it must in this coherent model. The exercise does not suggest that these diagnostic probabilities are observed market characteristics.

4.10 RESEARCH SYNTHESIS AND LITERATURE REVIEWS

4.10.1 Review questions, protocols, and search design

Literature review is not a chronological list of authors but an exercise in identifying what reliable evidence establishes, contradicts, or leaves uncertain about a defined question. A targeted critical review can be appropriate when the objective is to develop conceptual understanding across diverse theories; a systematic review requires explicit eligibility criteria, repeatable search methods, screening records, quality appraisal, and transparent synthesis. Scoping reviews map the nature and extent of evidence, while meta-analysis statistically combines suitable effect estimates under assumptions of meaningful comparability. These forms should not be conflated merely because a manuscript uses the expression “systematic” (Page et al., 2021).

For an enterprise question such as whether employee training improves productivity among smaller African firms, a defensible protocol should identify eligible populations, training interventions, comparators, outcomes, designs, period, language restrictions, sources, and planned effect measures. Search expressions need synonyms and indexing terms, including “microenterprise,” “SME,” “skills,” “management practices,” “training,” “productivity,” “Africa,” and relevant country names. Sources may include scholarly bibliographic databases, institutional repositories, working-paper series, evaluation repositories, and cited references. Search dates, platform settings, and exact strings should be recorded so that later researchers can reproduce or update the search.

4.10.2 Screening, extraction, and risk of bias

Transparent screening proceeds from retrieved records to deduplicated records, title and abstract decisions, full-text eligibility, and included studies, with reasons for exclusions documented according to the chosen review design. Data extraction identifies the study's country, sector, firm size, intervention, sample, outcome, time horizon, sampling frame, identification approach, effect estimate, uncertainty, funding, and major limitations. Studies are not automatically equivalent merely because they use the same outcome label. Productivity per worker, revenue growth, and subjective sales expectations represent different constructs, and comparability depends on the denominator, price adjustment, and observation period.

Risk-of-bias appraisal asks how allocation, deviations from intended treatment, missing outcomes, outcome measurement, and selective reporting may distort an estimate. An impressive sample size does not overcome an invalid comparison group. Publication bias may favor positive findings, while outcome

switching and specification search inflate apparent confidence. Evidence synthesis should inspect heterogeneity and plausible moderators rather than presume a single treatment effect transferable to every sector or country. Ioannidis (2005) provides a formal warning about the combined consequences of low power, bias, multiple testing, and selective publication; his argument does not imply that any particular empirical paper is false without an appraisal of its actual methods.

Critical appraisal must be appropriate to the design. Randomized evaluations demand scrutiny of random-sequence generation, allocation deviations, attrition, and measurement; observational evaluations place additional weight on confounding, selection, intervention classification, and time alignment; descriptive surveys require sampling-frame and response-process analysis. An evidence hierarchy that mechanically places every experiment above every observational study is methodologically unhelpful when the experiment measures the wrong outcome or applies to an irrelevant population. More importantly, a systematic-review label does not itself demonstrate that included studies are sufficiently comparable to justify a pooled estimate. Decisions about combining effects should be justified by interventions, populations, estimands, and outcome definitions, not solely by a statistical heterogeneity statistic.

4.10.3 Transparent synthesis and disagreement

The PRISMA 2020 statement is a reporting guideline for systematic reviews and includes items and flow diagrams intended to make review conduct visible (Page et al., 2021). It does not confer methodological quality by citation alone or authorize inventing identification and screening counts. Narrative synthesis is defensible where interventions and outcomes are too heterogeneous for pooling, provided it explains how studies were grouped and competing findings weighted. A meta-analysis should justify effect-size transformations, weighting, heterogeneity assessments, and sensitivity analyses; a statistically pooled result need not be the most practically transferable result to a small enterprise in Rwanda.

A disciplined review acknowledges the distinction between “no convincing evidence of benefit” and “convincing evidence of no meaningful benefit.” It should identify what future study would be informative, which design could discriminate competing hypotheses, and whether a new study's cost and ethical burden are justified. Research synthesis becomes a decision input only when the reported effects are translated into the intended local population, feasible implementation, possible costs, and relevant stakeholders. It cannot replace local measurement of whether the intervention was actually delivered as designed.

4.11 RESPONSIBLE ANALYTICS AND AI ASSISTANCE

4.11.1 Reproducibility, replicability, and reproducible decisions

Computational reproducibility requires that the same data, code, dependencies, and execution rules yield the reported analytical result. Replication tests an empirical claim with independently collected or otherwise independent evidence and suitable methods. Conceptual reproducibility also requires a third party to reconstruct how the business question became a variable definition, an analysis, and a recommendation. A PDF containing only polished charts is not sufficient if the analyst cannot identify the source table, filters, formulas, or version used. Wilkinson et al. (2016) supply useful data stewardship principles, although FAIR data do not guarantee correct inference or ethical use.

A minimal analysis repository includes read-only source copies where permitted, a dated data dictionary, reproducible scripts or transparent spreadsheet formulas, software versions, validation checks, a figure inventory, a decision memo, and a disclosure of transformations and exclusions. Sensitive microdata must not be publicly released merely to satisfy transparency rhetoric; access-controlled replication, de-

identified outputs, or synthetic demonstration data may be more appropriate. The controlling principle is that evidence should be auditable to the degree feasible without exposing individuals or violating contracts and law.

4.11.2 Algorithmic prediction and model validation

AI and machine-learning tools may detect patterns in large data, assist classification, generate forecasts, extract themes from documents, or help draft code. Their usefulness depends on the quality of training examples, alignment with the target decision, distribution stability, and safeguards against measurement and demographic bias. Training accuracy should be distinguished from performance on genuinely held-out observations; temporal or organizational leakage can make a model appear excellent while failing in deployment. Performance should be tested by subgroup and across plausible shifts, with attention to calibration, false positives and negatives, uncertainty, and consequences of error (Breiman, 2001; Tabassi, 2023).

A model that predicts customer loan repayment may still be unsuitable for approving or denying credit when protected characteristics, proxies, opaque explanations, or feedback effects produce unjustified exclusion. A low average error can conceal high error among less represented groups. Model performance must be examined against nonalgorithmic alternatives and the actual operational decision threshold, not only on a generic benchmark. The NIST AI Risk Management Framework is voluntary, cross-sector guidance that emphasizes governance, mapping, measurement, and management of AI risks; it is not itself Rwanda's binding law (Tabassi, 2023). The Organisation for Economic Co-operation and Development's (OECD) Recommendation on Artificial Intelligence, originally adopted in 2019 and amended in 2024, supplies an international normative reference whose jurisdictional status should not be overstated (OECD, 2024).

Model monitoring requires more than checking whether a single accuracy measure has deteriorated. Data drift concerns changes in the distribution of predictors; concept drift concerns changes in the relationship between predictors and the outcome; feedback effects arise when the model's own decisions reshape future training labels. For example, a credit model trained only on applicants who received loans lacks repayment outcomes for rejected applicants. Retraining on subsequent accepted borrowers can reinforce the original selection process and create unjustified claims of accuracy for excluded groups. Validation must specify the decision population, missing outcomes, fairness criteria, calibration, and practical consequences of intervention. Data statements documenting the populations and conditions of collection can improve transferability judgments, although they do not remove bias by themselves (Bender & Friedman, 2018).

4.11.3 Generative AI and research integrity

Generative models can assist with search query ideas, coding prototypes, table formatting, translation drafts, and explanatory illustrations. They can also hallucinate citations, merge conflicting sources, invent numerical results, overlook rare cases, and produce plausible but unverifiable syntheses. Human researchers remain accountable for every factual assertion, methodological decision, calculation, and bibliography entry. Data that are confidential, identifiable, embargoed, or subject to restrictive access terms should not be entered into unapproved external systems. AI-generated text cannot substitute for inspecting original law, primary methodological documents, or actual statistical metadata.

Responsible disclosure should identify material AI-supported activities and their purpose, such as drafting a reproducibility script or assisting language editing, subject to publisher policy. An AI assistant is not an authorial witness to unperformed research and cannot certify that a reference has been independently

verified. The relevant ethical duties are truthfulness, attribution, privacy, fairness, and transparent accountability, not whether prose resembles or evades a detection algorithm. Unsupported claims of independent research, surveys, model validation, or peer review are research-integrity failures regardless of the technology used.

A material distinction exists between assistance with a reproducible operation and substitution for a source of evidence. An AI system may generate code to compute a sample variance, but the analyst must inspect the code, test it against known inputs, and record the version executed. A model-generated bibliographic citation must be checked against a publisher or persistent identifier before use. A generated paragraph that plausibly describes an official statistic is not evidence that the statistic exists, and a polished synthetic image is not an authentic record of a real enterprise. Disclosure should be proportionate and informative, specifying what tasks were assisted, what evidence was independently checked, and whether confidential material was exposed. Automated fluency is neither a methodological guarantee nor a substitute for authorial accountability.

4.11.4 A proportionate governance framework

Analytics governance should assign a decision owner, data custodian, methodological reviewer, risk owner, and authority for intervention or shutdown. Before deployment, the team should document the model's intended use, prohibited uses, training and target populations, known limitations, validation results, fallback procedure, error escalation, and monitoring frequency. Oversight should be proportionate to consequence: automated stock-level recommendations can be piloted with routine checks, while employment discipline, lending exclusion, or safety decisions demand stronger scrutiny and meaningful avenues for human review. Periodic revalidation is essential when processes, populations, or incentives change.

The authority to deploy an analytics system should follow a documented sequence of challenge rather than a purely technical acceptance test. Management first establishes the legally and operationally permissible purpose, evaluates whether the proposed feature set is necessary, and compares an automated model with the best simpler alternative. Independent review then examines source legitimacy, target-population coverage, outcome validity, calibration and subgroup errors, cybersecurity exposure, and availability of contestation. A release decision records residual risks, accountable owners, rollback conditions, and actual monitoring capacity. Where impacts are irreversible or may affect livelihoods, small average accuracy gains cannot substitute for a meaningful explanation of error costs and human review. Model governance becomes evidence-based only when deployment decisions can be reopened in response to adverse outcomes rather than treated as settled by a one-time approval (Tabassi, 2023).

Proportionality is especially important for resource-constrained enterprises. A small Rwandan distributor may not need a dedicated machine-learning committee, but it does need a named record custodian, documented access permissions, a clear definition of the indicator used, a process for correcting wrong records, and an accountable manager authorized to suspend a faulty automated recommendation. Conversely, a large financial platform can inflict substantial harm through small error rates applied to many people and therefore warrants formal validation, incident tracking, independent assurance where appropriate, and structured appeals. The scale of oversight should reflect the consequences and reversibility of errors, not a fashionable distinction between “digital” and “traditional” business. No voluntary international framework should be mistaken for a substitute for applicable Rwandan law or sector-specific regulatory duties (OECD, 2024; Tabassi, 2023).

4.12 REPORTING, VISUALIZATION, AND COMMUNICATION

4.12.1 The architecture of a defensible evidence report

A serious management or scholarly report should make its argument inspectable rather than merely attractive. The opening identifies the decision, target population, objective, proposed comparison, and implications of choosing incorrectly. Methods describe where data came from, inclusion rules, measurement construction, analytic steps, model assumptions, and ethical constraints. Results distinguish observed values from estimates and uncertainties, and discussion considers competing explanations, external relevance, and action conditions. The conclusion recommends a feasible option only within the evidence's limits and identifies what would make management reconsider. This structure is compatible with APA's emphasis on transparent reporting, while specialist reporting guidelines should be applied where their study designs are relevant (von Elm et al., 2007; Page et al., 2021).

Reporting should avoid false numerical precision. An estimate such as 3.6000% is not necessarily more credible than 3.6%; an error bar without a label is not interpretable; and an association should never be described as an effect unless design warrants causal language. Academic APA style requires consistent author–date citations, clear tables and figure notes, and bibliographic correspondence. Businesses additionally need an executive summary that differentiates the recommendation, evidence strength, key uncertainty, financial implications, responsible officials, and monitoring trigger. These audiences differ in detail, not in standards of honesty.

4.12.2 Evidence tables and uncertainty statements

An evidence table should record the decision question, source, population, reference period, measure, estimate, uncertainty, design limitations, proposed action, and verification status. The table should distinguish direct evidence from proxy indicators and analytic judgment. Where uncertainty cannot be quantified, it should still be described concretely: for example, “the administrative series excludes unregistered businesses,” “the source does not publish a design-based standard error for this contrast,” or “the forecast assumes stable customer composition.” The absence of a confidence interval is not evidence of zero uncertainty.

A useful limitation section names constraints in a way that could change a decision: incomplete coverage; incompatible administrative classifications; lack of control group; unmeasured confounding; revisions; small effective sample; model instability; data-protection restrictions; or transfer from a different market. Merely inserting “further research is needed” is insufficient. Limitations should connect to their likely direction of bias where knowable, affected conclusions, and feasible remedies. Readers should be able to distinguish a high-confidence operational fact from a low-confidence explanation, forecast, or projection.

The language of uncertainty can be standardized without implying a spurious universal confidence scale. A report may identify sampling uncertainty with an interval; model uncertainty through the range across reasonable specifications; scenario uncertainty through a bounded set of economic conditions; and structural ignorance through explicit unknowns that cannot currently be quantified. For example, a delivery-time estimate may have a narrow sampling interval while remaining unreliable because records exclude failed deliveries. Presenting only the interval creates an illusion that the dominant uncertainty has been measured. The executive decision memo should identify which uncertainties are decision-reversing, which can be reduced at reasonable cost, and which must be governed through safeguards rather than more calculation.

4.12.3 Dashboard critique and the politics of indicators

Dashboards are useful for timely monitoring, yet they can make an enterprise seem controllable through whichever values happen to be continuously visible. Green and red threshold markers can conceal arbitrary standards, uneven denominator definitions, and the strategic manipulation of performance measures. A dashboard reporting rising sales but not increasing overdue receivables may reward commercial overextension; a workforce dashboard reporting output per employee but not injuries or unpaid overtime may produce ethically distorted decisions. Such failures are not caused only by poor visualization but by the institutional authority assigned to incomplete metrics.

A defensible dashboard incorporates indicator definitions, data freshness, segment visibility, uncertainty where appropriate, and clear ownership of corrective action. It should also provide links to the underlying evidence and signal when a metric has been redefined. Managers should periodically ask which consequential outcomes are not displayed, which people are excluded from measurement, and which behaviors the metrics encourage. The integration of research, management, and ethics is therefore substantive: good evidence can improve a decision only when governance permits inconvenient findings to be heard.

4.13 RWANDA-BASED SECONDARY-DATA APPLICATION AND INTERNATIONAL COMPARISON

4.13.1 Research question, dataset, and bounded estimand

An applied question for a business analyst is: How do selected reported operating characteristics differ between formal and informal non-agricultural business enterprises in Rwanda in 2024, and what can managers reasonably infer from those contrasts? The source is the NISR's official *Integrated Business Enterprise Survey (2024)* summary, published on its statistical-publications website with report and table downloads updated by March 2026. The publisher reports 278,060 business enterprises within the relevant survey's coverage and an informality rate of at least 87%. Its definition of formal businesses concerns registration with the Rwanda Revenue Authority and keeping regular business accounts; the covered activities are non-agricultural. The descriptive estimands below are published group percentages, not effects of becoming formal (NISR, 2026).

This distinction matters because formality is associated with size, location, financing, industry, customer base, capabilities, and other enterprise characteristics. Businesses selecting or able to achieve formal status may differ before any formalization-related effect occurs. Comparisons between two survey-defined categories can therefore identify disparities in observed characteristics but cannot tell managers whether formal registration would cause a business to export, adopt solar energy, or change waste-disposal practice. The previous chapters' emphasis on institutional context and enterprise heterogeneity supplies a substantive reason for preserving this distinction.

4.13.2 Published inputs, reproducible transformations, and results

Three pairs of percentages appear directly in NISR's published 2024 summary: 3.6% of formal versus 0.06% of informal enterprises engage in exports; 7.4% of formal versus 6.0% of informal enterprises use solar energy; and 78.1% of formal versus 51.7% of informal enterprises pay someone to dispose of waste. For each pair, calculate a percentage-point difference $\Delta pp = p_{formal} - p_{informal}$, where both p values are entered in percentage units, and a descriptive ratio $R = p_{formal}/p_{informal}$ when the latter is nonzero. This

requires no simulated observations or inferred survey weights. Reproduction is possible with the displayed inputs, a calculator, and the cited public webpage (NISR, 2026).

Table 4.3 | Official Rwanda IBES 2024 group percentages and original derived contrasts

Enterprise characteristic	Formal (%)	Informal (%)	Difference (percentage points)	Formal/informal ratio
Enterprises engaged in exports	3.6	0.06	3.54	60.00
Enterprises using solar energy	7.4	6.0	1.40	1.23
Enterprises paying for waste disposal	78.1	51.7	26.40	1.51

Source: NISR (2026), *Integrated Business Enterprise Survey (2024)*, publication summary; differences and ratios computed for this chapter. Values retain the source's reported rounding. The official summary does not supply sampling standard errors for these specific pairs; no confidence intervals, causal attribution, or statistical-significance tests are presented here. Solar use is an adoption measure, not the proportion of enterprise energy obtained from solar power. Waste-disposal payment is not equivalent to safe waste management, and export engagement is not the share of sales exported.

The arithmetic can be independently checked. Export difference: 3.60 − 0.06 = 3.54 percentage points; export ratio: 3.60/0.06 = 60. Solar difference: 7.4 − 6.0 = 1.4 percentage points; solar ratio: 7.4/6.0 ≈ 1.2333. Waste difference: 78.1 − 51.7 = 26.4 percentage points; waste ratio: 78.1/51.7 ≈ 1.5106. A 60-fold ratio sounds dramatic, but it arises partly because the comparison group's reported export rate is exceptionally small. Ratios and absolute percentage-point contrasts answer different questions and should be reported together rather than selectively using the more impressive one.

4.13.3 Interpreting the results without making unwarranted claims

The widest absolute difference in this selected set concerns payment for waste disposal at 26.4 percentage points; the widest relative ratio concerns export engagement at 60 to one. Neither is an intervention effect. Differences may reflect market access, scale, urban concentration, operational activity, ability to pay, and selection into formal status. The small solar contrast cannot establish the absence of an environmental financing gap, because the binary indicator does not measure installation capacity, equipment cost, or the proportion of energy use served. Moreover, the summary publishes rounded group estimates without the complex survey design variance required to test whether observed differences exceed plausible sampling error. A responsible manager would use these patterns to frame questions, not to advertise that formality increases export probability sixtyfold.

The survey's broader findings reinforce the importance of carefully defined evidence: NISR reports that 62.6% of surveyed manufacturing companies did not use their full capacity, with low demand and problems with reliable input supply among cited reasons; it also identifies gaps in foreign-language communication and advanced IT skills (NISR, 2026). Those observations motivate candidate investigations into market development, workforce capability, and input coordination, but they do not quantify the effect of any specific training, financing, or procurement policy. A firm considering intervention would need its own verified process data and an appropriate causal or forecasting design.

The comparison also illustrates why descriptive ratios require denominator scrutiny and a decision-oriented interpretation. The observed export rate of 0.06% among informal enterprises is very small, so

the 60:1 ratio is sensitive to rounding and to any definitional or measurement error in that group. The percentage-point contrast of 3.54 communicates a different feature: fewer than four additional enterprises per hundred are represented by the formal-group export proportion when the two reported percentages are treated arithmetically. Neither should be interpreted as a counterfactual number of new exporters caused by a registration policy. The solar-energy and waste-disposal items also illustrate that an apparently environmental metric may capture technology adoption or the payment of a service rather than measured emissions, waste safety, or ecological outcomes. A substantive policy recommendation would need better-defined outcomes and additional evidence about firms' constraints.

4.13.4 Cross-country transfer and the limits of comparability

Comparison with the World Bank Enterprise Surveys illuminates a methodological trap. The World Bank's standard formal-sector methodology covers eligible registered, partially privately owned businesses with at least five workers in selected sectors and uses stratified sampling by size, sector, and geographic region. It deliberately excludes many informal and micro-scale activities that are material to the NISR universe (World Bank, n.d.). Therefore a World Bank survey statistic from another African economy cannot be subtracted from the NISR's full non-agricultural Rwandan percentage as if they were observations of equivalent populations. A valid comparison would first align definitions, eligible industries, enterprise sizes, geography, survey period, questionnaire wording, treatment of nonresponse, and weights.

This is not an argument against cross-national analysis. It is a stronger requirement for it. After harmonizing sampling frames and definitions, analysts can compare characteristics and investigate institutional moderators such as export procedures, electricity reliability, digital infrastructure, financing conditions, and firm-size distributions. Where harmonization is incomplete, comparative work can still discuss design differences qualitatively and specify the unanswered empirical question. The general lesson is especially important for business education in Rwanda: the most visible international datasets may be strongest for formal firms while leaving the enterprises most numerous in the local economic landscape outside their inference frame.

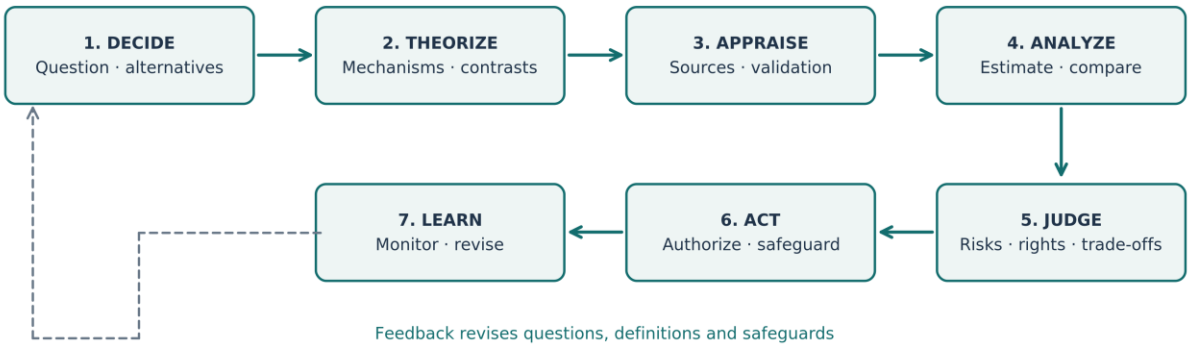
A limitation is the publication medium itself. Transcribing three percentages from an official summary allows the contrasts to be recalculated, but does not independently validate the survey's original sampling frame, weight calibration, questionnaire translation, nonresponse corrections, or raw data cleaning. Nor does it turn a rounded descriptive result into a confidence interval. The reproducibility claim made here is restricted to the transformation of visible source figures, with the substantive source conditions explicitly attributed to NISR (2026).

4.13.5 A transparent reproduction protocol

A reader can recreate Table 4.3 without proprietary software by opening the NISR 2024 IBES source page, recording the publication title, target reference year and webpage date, transcribing the three quoted pairs of group percentages, subtracting informal from formal values, dividing formal by informal values, and retaining source rounding. For reproducibility, the output should preserve the URLs, calculations, units, and any changes observed upon revisiting the page. The original raw microdata and design weights are not part of this exercise; consequently a reader can verify the arithmetic but cannot independently regenerate the agency's survey-weighted estimates from the displayed evidence alone. This distinction between computational reproduction and full empirical replication is material, not a minor methodological footnote.

4.14 INTERFUNCTIONAL INTEGRATION, ETHICS, AND RESEARCH OPPORTUNITIES

Figure 4.3 | Evidence-to-decision workflow for the integrated enterprise



Note. Original conceptual instrument developed for this chapter from evidence-based management, research-design, and decision-analysis principles. Arrows represent proposed decision stages, not empirically estimated causal effects.

4.14.1 An evidence-to-decision system for the integrated enterprise

The evidence-to-decision workflow in Figure 4.3 connects enterprise purpose with defensible research. A decision owner first specifies the question, affected stakeholders, alternatives, and threshold for action. The analyst identifies theories and sources, audits definitions and provenance, designs descriptive, predictive, causal, or decision analysis as appropriate, and communicates estimates alongside nonquantified uncertainties. The responsible manager then makes a documented choice, implements safeguards, and commissions follow-up evaluation. Feedback may revise the original question, decision model, data architecture, or policy. This is a proposed management instrument rather than a universally validated causal sequence.

The workflow works only if its functional interfaces are taken seriously. Marketing may propose expanding into a new segment, but its survey estimates must connect to operations capacity and financial contribution. Human resources may recommend training based on employee outcomes, but the design needs adjustment for role mix and interaction with process technology. Procurement may report that the cheapest supplier has the lowest invoice cost, while finance and operations assess disruptions, working capital, and service failures. Governance ensures that the source, assumptions, risks, and stakeholder consequences are visible to those authorized to decide. Later chapters specialize these functions; Chapter 4 supplies their shared inferential grammar.

4.14.2 Reproducibility and provenance checklist

The following checklist should accompany any substantial secondary-data recommendation, scaled to the sensitivity and cost of the proposed decision. Its function is to make a result inspectable, not to imply that checking boxes can substitute for methodological judgment. Missing items must be recorded with an explanation and an associated risk rather than silently presumed to exist.

Table 4.4 | Evidence reproduction and provenance checklist

Audit element	Check required	Verifiable evidence	Reviewer response
Decision specification	Population, intervention/comparison, outcome, horizon	Dated question and estimand	Pass / explain gap
Source origin	Publisher, URL/identifier, retrieval and version date	Archived source and metadata	Pass / explain gap
Measurement	Unit, codebook, denominator, exclusions, missing-value rules	Data dictionary and validation log	Pass / explain gap
Transformations	Formula, join, filtering, weighting, software/version	Executable script or transparent workbook	Pass / explain gap
Numerical checks	Totals, alternative calculations, outliers, sensitivities	Independent verification record	Pass / explain gap
Inferential warrant	Design, selection, assumptions, uncertainty	Methods and limitation statement	Pass / explain gap
Rights and security	Licence, authorized access, confidentiality, retention	Permission and access register	Pass / explain gap
Action accountability	Decision owner, alternative, trigger, review date	Signed decision log and monitoring plan	Pass / explain gap

Table 4.4 is an original audit checklist derived from the FAIR principles, statistical reporting guidance, and the evidence-to-decision framework developed in this chapter (Wilkinson et al., 2016; Wasserstein & Lazar, 2016). It is not a formal certification standard.

4.14.3 Ethics as a criterion of evidence quality

Business analytics has normative consequences because measurement privileges some experiences and excludes others. A profit-only decision model may make uninsured injuries or downstream pollution disappear from the table while they remain real burdens. An employee-performance algorithm may appear accurate in historical records that encode discriminatory supervision. A business that manipulates customer defaults can generate superior short-term conversion statistics without creating legitimate value. Normative judgment should be explicit: people should not be harmed or unfairly excluded merely because a dataset renders such outcomes invisible. Empirical evidence can estimate some consequences, but it cannot supply the final value criterion by which those consequences are judged.

Responsible evidence practice demands truthful description of data provenance, fair representation of affected populations, proportionate security, transparency of uncertainty, and accessible challenge procedures. In Rwanda and elsewhere, enterprises must verify the governing privacy, employment, consumer, and financial rules before processing protected records or automating consequential decisions. Ethical analysis also extends to the costs of poor inference: confidently repeating weak evidence can waste scarce public or enterprise resources, mislead investors, and burden vulnerable workers. A transparent "we do not know" can be more professionally useful than a precise but ungrounded forecast.

An ethics of evidence must also confront epistemic injustice: whose experiences count as valid evidence and who has authority to challenge an official interpretation? Customers without smartphones, informal traders absent from tax registers, workers whose grievances never reach an online portal, and rural enterprises poorly represented in digital transaction logs may disappear from analytical systems while remaining directly affected by their decisions. Their exclusion is simultaneously a measurement problem and a matter of institutional fairness. A responsible analyst should examine where data access, language, procedural power, and recourse determine who is visible, then state whether proposed policies could shift harm toward those omitted. This does not establish that every qualitative account is factually correct; it means the burden of evidence cannot be discharged merely by declaring unrecorded experiences nonexistent.

4.14.4 Researchable propositions and unresolved debates

Five propositions arise from the foregoing synthesis and should be understood as hypotheses, not validated effects. First, organizations using documented decision questions and source definitions may experience fewer metric-definition disputes than similar organizations lacking them, conditional on organizational scale and data maturity. Second, the relationship between analytics investment and operational performance may be moderated by complementary workforce capability and process reliability. Third, improved causal identification may alter the selection of managerial interventions independently of improvements in forecast accuracy. Fourth, transparent uncertainty and stakeholder involvement may strengthen the legitimacy of consequential decisions even when a single monetary performance index deteriorates. Fifth, research designs that explicitly account for informal-enterprise coverage may yield materially different development inferences from formal-sector-only sampling frames.

Each proposition has rival explanations and feasible research designs. A study of documentation could compare matched branches or exploit a staged rollout while checking selection and spillovers; capability complementarities could be assessed using suitably measured firm panels with credible identification; forecasting versus intervention analysis could be contrasted in multiple natural experiments; legitimacy would require carefully validated measures rather than mere public-relations sentiment; and frame effects could be evaluated by harmonizing overlapping populations across national and international surveys. These proposals invite multidisciplinary testing and should not be cited as findings of this chapter.

4.15 INTEGRATIVE SYNTHESIS AND CONCLUSION

The central intellectual contribution of this chapter is a distinction among four indispensable qualities of evidence. Relevance asks whether the inquiry answers the actual decision rather than a convenient proxy. Measurement integrity asks whether the data validly represent the stated unit, construct, and population. Inferential integrity asks whether descriptive, predictive, causal, or decision claims are supported by the design and its assumptions. Responsible use asks whether the resulting action can be justified to those who bear its risks, respects law and rights, and remains open to correction. None compensates automatically for failure in another. Precise estimation of an invalid construct is not strong research; a causal estimate of an irrelevant intervention is not useful management; and beneficial forecasts do not remove ethical obligations.

The Rwandan IBES example exposes why this discipline is practical. Officially reported formality-group differences can be replicated arithmetically and used to identify important business questions. Yet the same table cannot support a causal claim about registration, a confidence interval without variance information, or an unqualified international ranking. In the decision-tree exercise, transparent probability

assumptions make an expected-value recommendation possible while identifying the belief threshold that would reverse it. These examples show that managerial rigor consists not of extracting one definitive number but of clarifying the conditions under which evidence changes a justified choice.

The durable principle is therefore to build a revisable, accountable evidence system rather than fetishize the latest analytics instrument. New AI applications and expanding administrative datasets can reduce information costs and widen participation, but also multiply risks of biased coverage, opaque models, fabricated outputs, and misleading apparent certainty. Research competence gives managers a disciplined means to question those outputs. Chapter 5 turns to the functions and responsibilities of management, for which evidence-informed planning, delegation, monitoring, and correction are essential; later chapters apply the methodological foundation to specialized business decisions.

4.16 KEY TERMINOLOGY, EDUCATIONAL ASSESSMENT, AND FURTHER READING

4.16.1 Conceptual glossary

An estimand is the precisely defined quantity an inquiry seeks to learn about a population or intervention. A construct is a theoretically specified attribute, such as trust or managerial capability; construct validity concerns the support for a particular interpretation of its measurements. A sampling frame is the operational list or selection mechanism through which eligible units enter a study. External validity concerns the conditions under which results from those units can inform judgments about other populations, settings, or periods.

Selection bias is distortion arising when a study's inclusion process is associated with the outcome or its causes. Confounding occurs when a noncausal pathway produces or alters an observed exposure–outcome association. Endogeneity, in regression terminology, describes dependence between a regressor and the disturbance under the assumed model; it can arise from confounding, simultaneity, measurement error, or selection. A counterfactual is the potential outcome under an alternative action at the relevant decision point.

Missing at random is an assumption that, conditional on included observed information, the chance of missingness does not additionally depend on the missing value. A standard error summarizes an estimator's sampling variability under its sampling or statistical model. A confidence interval is generated by a procedure with a stated repeated-sampling coverage probability under its assumptions. A p-value quantifies the extremeness of an observed test statistic relative to a specified null model; it is not the probability that the null hypothesis is correct.

A forecast origin is the time at which predictors and other information must have been available to make a prediction. Expected monetary value is the probability-weighted average financial consequence of an action in a defined set of mutually exclusive states. Data provenance records the origin, transformations, versions, and permission conditions of evidence. Computational reproducibility is the ability to regenerate a reported result from disclosed inputs and procedures, which is distinct from reproducing the underlying data collection or independently replicating the empirical claim.

4.16.2 Learning questions and applied tasks

Foundational comprehension (LO1–LO3). Readers should explain, with a specific example, why the best available descriptive contrast is not automatically an intervention effect. They should compare an official enterprise survey, a government administrative register, and a company customer relationship management database in terms of unit of observation, frame coverage, measurement error, and admissible

inference. Finally, they should specify a decision-ready estimand for reducing late orders in a Kigali-based enterprise and defend the proposed measure and comparison.

Quantitative application (LO4–LO6). Using Section 4.6, students should independently reproduce the illustrative delivery-time mean, median, on-time proportion, sample variance, and standard deviation, then evaluate which indicators best represent service reliability. For Table 4.3, they should verify all NISR-based percentage-point contrasts and ratios, state the role of source rounding, and explain why the available evidence does not license a design-based p-value. They should recalculate the decision-tree EMV at a 0.40 probability of strong demand and identify nonfinancial conditions that could override the risk-neutral ranking.

Advanced critique (LO5–LO8). Assume a manager claims that a digital ordering system caused productivity growth because adopting branches improved more than nonadopters. Students should draw a causal diagram with at least three plausible pre-adoption confounders, identify one variable that must not automatically be controlled, and compare a suitable identification design with a weaker before–after analysis. They should propose a placebo or negative-control test, discuss measurement changes and selection, and assess transportability when microenterprises are absent and adoption coincides with regulatory change. An acceptable answer distinguishes a credible estimate from the ethics of imposing the policy.

Capstone-style integrative assignment (LO1–LO8). Students should develop a 2,500–3,500-word decision memorandum for a Rwanda-based enterprise comparing two delivery or market-access interventions. The memorandum must use a verifiable secondary dataset, dated provenance ledger, defined population and estimand, reproducible descriptive calculation, an uncertainty or sensitivity analysis, a decision tree or scenario comparison, and a substantive stakeholder-risk assessment. It must identify the evidence unavailable without primary collection, distinguish causal from descriptive claims, recommend a monitored action, and identify an explicit condition that would reverse the recommendation. Evaluators should reward methodological defensibility and truthful uncertainty rather than a predetermined commercial conclusion.

4.16.3 Educator guidance and assessment criteria

For introductory learners, evaluate the correct alignment of question, population, measure, denominator, and descriptive calculation. Intermediate students should be able to distinguish sample statistics from parameters, explain confidence and p-value misconceptions, and evaluate the limitations of association. Advanced learners should defend their causal assumptions, check source comparability, calculate decision thresholds, consider stakeholder harms, and specify conditions under which they would revise recommendations. A suggested 100-point rubric assigns 20 points to conceptual specification, 20 to source quality and provenance, 20 to analysis and reproducible calculations, 20 to causal and uncertainty judgment, and 20 to communication, ethics, and actionable monitoring. The rubric is pedagogical, not an accreditation instrument.

Educators can assess causal judgment without requiring specialist econometrics. In response to the digital-ordering capstone, a strong student distinguishes the policy contrast from a simple adoption association, identifies at least one plausible source of confounding, and explains which proposed check would address it. A stronger response also considers how attrition, treatment spillovers, and outcome redefinition could invalidate the result even if the proposed regression is correctly calculated. For the expected-value tasks, the benchmark answer should show each probability-weighted payoff, calculate the relevant threshold,

and explain how insolvency risk or stakeholder obligations could change the choice. Such assessment prioritizes defensible reasoning over complex software or unsupported numerical precision.

4.16.4 Selected further reading

Rousseau (2006) is a foundational conceptual treatment of evidence-based management; Shadish et al. (2002) and Hernán and Robins (2020) provide deeper treatments of design and causal identification; Cochran (1977) and Little and Rubin (2019) clarify sampling and missing-data problems; Hyndman and Athanasopoulos (2021) give applied forecasting guidance; Wasserstein and Lazar (2016) critically explain inferential reporting; and Page et al. (2021) define transparency expectations for systematic-review reporting. Readers applying these works should use the latest available corrected editions and documented source-specific methods, not a single methodological slogan.

For methodological progression, readers may begin with Rousseau (2006) and Briner et al. (2009) on the professional use of evidence, then study Shmueli (2010) on the distinction between explanation and prediction. Hernán and Robins (2020) provide a rigorous framework for causal identification, while Imbens and Rubin (2015) develop its potential-outcomes foundations. Hyndman and Athanasopoulos (2021) offer a practical and openly available treatment of forecasting and honest test-set evaluation. These works are complementary, not competing claims that a single technique can settle every enterprise problem.

REFERENCES

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-based management: Concept cleanup time? *Academy of Management Perspectives, 23*(4), 19–32. https://doi.org/10.5465/amp.23.4.19

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology, 183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC. https://miguelhernan.org/whatifbook

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

National Institute of Statistics of Rwanda. (2026). *Rwanda Integrated Business Enterprise Survey Report 2024* [Statistical report and online publication summary]. https://statistics.gov.rw/statistical-publications/business-establishment-finance-trade/business-establishment-finance-trade/integrated-business-enterprise-survey-2024

Organisation for Economic Co-operation and Development. (2024). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449; adopted 2019, amended 2024). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. *Harvard Business Review, 84*(1), 62–74, 133. https://hbr.org/2006/01/evidence-based-management

Rousseau, D. M. (2006). Is there such a thing as "evidence-based management"? *Academy of Management Review, 31*(2), 256–269. https://doi.org/10.5465/amr.2006.20208679

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. https://doi.org/10.2307/1884852

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLOS Medicine, 4*(10), e296. https://doi.org/10.1371/journal.pmed.0040296

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

World Bank. (n.d.). *Enterprise Surveys: Methodology*. Retrieved October 24, 2026, from https://www.enterprisesurveys.org/en/methodology