CHAPTER 17

OPERATIONS AND PRODUCTION MANAGEMENT

Sixbert SANGWA

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

PARENT BOOK

The Business Enterprise: Core Functions, Integrated Management, and Emerging Transformations

https://doi.org/10.65655/ocp.m30

CHAPTER DOI

10.65655/ocp.m30.c126

VOLUME 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

COPYRIGHT AND STATUS

© 2026 Sixbert SANGWA | CC BY 4.0

ABSTRACT

Operations and production management concerns the design, coordination and control of processes through which enterprises transform resources into dependable goods and services. Its analytical difficulty lies in managing interacting constraints: demand is uncertain, capacity is finite, variability creates queues, physical assets deteriorate, and apparently efficient decisions may impose costs on workers, customers and the environment. This chapter develops a critical and applied treatment of twelve approved operations domains, integrating operations strategy, process analysis, the theory of constraints, lean production, forecasting, queueing, maintenance and sustainable operations. Through a transparent, purposive synthesis of seminal research, peer-reviewed studies, standards and official secondary evidence, it interprets Rwanda's 2024 Integrated Business Enterprise Survey and 2026 national accounts without confusing sectoral aggregates with individual enterprise performance. An explicitly hypothetical Kigali food producer illustrates capacity, overall equipment effectiveness, bottleneck improvement, location economics and operational measurement; a service counter demonstrates the nonlinear consequences of high utilization under stated queueing assumptions. The chapter offers original process and accountability diagrams, an operations-metric register, management decision instruments, and a Rwanda–Toyota comparative interpretation. It argues that operational excellence is not the maximization of utilization or elimination of every buffer. Rather, it is a context-sensitive capacity to meet legitimate demand reliably, learn from variation, protect people and resources, and maintain economically and ethically defensible performance. Proposed integrative relationships are conceptual and require empirical testing.

Keywords: operations management; production systems; process mapping; demand forecasting; capacity planning; bottlenecks; queueing theory; lean operations; operational resilience; Rwanda.

LEARNING OUTCOMES AND CONCEPTUAL ROADMAP

On completing the chapter, readers should be able to (LO1) explain an operations system and critically compare its principal theoretical explanations; (LO2) map processes and identify constraints, rework, handoffs and flow-time risks; (LO3) build and evaluate demand forecasts with reproducible error measures; (LO4) distinguish design, effective and realized capacity and calculate utilization, throughput and equipment effectiveness; (LO5) evaluate facility location, layout, scheduling, buffers and automation as contingent decisions; (LO6) design an integrated operational scorecard that addresses quality, cost, delivery, flexibility, safety and resource productivity; (LO7) interpret Rwanda-specific evidence against international operational practice without unsupported generalization; and (LO8) defend a responsible operations improvement using explicit assumptions, alternatives, sensitivity tests and measurement safeguards.

The analytical progression begins with the scope and strategic meaning of operations (17.0–17.1), then follows the flow of work from process design and forecasting to capacity, space, scheduling and buffers (17.2–17.7). Technology, measurement, service operations, reliability and responsibility develop the control and adaptation problem (17.8–17.12). A Rwanda–international analysis, integrative conceptual propositions, glossary, assessment activities and verified references conclude the chapter. Chapter 16 established the promises made by sales and distribution; the present chapter asks how productive and service systems make those promises feasible. Procurement contracting belongs principally to Chapter 18, logistics and detailed inventory policy to Chapter 19, comprehensive quality systems to Chapter 20, project management to Chapter 21, costing to Chapter 23, digital infrastructure to Chapter 28 and enterprise-wide risk to Chapter 32.

17.0 INTRODUCTION: WHEN DEMAND EXCEEDS THE RELIABILITY OF A PROCESS

Consider a Kigali enterprise making shelf-stable food products for independent retailers and institutional buyers. The sales team celebrates a new monthly order; finance approves working capital; production managers know that the filling and sealing station is already unreliable. Raw-material preparation appears efficient and the kitchen has unused hours, yet packaging queues grow, small batches are reworked and dispatch becomes unpredictable. The owner proposes a second cooking vessel because the kitchen looks busy. A more exact diagnosis could show that the proposed investment increases unfinished work without increasing saleable output. This case is constructed solely for teaching. Its purpose is to isolate a recurring managerial error: treating utilization of separate resources as a proxy for performance of the entire value stream.

Operations management is the disciplined design, operation, improvement and responsible renewal of the processes by which resources are transformed into products and experiences. Production management is its historically manufacturing-focused component, although production also occurs in farms, professional services, hospitals, schools and digital platforms. Every operation has inputs, an arrangement of activities, a transformation logic, controls, and outputs. However, not every output is valuable, not every reported unit is saleable, and not every productive arrangement is sustainable. A restaurant prepares meals while customers participate in the service; a manufacturer may produce goods before the customer is known; an online service combines software infrastructure with human review. The appropriate unit of analysis is therefore the operational system, not merely a machine, employee or department (Slack et al., 2022; Frei, 2006).

The governing question is how organizations design and control processes that convert resources into reliable goods and services under uncertainty. Three connected judgements are necessary. The technical judgement asks what flow, capacity, time, safety and quality performance can be achieved. The economic judgement asks what options cost, which bottlenecks determine incremental return, and what liabilities accompany growth. The normative judgement asks whether deadlines, labor intensity, environmental burdens or customer exposure are justified. No isolated overall-equipment-effectiveness percentage, order count or automation forecast resolves all three. A defensible decision joins a realistic operations model with the capacity for correction, human voice and proportional risk controls.

17.0.1 Research design, source hierarchy and inferential limits

This chapter uses a purposive, critical secondary-source synthesis rather than a systematic review, original survey, plant audit or randomized intervention. Original operations scholarship was selected for distinctive mechanisms concerning manufacturing strategy, flow laws, constraints, variability, lean practices, maintenance and sustainability. Peer-reviewed empirical research was used to test where practice-based claims are stronger or weaker; primary institutional and professional sources were used for standards and Rwanda-specific descriptive context. Targeted searches on 2 November 2026 covered publisher and journal records for Skinner (1969), Little (1961), Kingman (1962), Ferdows and De Meyer (1990), Hopp and Spearman (2004), Shah and Ward (2007), McKone et al. (2001) and Kleindorfer et al. (2005), with official NISR, Toyota, ISO, ILO and UNIDO sources. Search expressions included “operations strategy manufacturing competitive priorities,” “Little law throughput lead time,” “lean production measures empirical,” “Rwanda

Integrated Business Enterprise Survey 2024 capacity utilization," and "operations management sustainable manufacturing." This is a disclosed targeted search, not a claim of exhaustive retrieval or PRISMA-compliant screening.

The evidence date is 2 November 2026. The National Institute of Statistics of Rwanda (NISR) reports statistics for its specified population and observation period, not audited capacity for any named business. Toyota's official production-system description documents its stated methods but cannot independently establish the effect size of those methods in Rwanda. Peer-reviewed studies support conclusions within their samples and designs; transfer depends on product variety, supply reliability, human skills, regulation, finance and infrastructure. All numerical teaching cases below are invented and transparently calculated. Their arithmetic can be reproduced from stated values; they are not disguised observations of a real enterprise. The review distinguishes operational accounting identities, approximations dependent on queueing distributions, causal claims supported by study designs, and normative proposals requiring managerial justification.

For reproducibility, the numerical examples report inputs, denominators and assumptions in the main text. A calculation can be independently recomputed without access to confidential firm data, but this does not make the resulting economic proposition an empirical estimate. The cases intentionally distinguish observed national statistics, theoretical identities, conditional stochastic models and constructed management choices. Searches were targeted rather than exhaustive; consequently, this chapter cannot quantify the prevalence of competing operations practices, resolve publication bias in the wider literature or claim that all relevant studies were retrieved. These limitations are substantive, because published factory studies often overrepresent formal organizations with record-keeping capacity, while Rwanda's smaller and informal businesses may face different constraints.

17.0.2 Historical development and the philosophical meaning of efficiency

The history of operations cannot be reduced to a succession from craft work to mass production to automation. Interchangeable parts, the division of labor, scientific management, standardization, statistical control and assembly lines gradually increased the visibility and measurability of tasks, while also creating concerns about worker discretion, quality and institutional power. Postwar operations research developed mathematical approaches to queues, scheduling and inventory; Japanese production research popularized just-in-time, autonomation and continuous improvement; service and digital operations later exposed the limits of models built around stable physical inventory. These methods coexist. A Kigali bakery may use handwritten planning, mobile payments and modern packaging equipment in the same workflow, without being accurately described by one technological generation.

Efficiency has at least three meanings that should not be confused. *Technical efficiency* compares feasible output with inputs; *allocative efficiency* asks whether resources are assigned to their highest-value available use under an assumed objective; and *institutionally legitimate efficiency* additionally evaluates whose costs are excluded from that objective. A process that reduces measured labor hours by transferring unrecorded work to unpaid family members is not necessarily socially more efficient. Likewise, requiring workers to bypass a machine guard to raise hourly throughput may increase immediate output while compromising safety and continuity. Skinner (1969) showed that manufacturing arrangements constrain strategic options; later research on cumulative capabilities questioned the universality of simple quality–cost trade-offs (Ferdows & De Meyer, 1990; Hayes & Pisano, 1996). The operational imperative is disciplined improvement under stated constraints, not an unqualified pursuit of speed.

The ethical consequences of standardization illustrate why the philosophy of efficiency remains disputed. Taylorist analysis could make hazardous or inconsistent work measurable and therefore improvable, yet time standards can also become instruments of coercive intensification when employees lack bargaining power. Reviewing early lean production research, Landsbergis et al. (1999) found reasons to question whether the practices described in automobile workplaces consistently increased workers' autonomy; their review also noted demands and work-intensity risks. Those findings should neither be universalized to every lean implementation nor ignored because a process achieves superior defect or inventory indicators. A mature

theory of operations must specify the distribution of control and gains, the mechanisms protecting workers, and whether reported improvement persists after accounting for fatigue, health and quality losses.

17.1 OPERATIONS AS A VALUE-CREATING FUNCTION

17.1.1 The transformation system and its boundaries

An operations system converts material, information, energy, labor, equipment time and supplier services into goods or service outcomes. Its design must specify what enters, which transformations occur, who authorizes movement between stages, what constitutes acceptable completion, and what happens to residuals. A packaged-food process may receive cleaned ingredients, packaging, utilities and customer specifications; preparation, cooking, filling and sealing alter these inputs; inspection and dispatch establish whether a saleable product exists. In a clinical service, the flow unit is an encounter and the output is not simply the number of patients seen: correct decisions, continuity, consent and waiting burden are part of the operational outcome. Defining the flow unit prevents misleading ratios, such as comparing gross kilograms processed with finished retail packs without adjusting for weight or yield.

Operational boundaries are managerial constructions. A producer may count only production time inside the factory, whereas the customer experiences lead time from order acceptance to receipt; managers may optimize the former while the latter deteriorates. The operating system must therefore be mapped at the level appropriate to the decision. The internal transformation boundary is necessary for line control; the order-to-delivery boundary is essential for customer promises; a life-cycle boundary is necessary when waste and energy burdens matter. A useful framework distinguishes a process's *physical flow*, *information flow*, *financial commitments* and *rights and obligations*. These are causally connected without having identical owners or measurement frequencies.

17.1.2 Competitive priorities and cumulative capability

Cost, conformance quality, dependability, speed, flexibility and responsible resource use commonly conflict in specific short-run decisions. More product variety may require additional setups; tighter delivery promises may need reserve capacity; a faster cycle may increase energy intensity or errors. Yet these are conditional relationships, not fixed laws of nature. Eliminating recurring defects can simultaneously reduce rework, lead time and unit cost. Ferdows and De Meyer (1990) argue that capabilities can accumulate, challenging a rigid trade-off interpretation. Their cumulative perspective must not be turned into a guarantee: resource-constrained enterprises can still face hard near-term budget, skill and equipment limits. The manager's task is to distinguish improvements that remove waste from choices that shift genuine trade-offs across stakeholders or periods.

Operations strategy links selected customer promises to investments in capacity, people, process technology, location, maintenance, planning and learning (Skinner, 1969; Hayes & Pisano, 1996). If customers value predictable delivery more than rapid customization, a stable repeating process and sensible capacity reserve may outperform equipment designed for maximum product changeability. If a public clinic must serve complex cases fairly, throughput objectives cannot displace triage and clinical safety. The operations strategy should therefore state the competitive or public-service priority, its measurable condition, and nonnegotiable legal or safety constraints rather than merely announcing “world-class efficiency.”

17.1.3 Theory and managerial interpretation

The theory of constraints (TOC), associated with Goldratt and Cox (1984), directs attention to the system element limiting throughput: identify the constraint, exploit it safely, subordinate nonconstraints, elevate capacity when justified and repeat the diagnosis as the limiting factor moves. TOC is a useful diagnostic discipline, but a constraint need not be an identifiable machine. It may involve regulatory release, unreliable power, insufficient demand, a specialist's authorization, or an upstream supplier. Lean thinking emphasizes waste reduction, stable flow, problem visibility and learning; Hopp and Spearman (2004) refine the

distinction by defining pull principally through limiting work in process and lean through minimizing the cost of buffering variability. Treating either approach as an unqualified mandate to eliminate reserves mistakes its mechanism. When supply or demand variance cannot be reduced economically, well-chosen inventory or spare capacity may be essential.

The empirical record also cautions against equating the adoption of operational tools with causal performance improvement. Shah and Ward (2007) identify several components of lean practice rather than a single universally measured treatment; plants scored as 'lean' may differ in training, supplier arrangements, management capability and production technology. In a randomized management intervention among large Indian textile plants, Bloom et al. (2013) reported a 17% first-year productivity increase through improvements including quality, efficiency and inventory management. The study is powerful evidence for the studied intervention and firms, not a transferable 17% uplift for a Rwandan SME. Its value for operations scholarship is more precise: deficient information and managerial routines can be binding constraints, and rigorous improvement evaluation requires an identified comparison rather than a post-adoption success narrative.

17.2 PROCESS MAPPING AND ANALYSIS

17.2.1 Process architecture, flow units and evidence

A process map is an explicit representation of the sequence, decisions, responsibilities and waiting states through which a defined unit moves. The analyst first chooses the unit and start/end events, then observes or documents activities, decision gates, queues, transfers, rework loops and exceptions. A *flowchart* makes the logical sequence visible, a *swim-lane map* shows accountability across departments, and *value-stream mapping* adds time, inventory and information. No map should be taken as a truthful observation solely because it is neat. It must be checked against transaction records, actual work practices, exceptions and workers' explanations. Formal procedures can be followed only under normal conditions, while informal workarounds dominate during shortages or peak loads (Spear & Bowen, 1999).

Process time is the actual time spent performing an operation; waiting time occurs when the unit cannot advance; transfer time is required for physical or informational movement; rework time corrects rejected work. Throughput time includes all relevant components between declared boundaries. Cycle time is ambiguously used in practice: it can mean time between consecutive completed units or processing duration for one unit. A defensible document defines the chosen meaning. Here, *output interval* denotes the average time between successive completions at a station, while *flow time* denotes time spent by one unit in the system. This avoids the mistaken conclusion that a four-minute bottleneck output interval implies only four minutes of end-to-end completion time.

17.2.2 Identifying bottlenecks and hidden constraints

In a serial process with no parallel stations, no starvation, identical output units and adequate supporting resources, the sustainable maximum rate is constrained by the slowest stage. If preparation requires 2.5 minutes per batch, cooking 3 minutes and sealing 4 minutes, nominal eight-hour capacities are 192, 160 and 120 batches respectively. The sealing station is the nominal physical bottleneck. Yet an operating system rarely meets these simplifying assumptions. Setup losses, breaks, material shortages, first-pass yield, maintenance and demand mix can shift the constraint. A station with the longest nominal cycle may not be the realized constraint when upstream variability or downstream inspection repeatedly blocks flow. Bottleneck diagnosis must therefore pair engineering standard times with actual time-stamped production, queues and reasons for stoppage.

An improvement at a nonconstraint can increase local efficiency without increasing final throughput. If preparation is raised from 192 to 220 batches per day while sealing remains capable of only 120, the extra preparatory capacity cannot raise finished output. Unless work is deliberately restrained, it will increase work in process, handling, spoilage exposure and working-capital needs. The practical lesson is not to neglect

nonconstraints, which may require preventive maintenance and protective capacity, but to evaluate investment at the level of end-to-end, good-quality production. The limiting factor should be reassessed after any genuine improvement, because the constraint migrates.

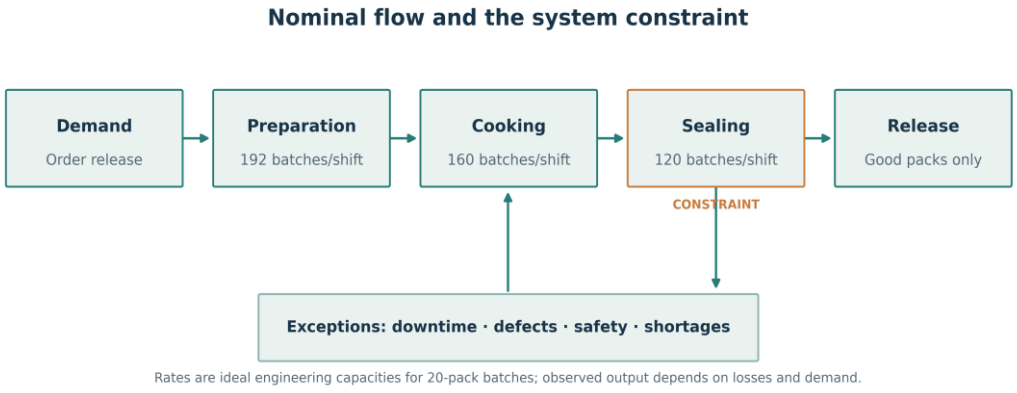
A process engineer should also distinguish the apparent bottleneck from the active bottleneck under the prevailing product mix. For the illustrative line, preparation, cooking and sealing have respective ideal rates of 192, 160 and 120 twenty-pack batches per eight-hour shift. The seal station sets the ideal whole-line rate, but this conclusion changes if cooking requires allergen-changeover cleaning, an inspection officer is available for only part of the shift, or the material recipe adds variable batch sizes. Figure 17.1 accordingly includes both engineering rates and a release and exception-feedback loop. The diagram is deliberately a nominal diagnostic, not a claim that actual plant flow always follows an uninterrupted serial sequence.

17.2.3 Little's law and the relation among work, throughput and time

Little (1961) established a powerful long-run flow relationship: $\mathbf{L = \lambda W}$, where L is average number of flow units in a stable system, λ is the average throughput rate in flow units per unit of time, and W is average time spent by a unit inside the same boundaries. In a stable packaged-food workflow averaging 240 packs in process and releasing 120 good packs per hour, average flow time equals 240 / 120 = 2 hours. This relationship does not identify why the time is long, and the averages must use matching unit boundaries and observation periods. It is not a universal formula for transient operations whose input rate persistently exceeds output or whose queues grow without bound. Nevertheless, it reveals a managerial consequence: increasing unfinished units without increasing throughput ordinarily lengthens time in the system.

The flow relationship is frequently confused with a prescriptive rule to minimize inventory to zero. Buffering can prevent starvation during material interruptions and protect essential service levels. The analytically sound objective is to choose the least costly and most responsible configuration of inventory, time and protective capacity under actual variability. Hopp and Spearman (2004) show why the meaning of pull becomes clearer when specified as a work-in-process control rather than as a ceremonial use of just-in-time terminology. Figure 17.1 expresses this chapter's operational accountability cycle; the diagram is an original synthesis rather than an empirically estimated causal model.

Figure 17.1. Operational process, bottleneck diagnosis and feedback



17.3 DEMAND FORECASTING AND OPERATIONAL PLANNING

17.3.1 Sources of demand uncertainty and forecast meaning

A forecast is a conditional statement about an uncertain future quantity given available information, a specified horizon and a prediction method. Operations require forecasts for materials, staffing, machine time, service appointments and capacity investment. Short-horizon predictions often rely on recent orders and known promotions, while long-horizon capital decisions depend on structural scenarios and institutional changes. Demand may combine trend, seasonality, calendar effects, sporadic institutional purchases,

promotions, weather and exceptional disruptions. Aggregated sales cannot always be treated as unconstrained demand: a stockout records what was sold, not necessarily what customers sought; rejected orders may reveal demand that the enterprise could not serve. Decisions built on sales data alone can therefore perpetuate under-capacity.

The *naïve forecast* carries forward the most recent observation; a moving average smooths recent values; exponential smoothing places declining weights on older observations; more sophisticated approaches can model trend, seasonality and covariates. Complexity should be justified by out-of-sample accuracy, available data and decision cost rather than prestige. Hyndman and Athanasopoulos (2021) emphasize exploring series properties, specifying training and test periods, and selecting error measures appropriate to the purpose. A model that minimizes average error across ordinary weeks may perform poorly at festival peaks when the cost of missed orders is greatest. Prediction intervals and alternative scenarios are indispensable where occasional shortages have nonlinear consequences.

17.3.2 Worked forecast and evaluation

Suppose monthly orders, in thousands of packs, are 10, 12, 11 and 14. A three-period moving-average forecast for month 5 is (12 + 11 + 14) / 3 = 12.33 thousand packs. This is a baseline, not a confidence bound and not a forecast that the factory can automatically satisfy. To demonstrate retrospective error evaluation, a one-step naïve forecast for months 2, 3 and 4 would be 10, 12 and 11 thousand packs, whereas actual values are 12, 11 and 14. Absolute errors are 2, 1 and 3, yielding mean absolute error (MAE) = (2 + 1 + 3) / 3 = 2 thousand packs. Weighted absolute percentage error (WAPE) = (2 + 1 + 3) / (12 + 11 + 14) × 100 = 16.22%. Because this is a four-observation illustrative series, it has no power to establish comparative superiority of one model or seasonality.

Managers should retain at least a simple benchmark, track signed errors to detect systematic underforecasting, evaluate by relevant product and horizon, and state how unusual periods are handled. Mean absolute percentage error can behave poorly when actual quantities are zero or near zero; WAPE can obscure weak performance in low-volume categories; a single aggregate measure may conceal severe errors for essential clients. Forecast quality must be judged against the decision: a 10% miss may be tolerable for a low-cost buffer but catastrophic for perishable regulated inputs. Records of demand censoring, promotional interventions and uncertainty should accompany forecasts, not be erased during management reporting.

17.3.3 From forecast to sales and operations planning

Sales and operations planning (S&OP) reconciles market expectations with resource feasibility across a medium-term horizon. In a mature process, commercial managers present demand and its uncertainty; operations assesses bottlenecks and feasible schedules; procurement reports input availability; HR assesses skill and shift constraints; finance tests cash, contribution and risk; executive authority resolves trade-offs. The process should produce an agreed, version-controlled plan, not an artificial “consensus forecast” created by averaging political demands. Lapide (2004) identifies disciplined meetings, clear preparation and cross-functional participation as important process conditions. The evidentiary limit is that S&OP quality cannot be established by the existence of meetings alone; adherence, exceptions, plan stability and outcomes must be tracked.

Scenario planning is preferable to false precision where weather, exchange rates, fuel costs or large contracts dominate variability. The base case can be compared with demand shocks, supplier interruptions and operational recoveries, using explicit trigger thresholds. A food producer might adopt base, downside and upside monthly order levels of 38,000, 32,000 and 44,000 packs, each paired with an admissible production, staffing and financing response. These values are hypothetical. A useful S&OP decision specifies the smallest feasible action for each state, the lead time to enact it, the constraint it relaxes and the evidence that will trigger reconsideration. No demand scenario should be silently treated as a realized sales commitment.

An effective review must distinguish forecast generation from decision approval. At a monthly S&OP meeting, commercial leaders might propose 44,000 packs in an upside state, while the current sealing process supports

only about 41,970 first-pass acceptable packs under the modeled operating factors. That difference cannot be erased by labeling the forecast 'agreed'. The legitimate options include verified overtime within lawful and safe limits, supplier-supported subcontracting, revised customer delivery dates, additional productive capacity, or declining low-contribution orders. Each option changes working capital, quality assurance and the distribution of risk. Management should record which trade-off was accepted and monitor the forecast separately from the feasibility of its chosen response.

17.4 CAPACITY PLANNING, UTILIZATION AND QUEUES

17.4.1 Capacity concepts and operational denominators

Design capacity is the theoretical maximum output under the stated engineering schedule and ideal conditions; *effective capacity* adjusts for planned maintenance, setups, staffing, quality controls and known operating constraints; *actual output* is the observed result in a specified period. A common teaching convention defines utilization as actual output divided by design capacity and efficiency as actual output divided by effective capacity, but terminology varies across organizations. An analyst must disclose denominator and time basis. A plant running at 95% of an assumed design capacity may experience serious overtime, quality losses or equipment failure; another with 65% utilization may intentionally reserve capacity for variable demand. There is no universal economically optimal utilization percentage independent of variability, risk, contracts and the cost of spare capacity.

Capacity planning spans different time horizons. Long-term decisions concern sites, equipment scale and product architecture; medium-term decisions concern shifts, staffing, outsourcing and maintenance windows; short-term decisions concern sequencing, overtime and rationing. The cost of undercapacity includes missed demand, long queues and reduced service reliability; the cost of excess capacity includes capital charges and ongoing facilities, insurance and staffing commitments. Capacity pooling can reduce the reserve required when demand variations are not perfectly correlated, whereas tightly specialized resources may provide high technical productivity but limited reallocation options. Choice therefore depends on demand composition and the economic consequences of shortage, not only average volume.

17.4.2 Reproducible capacity and overall equipment effectiveness

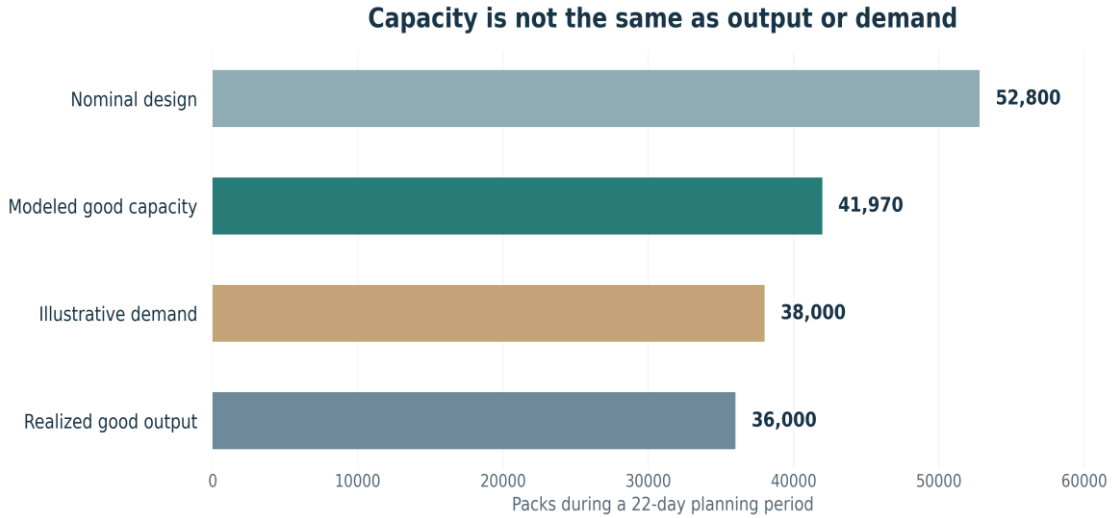
Take a hypothetical Kigali food-packaging line operating eight hours daily for 22 days per month. A sealing station takes four minutes per batch, each batch contains 20 packs, and nominal ideal output is (480 / 4) × 20 = **2,400 packs per day**, or **52,800 packs over 22 days**. Assume availability 90%, running-speed performance 92% of ideal, and first-pass good-pack fraction 96%. Overall equipment effectiveness (OEE), as defined for this exercise, equals availability × performance × quality = 0.90 × 0.92 × 0.96 = **79.488%**. Under the stated constant-mix assumptions, good-output capacity becomes 2,400 × 0.79488 = **1,907.712 packs per day**, or **41,969.664 packs per month**. This is a theoretical planning estimate, not observed production. The fraction may be used in calculations but finished packs must be counted as whole units in actual reporting.

Suppose actual first-pass good output is 36,000 packs during that month. Utilization relative to design capacity is 36,000 / 52,800 = **68.18%**, while achievement relative to modeled good-output capacity is 36,000 / 41,969.664 = **85.78%**. The different percentages describe different denominators; neither alone diagnoses why output was limited. Demand may be insufficient, staffing incomplete or packaging unavailable. Managers should disaggregate OEE losses, observe schedules and separate planned production from unplanned stoppages. If demand is 38,000 good packs, modeled good-output capacity appears sufficient, although the actual system's 36,000 output suggests an execution or scheduling gap. Buying an additional cooking vessel would be unjustified before identifying that gap.

A second productivity denominator makes the labor consequences inspectable. Assume, solely for this teaching calculation, four operators are present for all eight scheduled hours on each of the 22 days, with all their hours counted and no additional outsourced or unpaid labor. The line uses 4 × 8 × 22 = 704 operator-

hours; realized first-pass labor productivity is therefore 36,000 / 704 = 51.14 good packs per scheduled operator-hour. Reaching the same verified 38,000-pack demand with unchanged, safe hours would raise the ratio to 53.98. Neither figure is a national productivity benchmark or a claim about a real workforce, and neither captures capital, energy or indirect support labor. The comparison demonstrates why a productivity statement is meaningful only when its numerator, labor boundary and quality gate remain fixed.

Figure 17.2. Capacity denominators in the hypothetical Kigali production case



Note. Illustrative calculations: design = 52,800; good-output capacity = design × 0.90 × 0.92 × 0.96 ≈ 41,970; actual = 36,000; assumed demand = 38,000 packs. Not empirical company data.

17.4.3 Bottleneck improvement economics and sensitivity

Assume an engineering modification reduces the sealing standard from four to 3.5 minutes per batch without increasing risk or lowering yield. Its nominal capacity becomes (480 / 3.5) × 20 = 2,742.86 packs per day. Holding the illustrative OEE factors constant yields around **2,180.24 good packs per day**, above the baseline 1,907.71. Across 22 days, this is an additional modeled capability of approximately 5,996 good packs. Yet an actual forecast of only 38,000 packs against 36,000 achieved implies at most 2,000 incremental sales attributable to meeting this specific demand, unless new orders are demonstrated. If each extra sale contributes RWF 650 after truly variable costs, incremental monthly contribution is 2,000 × 650 = RWF 1,300,000, not 5,996 × 650. For a RWF 6,000,000 modification with incremental monthly support cost RWF 150,000, an undiscounted illustrative payback is 6,000,000 / (1,300,000 – 150,000) ≈ **5.22 months**. This is not a capital-budgeting recommendation: collection risk, returns, tax, lifetime, opportunity cost, service disruptions and the realized increment in sales could overturn it.

Sensitivity analysis exposes the assumption carrying the business case. If only 1,000 extra packs sell, net monthly incremental contribution becomes 1,000 × 650 – 150,000 = RWF 500,000 and simple payback becomes 12 months. If added output increases rejected packaging or injury exposure, the technical and ethical constraints fail even when nominal payback is attractive. An improvement proposal should therefore separate (a) physically achievable additional good units; (b) demand that can actually be served and collected; (c) full incremental costs; and (d) safety and legal thresholds. The distinction links this chapter to customer analysis (Chapter 14), pricing (Chapter 15), management accounting (Chapter 23) and financial investment appraisal (Chapter 24).

17.4.4 Queueing: why high utilization can undermine service

Queueing introduces a nonlinear element absent from simple average-capacity arithmetic. Kingman's (1962) heavy-traffic analysis explains why expected delay grows sharply as a variable system approaches saturation. As a fully specified illustrative *M/M/2* service model, suppose a counter has two parallel employees,

independent Poisson arrivals averaging eight customers per hour, exponentially distributed service durations averaging twelve minutes per customer, first-come-first-served discipline, unlimited waiting capacity and no abandonment. Each employee has service rate $\mu = 5$ customers/hour; aggregate nominal capacity is 10/hour and utilization $\rho = 8 / (2 \times 5) = 0.8$. Using the Erlang-C formula yields a probability of delay of approximately 0.7111 and mean queue waiting $W_q = 0.7111 / (10 - 8) = 0.3556$ hours, or **21.33 minutes**; mean total system time is 21.33 + 12 = **33.33 minutes**.

If arrivals were seven/hour with the same servers and assumptions, utilization falls to 0.7 and Erlang-C mean waiting becomes approximately **11.53 minutes**. A 12.5% reduction in the arrival rate lowers modeled waiting time by roughly 46%, demonstrating a nonlinear effect, not a prediction for an identified Rwandan service provider. Actual appointment systems, customer heterogeneity, batching, priority rules and staff breaks violate the simplified distributional assumptions and require other methods. No waiting-time target can be set responsibly without considering customer vulnerability and service safety. The practical point is that spare capacity at a variable service interface may create legitimate value through shorter delays, whereas maximizing measured staff occupancy can degrade the very experience being delivered.

The M/M/2 result can be reconstructed rather than accepted as a black-box software output. Set $a = \lambda/\mu = 8/5 = 1.6$ and $\rho = a/2 = 0.8$. The empty-system probability is $P_0 = [1 + a + a^2/\{2(1-\rho)\}]^{-1} = 1/9 = 0.1111$. The probability that an arrival waits is $C = [a^2/\{2(1-\rho)\}]P_0 = (6.4)(1/9) = 0.7111$. Finally $W_q = C/(2\mu-\lambda) = 0.35556$ hours, or 21.33 minutes. The calculation presumes stationary independent Poisson arrivals, exponential service times, identical servers and no abandonment. At $\lambda = 7$, the same construction gives $C \approx 0.5765$ and $W_q \approx 11.53$ minutes. It follows neither that waiting is always well modeled by these distributions nor that an observed queue should automatically be attributed to insufficient staffing.

17.5 FACILITY LOCATION AND LAYOUT

17.5.1 Location as an irreversible configuration of access and cost

Facility location decisions determine access to labor, suppliers, utilities, markets, transport links, skills and institutional services over a horizon often much longer than any individual forecast. A production site near Kigali can reduce delivery distance to some customers but may impose higher premises costs, limited expansion room or dependence on specific infrastructure. A site farther from the principal market may offer lower rent and closer agricultural inputs but expose finished products to higher transport, temperature-control or schedule risk. For a hospital or service desk, proximity, accessibility, disability access and continuity may be more important than the lowest land cost. Location must be evaluated against the actual service promise and full operational exposure, not merely a generalized comparison between urban and nonurban land prices.

Quantitative location screening can combine cost–volume analysis with weighted criteria, but neither method removes managerial judgment. Suppose two hypothetical sites have annual fixed costs of RWF 24 million and RWF 33 million respectively, and variable processing-related costs of RWF 760 and RWF 610 per saleable pack, holding all other comparable factors constant. At annual output Q, modeled costs are $\mathbf{C_A = 24{,}000{,}000 + 760Q}$ and $\mathbf{C_B = 33{,}000{,}000 + 610Q}$. The accounting crossover is $Q = (33{,}000{,}000 - 24{,}000{,}000) / (760 - 610) =$ **60,000 packs**. For output below 60,000, A is lower cost; above that, B is lower cost under the given assumptions. This result is only a screening identity. If B requires higher working capital, has unreliable water or adds delivery losses, the stated unit-cost function is incomplete. A robust location memorandum records land tenure, permits, safe water, power quality, occupational access, waste arrangements, expected growth and contingency alternatives.

Location should also be analyzed as a risk-portfolio decision. Concentrating all critical operations in one site may gain economies of scale but increase exposure to floods, localized power outages or social disruption. Dispersing facilities may reduce geographic concentration while creating duplication, governance complexity and transport costs. The business continuity analysis should identify the costs and lead times of restoring minimum essential service, not assume that a second facility is universally prudent. A site selected through a

narrowly financial spreadsheet can impose irreversible burdens on nearby communities or workers; environmental and social screening must therefore enter before investment approval, with obligations verified under current applicable local law.

17.5.2 Layout, process type and material handling

Layout translates the operating strategy into spatial and informational relationships. In a *product or line layout*, stations are sequenced around a high-volume repetitive flow; a *process or functional layout* groups similar equipment or specialist capabilities for flexible, diverse orders; a *cellular layout* brings resources together for product families with shared routings; and a *fixed-position layout* keeps a large immovable output in place while resources move to it. Services may use customer-facing zones, back-office cells, self-service interfaces and hybrid digital-physical arrangements. No layout is intrinsically superior; its effectiveness depends on demand volume and variety, variability, safety, setup costs, flow distance and future change. A highly specialized line may offer low unit processing cost at volume but become brittle when product variety or input dimensions change (Slack et al., 2022).

A practical layout study records travel distance, handling frequency, crossing flows, storage conditions, supervision, waiting areas, emergency access and infection or contamination pathways where relevant. A food-processing plant must prevent problematic intersections between raw and finished goods and preserve cleaning and inspection accessibility. A clinic must accommodate privacy and triage rather than optimize waiting-room seating as a simple throughput problem. Digital information layout also matters: a form requiring four approvals may create a virtual queue even when physical movement is efficient. Layout evaluation should combine visual mapping with time and safety observations, worker consultation and realistic change costs, including disruption during installation.

A common failure is choosing the minimum-distance layout without regard to adverse consequences. Compressing walkways can shorten material-handling time but obstruct evacuation; relocating a hazardous workstation closer to packaging can reduce movement while increasing contamination exposure. Conversely, redesign that reduces unnecessary lifting may lower both injury risk and handling waste. When comparing alternatives, managers should report at least cost, feasible output, time, flexibility, safety and resilience. Weighted scoring can discipline the discussion, but an unsafe option must not become “acceptable” simply because it receives a high aggregate score on cheaper criteria.

Facility alternatives also differ in option value. A low-cost site with no adjacent expansion space may become expensive once growth requires a second location; a modular layout could preserve flexibility at an initially higher installation cost. These effects warrant scenarios rather than a single extrapolated annual production volume. In a public-service context, relocation can improve managerial operating costs while increasing clients' transport time or excluding people with mobility limitations. Such uncompensated costs are relevant to the decision even when they do not appear in the facility's accounting ledger. Local consultation and transparent criteria therefore improve both informational quality and procedural legitimacy.

17.6 SCHEDULING AND RESOURCE ALLOCATION

17.6.1 From aggregate capacity to executable commitments

Scheduling assigns specific work to limited equipment, personnel, space and time within precedence, availability and deadline constraints. Aggregate capacity calculations can show that total monthly output is possible while concealing infeasibility on particular days or shifts. Orders may compete for one specialist machine, packaging format or quality-release officer. A credible schedule specifies job quantities, release times, required sequence, setup families, completion deadlines, resource calendars and contingency authority. Production control then compares actual progress with the committed schedule and records reasons for replanning. Planning rigidity is inappropriate where suppliers or patient demand are volatile, but continuous improvised rescheduling can also undermine reliability and conceal chronic undercapacity.

Common sequencing rules optimize different objectives. First-in-first-out privileges arrival order; earliest-due-date may improve deadline protection; shortest-processing-time tends to reduce mean completion time when all jobs are immediately available on one machine and other simplifying assumptions hold; critical-ratio rules compare remaining time to required work. There is no ethically neutral “optimal schedule” without an objective and constraints. A hospital cannot use shortest service time to displace urgent cases; an enterprise may need a legally mandated rest break even when a solver would otherwise assign overtime. Scheduling should represent skill coverage, fatigue, maintenance windows and customer priority rights as explicit constraints, not as discretionary exceptions after optimization.

17.6.2 A transparent sequencing illustration

Five hypothetical jobs require 2, 6, 3, 7 and 4 processing hours on one machine; all are ready at time zero, without setup or interruption. First-in-first-out in that order generates completion times 2, 8, 11, 18 and 22 hours, with mean flow time **12.2 hours**. Ordering jobs by shortest processing time, 2, 3, 4, 6 and 7 hours, generates completion times 2, 5, 9, 15 and 22 hours, with mean **10.6 hours**. Thus the latter reduces average flow time by 1.6 hours in this special case. It does not establish better on-time delivery if due dates differ, nor equitable treatment of long, urgent or socially essential jobs. The correct decision requires a stated service policy and sensitivity to actual arrival times, sequence-dependent setups and priority exceptions.

Finite-capacity scheduling should not confuse jobs on a spreadsheet with available capacity in a shift. If an eight-hour shift includes a planned 30-minute sanitation shutdown and a required break, assigning eight net production hours would generate systematic lateness. A practical scheduling system maintains versioned calendars and exceptions, assigns ownership of changes, and reports adherence both to the frozen schedule and to the originally promised customer date. A revised due date may make internal schedule attainment appear high without restoring the customer's experience. Distinguishing commitment changes from actual performance is therefore essential.

17.6.3 Bottleneck sequencing, flexibility and human constraints

At a constrained station, sequencing choices have system-wide effects because upstream work waits for the resource and downstream resources may be starved. Grouping similar products can reduce setup time but delay small urgent orders. Releasing all orders at once may create excessive work in process; releasing only when the bottleneck has available capacity can stabilize flow, provided input reliability is adequate. Cross-training offers flexibility, but “flexible staffing” can conceal overwork if minimum staffing, competence and safe working hours are ignored. Boudreau et al. (2003) demonstrate why the interface between operations and human resource management is analytically central: human responses are not simply random noise added to mathematical machinery.

Detailed project-network techniques are developed in Chapter 21. Here it is sufficient to distinguish precedence-constrained production jobs from projects: operational scheduling repeats interdependent resource assignments, while a one-off capital installation may require a project critical path before it can change productive capacity. A practical production plan records who owns dispatch authorization, shortage responses, priority overrides and the restoration of normal scheduling after disruptions. Escalation mechanisms prevent urgent sales requests from silently overriding health, safety or inspection duties.

For a more complete scheduling diagnosis, managers should record both flow and lateness distributions instead of only the average completion time. In the five-job illustration, the shortest-processing-time sequence reduces average flow time, but the last job still completes at hour 22 and may violate an earlier contractual due date. If a job requires a different sealing temperature or package type, setup times become sequence-dependent, and the apparent ranking can reverse. Robust scheduling therefore requires a calendar-accurate model, realistic setup families, transparent exception rules and a managerial authority that accepts the consequences of priority changes. Human availability cannot be represented as an unlimited fungible machine-hour resource.

17.7 INVENTORY–PRODUCTION RELATIONSHIPS

17.7.1 Why inventory and spare capacity are both buffers

Inputs, work in process (WIP) and finished goods serve different operational purposes. Raw-material availability decouples production from supplier delivery; WIP can protect downstream resources against short-term interruptions; finished-goods stocks decouple customer service from manufacturing time. All inventories consume finance, storage, handling and information capacity, and some deteriorate or become obsolete. Spare capacity is an alternative or complement to stock: excess staffing or equipment availability can absorb variation without holding as much physical inventory. Time is also a buffer, since promising a longer lead time gives a system room to respond. Hopp and Spearman (2004) characterize lean in terms of controlling the cost of buffering variability, a more useful framing than declaring all inventory inherently wasteful.

A flow constraint changes the meaning of inventory. An extra ton of input material does not improve shipped output when downstream processing is already saturated; instead it can raise holding costs or spoilage. At the same time, zero input stock may stop the bottleneck during ordinary supply interruptions. The control objective is therefore to provide deliberately sized buffers at points where their marginal reliability benefit is justified. The appropriate size depends on replenishment lead time, demand variability, supplier reliability, shelf life and the financial consequences of shortage. Detailed reorder-point, economic-order-quantity and network inventory models belong to Chapter 19; this chapter owns their production-flow implications.

17.7.2 Push, pull and hybrid control

Push systems release work according to a plan, whereas *pull* systems authorize replenishment or work release based on a downstream consumption signal or a cap on WIP. These categories have variants; an actual enterprise may use planned purchases upstream, a WIP-limited production cell and order-specific packing downstream. Hopp and Spearman (2004) caution against defining pull solely by the presence of a kanban card. A genuine pull-control rule must say what quantity can enter the system, when release is permitted, and who maintains the signal's integrity. A kanban label applied to an unlimited release process does not create the expected flow discipline.

A practical work-in-process limit should be grounded in the quantity the constrained line can release without either starving downstream work or congesting the floor. For example, an operational pilot may cap released work at a defined number of batches, monitor the queue immediately before sealing, and authorize more work only when the same number has cleared the control boundary. This is a control protocol, not a universal numerical kanban prescription; the limit must be revised when lead-time variation, planned maintenance, the product mix or material availability changes. An excessive cap can generate starvation and lost demand, whereas a nominal cap that staff routinely bypass does not actually restrain flow.

In a Rwandan SME dependent on agricultural ingredients, the stability assumptions behind low-stock systems may be difficult to sustain during seasonal shortages, inconsistent input quality or delayed transport. This does not invalidate lean principles. It implies that variance reduction, supplier development, prudent buffer design and truthful forecasts precede aggressive stock reduction. Internationally benchmarked inventory turns without comparable product mix, financing costs and procurement lead times provide weak evidence. Businesses should also quantify the cost of food spoilage, discarded packaging, emergency purchases and customer service failures before claiming a reduction in inventory as a net improvement.

17.7.3 Work-in-process discipline and financial consequences

WIP is a physical queue and an accumulation of cash commitments. A growing production queue ties up purchased inputs, wages and overhead before customers pay. If the average WIP is 500 batches and output is 125 batches per day, Little's law implies an average four days in process under stable-flow assumptions. Reducing releases to align with the bottleneck, while protecting it from starvation, can shorten that time without necessarily requiring an additional machine. But finance should not presume that every reduction in

WIP becomes immediate cash: contract terms, material purchase quantities, minimum pack sizes and accounting policies matter. The operational measure must be linked to cash-conversion realities rather than translated into a fabricated financial benefit.

The practical production–inventory interface requires one authoritative material status, clear ownership of physical counts, quarantine rules for suspect goods, traceability where safety requires it, and agreement on production priorities. A shared spreadsheet may suffice for a small operation if changes are controlled and reconciled; a costly enterprise-resource-planning platform cannot repair ambiguous product codes or unreliable counting. The wider data architecture is examined in Chapter 28. Inventory accuracy and service consequences matter because an apparently efficient production schedule may be unexecutable if the recorded inputs do not exist.

17.8 PROCESS TECHNOLOGY AND AUTOMATION

17.8.1 Technology as a sociotechnical redesign, not a machine purchase

Operational technology includes processing machinery, sensors, programmable controls, industrial software, barcode identification, condition monitoring, computer vision and robotics. Technology selection must begin with the task and bottleneck to be addressed. If errors originate in unclear labels and inconsistent master data, installing a robot may increase the speed at which errors propagate. If the constraint is a legally required manual review, automating an unconstrained preparatory operation may simply expand a queue. Modern information systems can improve visibility and traceability, but the reliability of data definitions, staff judgment, maintenance and permissions determines whether such visibility improves decisions. Operational transformation is therefore sociotechnical: a change in machines also changes responsibilities, competencies, failure modes and governance (Boudreau et al., 2003; Slack et al., 2022).

The relevant alternatives range from work-method redesign and simple jigs to automation or outsourcing. A make-or-buy comparison must specify total acquisition and installation cost, required utilities, spare parts, vendor support, training, cybersecurity, consumables, downtime during installation, operational lifespan and disposal obligations. Two machines with identical nameplate throughput may have very different economic performance when one requires imported proprietary components or specialized calibration. For a small enterprise, serviceability and continuity may outweigh marginal differences in ideal speed. The supplier's benchmark under favorable test conditions should be treated as vendor evidence to be validated through acceptance trials rather than automatically as the firm's forecast.

17.8.2 Digital production systems, sensors and industrial analytics

Industrial Internet of Things (IoT) systems collect operational signals such as temperature, vibration, energy draw, count, stoppage and quality exceptions. Properly governed, these observations enable condition monitoring, traceability and earlier diagnosis of failures. Yet timestamp drift, sensor calibration, network outages and ambiguous machine states can create false precision. A dashboard showing “90% availability” is not useful until managers can reproduce the numerator and denominator from approved downtime definitions. Predictive models trained on another factory's equipment may perform poorly under different loads, raw materials and maintenance patterns. AI-assisted operations can help forecast failures or scheduling conflicts, but model recommendations require validation and accountable human override, particularly when safety or food integrity is at stake. Specialist AI lifecycle and security controls belong to Chapter 29.

An incremental implementation pathway is often more defensible than wholesale digitalization: define the operational decision, establish a clean manual baseline, identify minimum data required, pilot with a holdout or comparable historical period where feasible, measure error and unintended effects, then authorize expansion. A safety-critical automatic shutdown requires a higher evidentiary standard than a nonbinding recommendation to move a low-risk order. Adoption must include a fallback manual procedure and the capacity to maintain data and hardware over time. Otherwise, a system built to raise efficiency may become a single point of operational failure.

17.8.3 Implementation economics and fairness

Technology value should be evaluated by incremental good-output contribution, service reliability, risk reduction and lifecycle cost, not by vendor claims of “hours saved” alone. Time released from an operation creates financial benefit only if it enables valuable redeployment, avoids a genuinely variable cost, prevents losses or supports attainable sales. Savings cannot be counted twice as both increased throughput and eliminated wages if staff remain employed to deliver that throughput. Equally, labor reduction is not ethically justified simply because a system automates a task: training opportunities, consultation, safe redeployment and lawful employment obligations must be examined. Chapter 12 develops workforce learning and Chapter 11 employment rights. The operational argument is that technology acquires value through an accountable redesigned work system rather than through adoption itself.

A technology business case is especially vulnerable to double counting. Suppose a sensor system reduces unscheduled downtime and the line sells additional good packs. Incremental contribution should be calculated from verified added and collectible units, net of consumables, recurring licences, training, maintenance and the installation outage. If the model also claims a wage saving for the same operators who are needed to produce those extra packs, that saving is not independently realizable. Sensitivity should include vendor response time, false alarms, production mix and a downside case in which throughput fails to change. Approval ought to follow a pilot with documented acceptance criteria, independent safety assessment where relevant, cybersecurity responsibility and a funded fallback method. Capital valuation is treated in Chapter 24; the operational discipline here is to establish credible physical and organizational effects before applying a discount rate.

17.9 OPERATIONS PERFORMANCE MEASUREMENT

17.9.1 Measuring the system without rewarding local distortion

Performance measurement translates operating commitments into explicit evidence. Measures should be selected from the causal and ethical logic of the process, not merely from what the current system captures conveniently. The six common perspectives of output, quality, speed, dependability, flexibility and cost should be complemented by safety and resource productivity where material. An enterprise may raise its output-per-labor-hour indicator by deferring maintenance, accumulating uninspected goods, using excessive energy or requiring undocumented overtime. The apparent productivity improvement would be a measurement artifact or an unsustainable transfer of costs. The correct question is what output is counted, whether it is acceptable, which input is included, what time window applies, and what harms remain outside the ratio.

Table 17.1. Core operational metrics and interpretation controls

Measure	Definition and calculation	Essential interpretation caution
Good-output throughput	First-pass acceptable units released / period	State product mix and unit; avoid counting defective or unreleased work
Design-capacity utilization	Actual good units / theoretical design units × 100	Reveals denominator-specific loading, not overall efficiency or welfare
Schedule adherence	Committed operations completed as scheduled / scheduled operations × 100	Record frozen schedule and approved changes; cannot replace customer on-time delivery
First-pass yield	Units accepted without rework / units initially processed × 100	Define inspection gate and whether rework is counted twice
Flow time	Mean elapsed interval from agreed process entry to exit	Include queues, transfers and returns; separate process time

Measure	Definition and calculation	Essential interpretation caution
Labor productivity	Acceptable output / direct and relevant indirect labor hours	Correct for quality, case mix, outsourced labor and unpaid effort
Availability	Actual running time / planned production time × 100	Specify planned stops and classify downtime consistently
On-time-in-full (OTIF)	Orders delivered on promised date and quantity / eligible orders × 100	Use original customer commitment, not revised promise alone
Energy intensity	Metered energy use / good units produced	Control product mix, load, source and waste streams
Safety performance	Exposure-based incidents plus preventive leading indicators	Low reported incidents can reflect weak reporting rather than low risk
Defect fraction	First-pass rejected units / total units initially processed × 100	Define acceptance gate, yield, rework and defect categories consistently
Upper-tail service waiting	90th or 95th percentile of arrival-to-service delay	Declare case mix, inclusion of abandonment and queue priority

Note. This table is an original measurement instrument developed from operations-management concepts (Slack et al., 2022), flow reasoning (Little, 1961), maintenance evidence (McKone et al., 2001), and safety principles (ILO, 2022). Indicators require documented definitions and organizational validation; no universal benchmark is proposed.

17.9.2 A balanced operational scorecard and metric governance

A robust scorecard uses a manageable hierarchy: strategic promises, operational outcome indicators, process diagnostic indicators and leading safeguards. The strategic layer might specify reliable food supply and sustainable cost. Outcome measures can include first-pass good packs, original-promise OTIF, customer complaints, cash-constrained contribution, and harm indicators. Diagnostic measures can then track bottleneck availability, sanitation downtime, label errors, input shortages and changeover duration. Leading controls include preventive-maintenance completion, operator training, calibration status and unresolved hazards. Each measure needs an owner, source, unit, data lineage, review frequency, threshold and authorized response. An operational KPI without a defined action may be informational but not yet a control.

Measurement also creates behavioral incentives. If rewards are attached solely to good output, workers may conceal defects, narrow product mix or postpone proper cleaning. If call-center staff are evaluated only on average handling time, complex customers can be disadvantaged. Controls should combine result measures with process checks, worker voice, exception audits and independent customer evidence. Goodhart-like distortion becomes more likely when targets are treated as exact representations of complex organizational purposes. A defensible evaluation therefore considers whether improvement remains after adjusting for mix, demand, quality and exposure and whether persons affected by the process can challenge the data.

17.9.3 Diagnosing variation and establishing causal improvement

Performance varies because of random events, systematic changes and measurement errors. A manager observing lower defects after a new machine does not automatically know whether the machine caused the improvement: product mix, training, supplier inputs or inspection sensitivity may also have changed. The evaluation should specify a baseline period, comparable output unit, intervention date, exposure denominator and competing explanations. Where feasible, staggered rollout or a matched comparison line is preferable to an uncontrolled before-and-after claim. Process-control principles can help detect unusual variation, but detailed statistical quality control is developed in Chapter 20 and causal inference in Chapter 4. Here the managerial contribution is to ensure that improvement claims retain a counterfactual interpretation and that uncertainty is not hidden behind attractive charts.

An operational evaluation can be specified as a small decision protocol. Define the target estimand as the change in first-pass acceptable output per scheduled production hour attributable to the proposed maintenance procedure, not the change in nominal machine speed. Fix the outcome definition before the pilot; log daily input quality, product mix, staffing, downtime and rejected packs; and compare the treated line with a credible unaffected line or with repeated pre-intervention observations if comparison capacity is unavailable. Weekly control charts may reveal instability, but their use does not automatically establish a causal effect. A reduction in defects following a new check could simply reflect fewer batches inspected. An improvement claim is stronger when independent inspection coverage, customer returns and downtime support the same interpretation.

17.10 SERVICE OPERATIONS AND CUSTOMER EXPERIENCE

17.10.1 Customer participation and service variability

Service operations differ from many manufacturing systems because the customer may be present, provide essential information, contribute work and experience waiting or treatment as part of the output. A customer arriving late, bringing incomplete documents or requiring an interpretation changes the work requirement. Frei (2006) identifies multiple sources of customer-introduced variability, including arrival, request, capability, effort and preference. These sources should not be treated automatically as faults to be eliminated. A public institution must remain accessible to people with different capacities; a hospital must allow complex cases; a university cannot grade its service purely by the number of students processed per hour. The design question is how to absorb legitimate variation through appointments, triage, flexible staffing, accessible information and differentiated service pathways without unfair exclusion.

Services have both *frontstage* and *backstage* activities. A customer sees reception, communication and resolution, but the outcome depends on internal records, verification, approvals, payment and specialist support. A service blueprint can make these layers and their handoffs visible. If a bank changes its front-counter staffing while back-office authorization remains saturated, observed queue movement may simply shift delay to a less visible stage. Digital self-service can reduce repetitive work for digitally capable clients but transfer burdens to customers who lack access, language support or reliable connectivity. Service design must therefore count failed attempts, abandoned requests, accessibility barriers and recovery time alongside completed transactions.

17.10.2 Waiting, priority and service recovery

Service capacity matching requires more than averaging customers per employee. Demand peaks, appointment no-shows, service-time variation, staff breaks and urgent exceptions determine waiting experience. The queueing example in Section 17.4 illustrates why an operation can have nominal capacity above average arrivals while still producing unacceptable delays. Alternatives include adding temporary capacity at predictable peaks, reducing administrative steps, separating simple from complex work, managing appointments or improving information before service begins. Each alternative has distributional effects. An exclusive fast queue may raise revenue while disadvantaging urgent or low-income users; prioritized triage is appropriate when it reflects legitimate need rather than arbitrary purchasing power.

Service recovery is the operational response when delivery fails: acknowledge the problem, protect safety, determine an appropriate remedy, correct records and learn from root causes. The promise should be proportionate and enforceable. A manager who automatically offers replacements without verifying inventory may increase customer disappointment; one who measures complaint closure by marking tickets resolved before remedy may manufacture apparent service quality. Customer-experience evidence should be linked to operational events such as lost orders, delayed authorizations and wrong documentation. Chapter 14 explains customer research, while Chapter 16 addresses communication and channel promises; this chapter concerns the actual reliability and recovery architecture that those promises require.

17.10.3 Designing equitable service productivity

Frei's (2006) service-operations analysis is more directly relevant: customer-introduced variability can sometimes be accommodated through better process architecture rather than suppressed by making services inaccessible. Such changes may increase productivity and access simultaneously, but their effects depend on which variation is legitimate, which process steps are standardized and which risks shift to customers.

The distinction between operational equality and equitable service is particularly important for public-facing and essential services. Identical processing-time targets for every user may appear impartial while penalizing clients who need interpretation, disability accommodation, complex verification or confidential support. Managers can separate standardized administrative steps from legitimately variable professional judgment, then measure outcomes within relevant case categories. For example, an average five-minute reduction may coexist with a rising 90th-percentile wait among vulnerable users. Such a distributional comparison should be interpreted only after recording case mix and privacy-protective group sizes; it is a diagnostic requirement, not a justification for intrusive identification or disparate treatment.

17.11 RELIABILITY, MAINTENANCE AND CONTINUITY

17.11.1 Reliability as a time-dependent operational capability

A machine's capacity is not fixed by its label once it enters service. Wear, calibration drift, contamination, operator error, electrical interruption and repair delays alter productive availability and output quality. Reliability engineering studies the probability and consequences of successful operation over a specified period; maintenance organizes inspection, servicing, repair, replacement and learning. *Corrective maintenance* responds after failure; *preventive maintenance* uses planned intervals or conditions; *condition-based maintenance* acts on measured asset state; and *predictive maintenance* employs models to estimate future failure risk. Each approach is rational in some settings: an inexpensive noncritical tool may be repaired on failure, while a safety-critical seal or cold chain requires much stronger preventive controls.

An asset register should identify equipment, criticality, age, maintenance history, responsible persons, spares, support contracts and failure consequences. ISO 55001:2024 presents asset-management-system requirements that connect asset decisions with value and organizational objectives (International Organization for Standardization, 2024). Such standards support management discipline but are not, by themselves, evidence that a particular machine is safe or that the enterprise is legally compliant. McKone et al. (2001) report significant relationships between total productive maintenance practices and manufacturing outcomes in their studied settings, including indirect relationships involving just-in-time practices. Their observational structural-equation design cannot establish universal causal benefits or eliminate the need for local evaluation.

17.11.2 Maintenance indicators and criticality-based prioritization

Reliability indicators include mean time between failures (MTBF), mean time to repair (MTTR), availability, repeated fault rates and overdue preventive work. If a machine runs for 180 hours between four recorded failures in a specified period, an illustrative operating-time MTBF is 180 / 4 = **45 hours**. If total corrective downtime is ten hours, MTTR is 10 / 4 = **2.5 hours**, and a simple steady-state availability approximation is MTBF / (MTBF + MTTR) = 45 / 47.5 = **94.74%**. The computation assumes comparable failure episodes and excludes planned downtime; it is not equivalent to line OEE or total calendar availability. Definitions must be standardized before comparing assets or periods. Failure probabilities can differ by operating regime, so averages are diagnostic summaries rather than permanent physical constants.

A useful prioritization begins with the consequence of failure, not only its frequency. An infrequently failing critical safety valve may deserve stronger maintenance attention than a frequently failing but noncritical office device. A risk register should identify safety, service, financial, environmental and regulatory consequences, current controls, detection options, spare-part availability and acceptable restoration times. Redundancy,

condition monitoring and preventive inspection each have costs and failure modes. Operators should have clear authority to stop unsafe equipment without fear of retaliation, while managers distinguish genuine safety stops from avoidable breakdowns through proper review rather than suspicion of staff motives.

17.11.3 Continuity, disruption and recovery decisions

Operational resilience is the ability to maintain or restore essential processes within tolerable harm, time and cost when disturbances occur. This requires mapping dependencies on electricity, water, transportation, digital access, critical people, materials and regulatory permissions. A continuity plan prioritizes outputs that must be preserved, specifies safe shutdown and restart procedures, establishes alternative resources and rehearses communication. A generator is not a complete continuity plan if fuel, maintenance, ventilation or changeover safety is unaddressed. Outsourcing a critical function can diversify capacity yet introduce new supplier dependence. Because detailed enterprise-wide risk management belongs to Chapter 32, the present concern is process-level recovery: minimum safe operating state, restoration sequence, validation of product quality and accountability for lessons learned.

Continuity should be expressed in operationally testable terms. Recovery time objective (RTO) is the target duration for restoring a defined essential process, while recovery point objective (RPO) expresses the maximum tolerable interval of data loss in an information-dependent process. A bakery may require validated temperature records and retained batch identities before releasing product after a power interruption; restoring electricity alone is not recovery. For a digital service counter, an apparently functional screen may still conceal missing transactions requiring reconciliation. These objectives are internal planning assumptions that must be justified against customer, safety and legal requirements, not copied from generic templates.

A meaningful rehearsal should test one plausible disruption pathway end to end. Managers would simulate a sealing-machine failure during a peak order period, identify safe shutdown and product quarantine, confirm a spare-part source, trace the effect on the next delivery commitment and authorize restart only after required checks. The exercise should document actual restoration time, missing information, role ambiguity and the cost of temporary alternatives. A business unable to finance full geographic redundancy may nevertheless improve resilience through reliable maintenance records, cross-training, calibrated spare parts and a candid customer notification protocol. The resulting preparedness is demonstrated by tested recovery capability, not by the mere existence of a continuity document.

17.12 RESPONSIBLE AND SUSTAINABLE OPERATIONS

17.12.1 People, prevention and legitimate operational limits

Operational decisions redistribute exposure to physical, psychosocial and financial risks. Managers decide line speed, staffing, shift pattern, guarding, cleaning time, training access and whether a near miss is investigated. A process that reaches target throughput only when workers disregard safe operating procedures is not reliably capable at that target. The International Labour Organization (ILO, 2022) recognized a safe and healthy working environment as a fundamental principle and right at work. ISO 45001:2018 offers a management-system framework including leadership, worker participation, hazard assessment and improvement; the standard's existence does not replace national legal verification, competent technical assessment or actual prevention. The appropriate operational measure is output achieved *within* nonnegotiable safety and rights constraints.

Human dignity matters even when a worker's exposure is not captured in an injury count. Fatigue, repetitive strain, coercive overtime, exclusion from reporting channels and opaque algorithmic assignments may undermine productivity and lawful work over time. Worker consultation is operationally valuable because frontline knowledge exposes rare failures and workarounds not visible in dashboards, but employee voice is also a legitimate interest independently of financial returns. Chapter 11 develops the enforceable employment relationship, while this chapter requires every production design to make safe work an explicit feasibility condition rather than an afterthought.

17.12.2 Resource productivity and the limits of green labels

Sustainable operations evaluates energy, water, material losses, emissions and disposal across a meaningful system boundary, in addition to economic output and human consequences. Kleindorfer et al. (2005) establish the importance of linking environmental, health and safety considerations with operational design rather than treating them as external reporting issues. ISO 50001:2018 provides an energy-management-system framework; it does not certify that every product made by a participating organization is environmentally benign. A facility reporting lower kilowatt-hours per good pack may still consume more total energy if volume increases, or may have shifted energy-intensive steps to suppliers. Both intensity and absolute impacts should therefore be reported where relevant, with clear scope and method.

A numerical intensity measure makes the boundary problem explicit. If the constructed processor consumes 1,200 kWh of metered electricity to produce 36,000 first-pass acceptable packs during a month, recorded intensity equals 0.0333 kWh per pack. If output rises to 38,000 packs while electricity increases to 1,250 kWh, intensity falls to approximately 0.0329 kWh per pack even though total use rises by 50 kWh. Both results can be true and neither establishes a life-cycle emissions reduction without the energy-source mix, displaced processes and relevant upstream impacts. Responsible reporting should therefore disclose absolute resource flows, intensity ratios, system boundaries and credible uncertainty rather than selecting whichever metric appears more favorable.

ISO 14040:2006 sets out the goal-and-scope, inventory, impact-assessment and interpretation structure for life-cycle assessment (International Organization for Standardization, 2006). Its application requires specifying a functional unit, inventory boundaries, allocation choices and uncertainty rather than calculating only the factory's electricity bill. Even a narrowly defined cradle-to-gate study can be informative if its limitations are made explicit; it should not be portrayed as a complete cradle-to-grave account. The comparative question is whether a proposed operational change reduces the relevant burden for a functionally equivalent product, rather than merely moving emissions or waste into the supplier's activities.

Material-yield improvement can reduce both costs and waste. Suppose a hypothetical processor receives 1,000 kilograms of inputs and obtains 820 kilograms of acceptable product. Material yield is 82%, while the 180 kilograms difference includes moisture changes, trims, rejected material and other losses that must be identified separately. Describing all 180 kilograms as avoidable waste would be incorrect without a material balance and process analysis. Likewise, recycling a material does not automatically establish net environmental improvement if collection and reprocessing generate substantial burdens. A cradle-to-gate or life-cycle assessment, where feasible, offers a stronger basis for claims than an isolated packaging label. Detailed circular-economy frameworks are developed in Chapter 31.

17.12.3 Rwanda-specific resource and safety implications

The NISR 2024 Integrated Business Enterprise Survey reports limited solar use, at 7.4% of formal and 6.0% of informal businesses, according to its published summary (NISR, 2026a). These proportions are descriptive for the survey's definitions, not evidence that every business without solar is financially inefficient or unwilling to adopt renewable energy. Economic suitability depends on tariffs, load shape, capital access, equipment support and alternative energy reliability. The same survey reports different patterns of paid waste disposal between formal (78.1%) and informal (51.7%) businesses. These figures should not be interpreted as full compliance rates or direct estimates of ecological harm; they identify a measurement and service-access question requiring closer analysis.

Food manufacturing demands attention to sanitation, ingredient traceability, contamination control and response to suspected unsafe batches. Rwanda Standards Board (n.d.) describes RS 184 as a HACCP-based food-safety system standard that applies to relevant food-chain organizations. The description supports the relevance of hazard-based design, but it is not sufficient to assert that a hypothetical facility holds certification or meets all Rwanda FDA requirements. The World Health Organization's five keys to safer food provide general prevention guidance (World Health Organization, n.d.), while the precise legal and technical obligations of an enterprise require current source verification in its subsector. Operations managers should

therefore maintain documented control points, verification records, withdrawal/recall procedures and clear duties for stopping distribution when safety is uncertain, without overstating what a generic operating checklist can guarantee.

Table 17.2. Responsible operations decision test

Decision dimension	Evidence required before authorization	Stop or revise condition
Customer and mission	Demand segment, promised delivery, demonstrated use and customer consequences	Plan depends on deceptive claims or unserviceable commitments
Process feasibility	Flow map, bottleneck capacity, input and labor availability, quality gate	Required output exceeds safe feasible good-output capacity
Economic viability	Incremental contribution, total costs, cash exposure and sensitivity	Decision fails plausible downside or creates unfinanceable obligations
People and safety	Hazard assessment, competent staffing, consultation, lawful work conditions	Uncontrolled serious hazard or coercive labor requirement
Environmental stewardship	Energy, water, material flows, waste destination and relevant compliance	Material harm is unmeasured or knowingly shifted outside boundary
Evidence and accountability	Baseline, owner, data provenance, control thresholds and review date	No credible mechanism to detect failure, correct records or remedy harm

Note. Original integrative decision test based on the chapter's theory and reviewed standards (ILO, 2022; International Organization for Standardization, 2018a, 2018b; Kleindorfer et al., 2005). The table proposes a governance protocol; it is not a legally authoritative test or empirically validated scale.

17.13 RWANDA-BASED APPLICATION AND INTERNATIONAL COMPARISON

17.13.1 What official national evidence does and does not show

Rwanda's business-enterprise structure is an essential boundary condition for operations theory. NISR's Integrated Business Enterprise Survey 2024 describes a business landscape with at least 87% informality in the published summary, and reports that **62.6% of manufacturing companies did not use their full capacity** in the survey reference year (NISR, 2026a). The summary identifies weak demand and input-supply reliability among reported reasons for underutilization. A firm may therefore remain idle because its market cannot absorb output, not because equipment is technically deficient. An indiscriminate policy of capital subsidies or additional machinery would fail to address demand, market access or supplier consistency in some cases. The percentage is a descriptive survey result; it does not measure idle machine-hours per industry, prove that a specific firm is mismanaged, or estimate causal effects of government policy.

In the second quarter of 2026, NISR reported real GDP growth of 9.4% year on year, industry growth of 18% and manufacturing growth of 10%; within manufacturing, food processing grew by 2% in the published activity breakdown (NISR, 2026b). These measures have different economic scope and must not be conflated. Sectoral output growth does not establish capacity utilization, demand for one brand, unit profitability or the reliability of a named factory. A prudent Rwandan operations analyst joins national context to enterprise-specific evidence by specifying products, orders, seasons, process constraints and financial terms. Growth statistics supply a context for inquiry rather than an operational forecast. Likewise, a national energy or labor indicator should not replace on-site meters, contractual power data or actual skills audits.

Comparative small-firm scholarship helps interpret what these constraints mean without extrapolating from national aggregates. McKenzie and Woodruff (2017) examined managerial practices such as stock-keeping,

recording transactions, marketing and financial planning across micro and small businesses in seven developing countries, including Ghana, Kenya and Nigeria. They found associations between better practices and firm outcomes, but their observational estimates are not randomized treatment effects and the sampled countries do not include Rwanda. Read alongside the Indian textile experiment of Bloom et al. (2013), this evidence suggests a defensible but testable priority: establish reliable everyday operating records before assuming that sophisticated capital equipment will resolve an enterprise's main constraint. The policy relevance lies in comparing mechanisms and constraints, not importing a universal productivity coefficient or treating all African enterprises as equivalent.

The enterprise survey's sampling frame also constrains comparison. NISR's IBES covers defined non-agricultural business activity, and its operational classifications must not be equated with a census of all farms or all unregistered productive work (NISR, 2026a). The reported 62.6% refers to manufacturing enterprises indicating incomplete utilization, not an estimate that 62.6% of Rwanda's manufacturing capacity lies idle. Its denominator counts enterprises rather than homogeneous machine-hours and does not state a common amount of unused productive potential. Further research would need firm-level orders, asset hours, utilization definitions and product mix before estimating a national capacity gap or recommending equipment investment.

The United Nations Industrial Development Organization (UNIDO, 2024) places industrial capabilities, skills, innovation and resource efficiency within a wider development agenda. This supports examining operational productivity as an institutional and technological capability rather than a problem of worker effort alone. However, international policy narratives must still be tested against local transaction costs, market size, financing and implementation capacity. A small food processor may achieve more from stable supplier specifications, safe cleaning routines, better demand records and an achievable maintenance plan than from expensive autonomous production technology. The ranking of interventions depends on the particular bottleneck and evidence, not a presumption that simple practices are always sufficient.

17.13.2 Hypothetical Kigali food enterprise: operational diagnosis

The constructed food enterprise has one eight-hour production shift and 22 nominal workdays per month. Its preparation, cooking and sealing standards are respectively 2.5, 3 and 4 minutes per 20-pack batch. The bottleneck is sealing at 120 batches/day before losses. Baseline good-output capacity is approximately 41,970 packs/month under the separately stated availability, performance and quality factors. Actual first-pass good output is 36,000 packs, and plausible demand is 38,000 packs. Because the modeled capacity exceeds demand while observed output falls short, the case cannot justify buying capacity merely from order counts. A first managerial intervention is to reconcile daily time losses, sealing stoppages, input shortages, planned versus actual production time and rejected packs, while obtaining evidence that orders are genuine, collectible and legally serviceable.

A disciplined thirty-day pilot would begin by approving one consistent unit definition and recording hourly starts, good completions, scrap, duration of stoppages, reasons for interruption, staff hours, energy use and missed orders. The production supervisor and worker representatives should review the first week for incorrect definitions; a maintenance specialist should test whether seal temperature instability or delayed changeovers accounts for recurrent interruptions. The team should separately examine market feasibility and packaging supply, because improved cycle speed cannot correct absent demand or materials. The intervention's first test is whether the process can meet the known 38,000-pack requirement safely and dependably with existing assets. Only after that result is reproducibly demonstrated should management consider the RWF 6 million equipment modification modeled earlier.

The case's financial and human dimensions must remain linked. Delivering 2,000 extra packs at RWF 650 contribution each yields RWF 1.3 million before incremental support cost, but this depends on saleable demand and cash collection. If output improvement requires unpaid overtime or greater contamination risk, the decision has failed the responsible feasibility test even if contribution rises. If reducing late deliveries also lowers customer complaints and emergency transport, those may be additional benefits, but they require

independent measurement and cannot be invented into the calculation. A credible post-pilot review compares observed output and complaints to the pre-intervention baseline, documents demand and material conditions, and treats uncontrolled coincident changes as alternative explanations.

17.13.3 Hypothetical Kigali service operation: a different form of constraint

An illustrative walk-in service center has two employees, each taking twelve minutes per eligible transaction on average. Under the deliberately stylized M/M/2 assumptions in Section 17.4, an arrival rate of eight customers per hour produces predicted mean queue waiting of about 21.33 minutes despite nominal aggregate capacity of ten customers per hour. A manager relying solely on 80% occupancy may overlook that waiting burden. Capacity is not the only remedy: better appointment communication, separation of simple and complex services, clear documentation before arrival, and suitable peak staffing may reduce variance or actual work. These measures must be tested locally rather than selected from a simplified queue model alone.

If real service time has a heavy right tail or customers have priority needs, an M/M/2 model may underestimate or mischaracterize delay. Its assumptions are useful precisely because they are visible and falsifiable. A service manager should record arrival times, service durations, abandonments, staffing, peak patterns and fairness-relevant categories with proportionate privacy safeguards. The service case also illustrates why total productivity cannot be reduced to quick closure: an accurate, respectful and accessible resolution may legitimately require more time. The contrast with food production shows that a common flow vocabulary can illuminate different operations without forcing the same technical intervention.

17.13.4 Documented Toyota production-system comparator

Toyota Motor Corporation describes its Toyota Production System as resting on two pillars, *jidoka*, the immediate detection and response to abnormality, and *just-in-time*, synchronizing production with what is needed when needed (Toyota Motor Corporation, n.d.). Its official account also emphasizes worker involvement in continuous improvement. Spear and Bowen (1999), analyzing the system as a structure of explicit work specifications, direct connections, clear pathways and experimental improvement, interpret its effectiveness as organizational learning rather than the mere possession of kanban cards. This provides a documented international comparison. The evidence does not justify attributing a particular unverified financial return to Toyota or claiming its arrangements can be copied without adaptation.

For the hypothetical Kigali processor, the transferable principle is to make abnormalities observable and authorize competent workers to stop, diagnose and correct unsafe or defective flow. The less readily transferable feature is the assumption of reliably synchronized replenishment, high-quality supplier communication, sophisticated technical support and stable infrastructure. A low-stock arrangement that depends on punctual deliveries may amplify rather than reduce risk where transport and input reliability are variable. The Toyota comparison therefore supports learning from underlying mechanisms while rejecting superficial imitation. It also underscores a tension: “lean” practices can strengthen worker problem-solving when authority and training are real, but intensify work or conceal vulnerability when implemented solely as labor compression.

The comparison must also address the employment side of process design. Toyota presents jidoka and kaizen as involving human judgment and abnormality response, whereas the independent review of lean work organization by Landsbergis et al. (1999) describes evidence of intensified demands in some automotive settings. These are not logically inconsistent observations: worker authority can be genuine in one design and merely rhetorical in another. The transferable criterion is whether frontline employees have the time, training and protected discretion to signal defects and stop unsafe flow, and whether managers repair the process rather than punish the messenger. Cost or inventory statistics alone cannot demonstrate that this criterion has been met.

Table 17.3. Context-sensitive comparison of the illustrative enterprise and documented Toyota principles

Analytical aspect	Constructed Kigali operation	Documented Toyota production principles	Transferability test
Process visibility	Simple daily stoppage and first-pass yield log	Explicit abnormality detection and visible process signals	Can workers record and act on accurate exceptions?
Flow control	Bottleneck-limited release with selected material buffers	Synchronization of production to need, just-in-time	Are inputs sufficiently reliable to reduce stock safely?
Defect prevention	Seal integrity checks and stop authority	Jidoka and immediate correction of abnormality	Is stop authority supported by skills and management?
Improvement discipline	Thirty-day pilot with documented outcomes	Repeated hypothesis testing and continuous improvement	Are baselines, resources and review time available?
Governance context	SME finance, skills and infrastructure constraints	Firm-specific accumulated processes and supplier network	Which capabilities must be built before imitation?

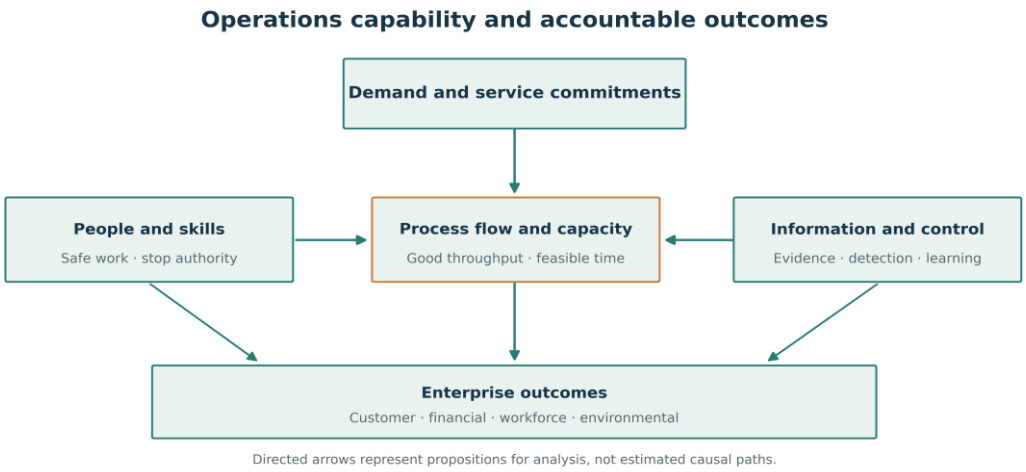
Note. The Kigali case is hypothetical; Toyota's principles are drawn from Toyota Motor Corporation (n.d.) and critically interpreted using Spear and Bowen (1999). The comparison does not report measured performance differences.

17.14 INTEGRATIVE DISCUSSION, CRITICAL SYNTHESIS AND RESEARCH AGENDA

17.14.1 An accountable operations-capability synthesis

The preceding analysis suggests an operations system should be understood as an adaptive configuration of *demand commitments*, *flow and capacity*, *resource and labor capability*, *information and control*, and *institutional and ecological constraints*. Outcomes emerge through these interfaces rather than as a simple sum of individual machine efficiencies. Marketing promises determine required service, but production defines what can be delivered; procurement contracts affect input consistency, while inventory buffers absorb some variability; HR practices determine whether staff possess the capabilities and authority to detect abnormality; finance defines the feasible capital and working-capital envelope. The resulting synthesis is conceptually compatible with the broader enterprise-systems perspective established in Chapter 1, but its suggested causal links remain propositions for empirical inquiry, not an independently validated model.

Figure 17.3. Proposed operations-accountability synthesis



Note. Original conceptual synthesis drawing on the operations strategy, flow, human resources and sustainable operations literature discussed in Sections 17.0–17.14. Relationships are analytical propositions, not statistically validated paths.

This framework has an important epistemological implication. Operations indicators are not neutral representations of an object independent of management; they depend on boundaries, counting rules and intervention choices. A "capacity deficit" can be a false diagnosis caused by demand volatility, missing skills, erroneous product classifications, unrecorded downtime or socially necessary safety controls. A "productivity gain" can arise from reclassifying defects or moving unpaid labor outside the accounting boundary. Consequently, diagnostic accountability must accompany managerial authority. Operations should retain the possibility that its favorite model, process map or improvement technique has been falsified by evidence, and that a legitimate stakeholder constraint changes the admissible objective.

17.14.2 Cross-functional consequences and irreducible trade-offs

First, the link to marketing and commercial execution is reciprocal: sales cannot make dependable promises without credible available-to-promise capacity, while operations cannot optimize its product mix without defensible demand signals and customer-value priorities (Chapters 13–16). Second, the link to procurement and logistics is temporal: supplier reliability affects buffers and schedules, while a production release plan changes warehousing and delivery requirements (Chapters 18–19). Third, the link to employees and finance is simultaneously causal and normative: staffing, safety, competence and maintenance alter good-output capacity, while investment, pricing and cost allocation determine which improvements are economically sustainable (Chapters 9–12 and 22–24). Digitization can support all these exchanges but cannot substitute for agreed definitions and decision rights (Chapters 28–29).

Some apparent trade-offs are artifacts of badly designed processes and can be reduced through better coordination. Fewer defects may improve speed and material yield at the same time; faster information exchange may reduce both waiting and unnecessary inventory. Other trade-offs persist because the system is truly constrained: a small enterprise may be unable to finance all desired resilience measures immediately, and a public service cannot simultaneously guarantee instant service for everyone with fixed specialist staff. Critical scholarship should resist both fatalistic "everything is a trade-off" reasoning and promotional "every efficiency measure benefits everyone" claims. Operations decisions are defensible when they make boundary conditions, winners and losers, legal limits and uncertainty explicit.

17.14.3 Testable propositions and empirical agenda

Proposition 17.1 (constraint-focused investment): After controlling for demand, product mix, baseline capacity and financing conditions, investments directed at a verified system constraint will be associated with larger improvements in good-output throughput than equal-cost investments at unconstrained resources. The proposition requires longitudinal process data and explicit treatment of managerial selection into investments; simple comparisons of "lean" adopters and nonadopters would be confounded.

Proposition 17.2 (variability and reserve capacity): In service operations with comparable variability and service standards, lower planned resource utilization, or equivalent variability-reduction mechanisms, will be associated with disproportionately shorter waiting times near saturation. A credible test must measure actual arrival and service processes, peak congestion and customer abandonment rather than apply heavy-traffic formulas without checking their assumptions.

Proposition 17.3 (worker voice and abnormality response): The relationship between operational standardization and quality will be moderated by worker competence, authorized stop-and-correct procedures and protection against retaliation. Research could use multi-site longitudinal interventions and independent safety and quality measures, recognizing that better incident reporting may initially raise recorded near misses.

Proposition 17.4 (context and lean transfer): The effect of low-inventory flow controls on service reliability will depend on supplier reliability, transport variability, financing conditions and product

perishability. This invites context-rich African manufacturing and service studies that directly measure these moderators rather than treat formalization or geography as sufficient proxies.

Proposition 17.5 (sustainable output): Improvements in good-output yield will be more robust predictors of environmentally adjusted operational performance when absolute energy, water and material impacts are measured alongside intensity ratios. Tests need system boundaries that detect outsourced impacts and avoid equating unit-cost savings with ecological benefit.

A feasible research programme would combine official secondary data with well-documented enterprise-level operating records where lawful access permits, and might use matched longitudinal comparisons or staggered process interventions. Systematic reviews should separately classify field experiments, observational studies, analytical queueing models and practitioner case descriptions. Data gaps in informal enterprises, service organizations and African regional supply networks remain consequential. Without comparable measurement of quality, time, work and environmental burden, cross-country rankings of operations excellence may reward reporting capacity rather than underlying performance.

17.14.4 Chapter synthesis and conclusion

Operations is the function through which enterprise commitments confront physical, temporal, human and institutional limits. Its fundamental concepts are durable: every operation transforms inputs, every flow has constraints, demand is uncertain, variability needs buffering, assets need maintenance, and measures need well-defined denominators. The technical means of addressing these problems change through automation, analytics and new operating models, but the underlying need for judgment does not disappear. In Rwanda, where the national survey documents widespread informality and substantial reported manufacturing capacity underutilization, a blanket prescription for more machinery or leaner inventory cannot substitute for a diagnosis of demand, input reliability, skills, finance and actual process variation.

The most important educational conclusion is that local optimization and enterprise value are different objects. A fully occupied machine can generate excessive queues; a faster preparatory station can increase unfinished work; a low inventory ratio can conceal inadequate protection against predictable interruptions; a high output count can coexist with unsafe labor or defective goods. Effective operations management therefore coordinates reliable customer benefit, genuinely saleable throughput, stable financial performance, human protection and ecological responsibility through explicit decision rights, sound data and revisable hypotheses. It prepares the conceptual foundation for procurement (Chapter 18), supply-chain and inventory systems (Chapter 19), quality management (Chapter 20) and project delivery (Chapter 21), while retaining its own distinct ownership of process design, productive capacity and operational control.

Seven connected insights follow from the analysis. Operations performance begins with a defensible definition of the good or service delivered; effective capacity depends on the process constraint rather than isolated station efficiency; forecasting must be separated from the financial and technical ability to meet demand; variable flows need suitable buffers and capacity reserves; automation creates value only through a validated sociotechnical redesign; performance indicators need explicit counterfactuals, rights and environmental boundaries; and institutional context determines whether apparently successful international practices can be replicated responsibly. These conclusions are conceptual guides for disciplined inquiry, not universal prescriptions independent of enterprise circumstances.

17.15 KEY TERMINOLOGY AND PROFESSIONAL INSTRUMENTS

Table 17.4. Essential terminology and distinctions

Term	Operational meaning and necessary qualification
Bottleneck	Resource or policy constraining system throughput under specified conditions; can move after improvement

Term	Operational meaning and necessary qualification
Capacity	Feasible maximum output over a declared period; distinguish design, effective and observed output
Cycle/output interval	Time between completed units at a station; distinguish from an individual unit's flow time
First-pass yield	Acceptable units without rework divided by initial units at a defined control gate
Flow time	Elapsed time a unit spends within an explicitly bounded process
Just-in-time	Synchronization of production or delivery to requirement; not a guarantee that zero inventory is prudent
Little's law	Long-run identity $L = \lambda W$ for a stable flow with matched boundaries and units
OEE	Availability × performance × quality under documented and consistent operational definitions
Pull control	Work release constrained by downstream demand or WIP authorization, not merely a visual label
Queue	Accumulation awaiting service or processing; highly sensitive to variability near saturation
Service recovery	Authorized diagnosis and appropriate remedy for service failure, including corrective learning
Throughput	Rate of completed flow units; specify whether gross or first-pass acceptable output
WIP	Units that have entered but not completed the defined productive process
Effective capacity	Practical feasible output under the specified calendar, planned losses, staffing, mix and controls
S&OP	Sales and operations planning: documented reconciliation of expected demand with attainable supply, finance and commitments
Takt time	Available production time divided by required output; demand-paced reference, not a measured station processing time
RTO	Recovery time objective: approved target duration for restoring an essential operational process

Note. Terminology synthesizes Little (1961), Hopp and Spearman (2004), Slack et al. (2022) and the operational definitions explicitly adopted in this chapter.

An implementable operations decision record should contain the decision date, problem statement, product or service unit, boundary, baseline, supporting evidence, alternative explanations, assumptions, nonnegotiable legal and safety requirements, options and full costs, threshold for success, accountable approver, pilot period, stop conditions and review date. For each formula, the record must identify the period, data owner, unit, denominator and rounding convention. For each external claim, it should preserve the issuing source and observation period; for a hypothetical case, it should retain the label “illustrative.” This

modest instrument is intentionally usable by a small enterprise with basic records, although larger operations may implement the same principles through integrated information systems.

17.16 LEARNING, APPLICATION AND EVALUATION PACKAGE

17.16.1 Foundational and conceptual assessment

Question 1 (LO1–LO2). Distinguish operations management from production management and show, using one manufacturing and one service example, why the appropriate process boundary changes the meaning of productivity. Explain why high utilization is not sufficient evidence of effective operations.

Question 2 (LO1, LO5). Compare Skinner's manufacturing-strategy argument, the cumulative-capabilities interpretation of Ferdows and De Meyer, the theory of constraints and Hopp and Spearman's account of pull. Identify the assumptions and strongest criticisms of each approach, rather than merely listing their prescriptions.

Question 3 (LO3). A firm forecasts monthly product orders using a three-period moving average. State how the method handles trend, structural shocks and stockout-censored historical sales; propose a defensible evaluation protocol and explain why forecast error cannot be inferred from one attractive month.

17.16.2 Numerical and analytical exercises

Exercise 1: Bottleneck and capacity (LO2, LO4, LO8). A process has three serial stations taking 5, 3 and 4 minutes per batch. Work is scheduled for 450 available minutes daily, and 12 packs are produced per batch. First-pass yield is 95%, with no other losses. Calculate nominal station capacities in batches/day, theoretical bottleneck good-output capacity in packs/day, and expected daily good units under the stated simplified assumptions. Then assess whether spending on the second station would raise end-to-end throughput and list data needed before approval.

Exercise 2: Forecast and service (LO3–LO4). Recorded weekly customer orders are 75, 80, 70 and 95 units. Calculate a three-week moving-average forecast for week 5 and a naïve one-step forecast's MAE for weeks 2–4. For an M/M/1 system with arrivals $\lambda = 4$/hour and service rate $\mu = 5$/hour, calculate utilization, average waiting in queue and average time in system; explain why these values should not be used as unconditional predictions for a real public service.

Exercise 3: Location and responsibility (LO5, LO6, LO8). Two candidate facilities have costs $C1 = 18{,}000{,}000 + 550Q$ and $C2 = 24{,}000{,}000 + 430Q$ (RWF/year). Identify the crossover volume and recommend a location for 40,000 annual units while making explicit three nonfinancial constraints that could reverse the recommendation.

17.16.3 Case-based critical and integrative assignments

Case assignment (LO2–LO8). Using the constructed Kigali processor in Section 17.13, develop a 2,000-word operations diagnostic memorandum containing a flow map, capacity reconciliation, loss register, forecast assumptions, worker and safety review, alternatives appraisal, trial design, KPI definitions and a stop-or-scale decision. Your conclusion must separate physically available capacity, expected additional sales, net contribution, rights and environmental requirements. Include at least one plausible explanation that would falsify your preferred intervention.

Advanced seminar (LO6–LO8). Compare the claims made by Toyota Motor Corporation with the interpretations of Spear and Bowen and Hopp and Spearman. Discuss whether a manufacturing system can be transferred across institutional settings by copying observable tools. Use Rwanda's IBES 2024 results carefully, without turning a survey percentage into a causal estimate or treating an invented firm as a national sample.

Capstone exercise (LO1–LO8). Design an operations system for a small enterprise combining an affordable packaged-food line and a direct customer-collection service. Deliver a defensible month-by-month capacity plan, simple service queue policy, facility/layout justification, maintenance and data protocols, operational scorecard, ethical safeguards and comparison with one international practice. Record all assumptions and design a monitoring approach that could reveal a mistaken forecast or unsafe workflow.

17.16.4 Educator's calculation notes and performance criteria

For Exercise 1, nominal station capacities are 450/5 = 90, 450/3 = 150 and 450/4 = 112.5 batches/day; Station 1 constrains at 90 complete batches, corresponding to 90 × 12 × 0.95 = **1,026 first-pass good packs/day** in expectation under the simplified model. A fractional capacity value is theoretical rather than a feasible partial physical batch. Adding capacity to Station 2 will not increase the stated system bottleneck rate. Students should earn more credit for identifying yield, variability, starvation, labor and planned-break assumptions than for arithmetic without interpretation.

For Exercise 2, the three-week average is (80 + 70 + 95)/3 = **81.67 units**. One-step naïve forecasts for weeks 2–4 are 75, 80 and 70, against actuals 80, 70 and 95; MAE = (5 + 10 + 25)/3 = **13.33 units**. For the stylized M/M/1 queue, $\rho = \lambda/\mu$ = **0.8**; $W = 1/(\mu-\lambda)$ = **1 hour**, $Wq = \lambda/[\mu(\mu-\lambda)]$ = **0.8 hours = 48 minutes**. These values depend strongly on exponential and steady-state assumptions; an applicant who identifies this limitation demonstrates the intended higher-order competence.

For Exercise 3, C1 = C2 at Q = 6,000,000/120 = **50,000 units**. At 40,000 units, C1 = RWF 40 million and C2 = RWF 41.2 million; Site 1 is lower modeled cost by RWF 1.2 million. A safe-water deficit, inaccessible transport, labor risk or critical utility unreliability could reverse the admissible decision. A suggested rubric weights conceptual and theoretical reasoning (25%), correct quantitative methods and interpretable units (25%), evidentiary quality and contextual transferability (20%), integrated ethical and financial judgment (20%), and clarity, reproducibility and citation discipline (10%). These weights are proposed teaching guidance, not an externally validated assessment instrument.

17.16.5 Further reading

Readers seeking deeper technical treatment should consult Hopp and Spearman (2004) for variability and pull controls; Little (1961) and Kingman (1962) for flow and delay; Hyndman and Athanasopoulos (2021) for responsible forecasting; Shah and Ward (2007) for lean measurement; McKone et al. (2001) for the limits of maintenance–performance evidence; and Kleindorfer et al. (2005) for the foundations of sustainable operations scholarship. NISR's official IBES 2024 publication and the 2026 quarterly national accounts provide contextual evidence that should be read alongside their definitions and reference periods rather than as proxies for the operating records of an individual organization.

17.17 REFERENCES

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does management matter? Evidence from India. *The Quarterly Journal of Economics, 128*(1), 1–51. https://doi.org/10.1093/qje/qjs044

Boudreau, J., Hopp, W., McClain, J. O., & Thomas, L. J. (2003). On the interface between operations and human resource management. *Manufacturing & Service Operations Management, 5*(3), 179–202. https://doi.org/10.1287/msom.5.3.179.16032

Ferdows, K., & De Meyer, A. (1990). Lasting improvements in manufacturing performance: In search of a new theory. *Journal of Operations Management, 9*(2), 168–184. https://doi.org/10.1016/0272-6963(90)90094-T

Frei, F. X. (2006). Breaking the trade-off between efficiency and service. *Harvard Business Review, 84*(11), 93–101. https://hbr.org/2006/11/breaking-the-trade-off-between-efficiency-and-service

Goldratt, E. M., & Cox, J. (1984). *The goal: Excellence in manufacturing*. North River Press.

Hayes, R. H., & Pisano, G. P. (1996). Manufacturing strategy: At the intersection of two paradigm shifts. *Production and Operations Management, 5*(1), 25–41. https://doi.org/10.1111/j.1937-5956.1996.tb00383.x

Hopp, W. J., & Spearman, M. L. (2004). To pull or not to pull: What is the question? *Manufacturing & Service Operations Management, 6*(2), 133–148. https://doi.org/10.1287/msom.1030.0028

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

International Labour Organization. (2022). *A safe and healthy working environment is a fundamental principle and right at work*. https://www.ilo.org/publications/safe-and-healthy-working-environment-fundamental-principle-and-right-work

International Organization for Standardization. (2006). *ISO 14040:2006: Environmental management—Life cycle assessment—Principles and framework*. https://www.iso.org/standard/37456.html

International Organization for Standardization. (2018a). *ISO 45001:2018: Occupational health and safety management systems—Requirements with guidance for use*. https://www.iso.org/standard/45001

International Organization for Standardization. (2018b). *ISO 50001:2018: Energy management systems—Requirements with guidance for use*. https://www.iso.org/standard/69426.html

International Organization for Standardization. (2024). *ISO 55001:2024: Asset management—Asset management system—Requirements*. https://committee.iso.org/sites/tc251/home/projects/published/iso-55001.html

Kingman, J. F. C. (1962). On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B (Methodological), 24*(2), 383–392. https://doi.org/10.1111/j.2517-6161.1962.tb00465.x

Kleindorfer, P. R., Singhal, K., & Van Wassenhove, L. N. (2005). Sustainable operations management. *Production and Operations Management, 14*(4), 482–492. https://doi.org/10.1111/j.1937-5956.2005.tb00235.x

Landsbergis, P. A., Cahill, J., & Schnall, P. (1999). The impact of lean production and related new systems of work organization on worker health. *Journal of Occupational Health Psychology, 4*(2), 108–130. https://doi.org/10.1037/1076-8998.4.2.108

Lapide, L. (2004). Sales and operations planning, part I: The process. *The Journal of Business Forecasting, 23*(3). https://ibf.org/knowledge/jbf-articles/sales-and-operations-planning-part-i-the-process-671

Little, J. D. C. (1961). A proof for the queuing formula: L = λW. *Operations Research, 9*(3), 383–387. https://doi.org/10.1287/opre.9.3.383

McKenzie, D., & Woodruff, C. (2017). Business practices in small firms in developing countries. *Management Science, 63*(9), 2967–2981. https://doi.org/10.1287/mnsc.2016.2492

McKone, K. E., Schroeder, R. G., & Cua, K. O. (2001). The impact of total productive maintenance practices on manufacturing performance. *Journal of Operations Management, 19*(1), 39–58. https://doi.org/10.1016/S0272-6963(00)00030-9

National Institute of Statistics of Rwanda. (2026a). *Integrated Business Enterprise Survey (2024)*. https://www.statistics.gov.rw/statistical-publications/business-establishment-finance-trade/business-establishment-finance-trade/integrated-business-enterprise-survey-2024

National Institute of Statistics of Rwanda. (2026b). *GDP national accounts (Q2, 2026)*. https://statistics.gov.rw/statistical-publications/gross-domestic-product/gdp-national-accounts-q2-2026

Rwanda Standards Board. (n.d.). *Training services: RS 184, Food safety system based on hazard analysis and critical control points (HACCP)*. https://www.rsb.gov.rw/standards/training-services

Shah, R., & Ward, P. T. (2007). Defining and developing measures of lean production. *Journal of Operations Management, 25*(4), 785–805. https://doi.org/10.1016/j.jom.2007.01.019

Skinner, W. (1969). Manufacturing—Missing link in corporate strategy. *Harvard Business Review, 47*(3), 136–145. https://hbr.org/1969/05/manufacturing-missing-link-in-corporate-strategy

Slack, N., Brandon-Jones, A., & Burgess, N. (2022). *Operations management* (10th ed.). Pearson.

Spear, S., & Bowen, H. K. (1999). Decoding the DNA of the Toyota Production System. *Harvard Business Review, 77*(5), 96–106. https://hbr.org/1999/09/decoding-the-dna-of-the-toyota-production-system

Toyota Motor Corporation. (n.d.). *Toyota Production System*. https://global.toyota/en/company/vision-and-philosophy/production-system/

United Nations Industrial Development Organization. (2024). *Industrial Development Report 2024: Turning challenges into sustainable solutions: The new era of industrial policy*. https://www.unido.org/idr/idr2024

World Health Organization. (n.d.). *Five keys to safer food*. https://www.who.int/activities/promoting-safe-food-handling/five-key-to-safer-food