PRACTICAL RESOURCE / AI PRODUCT MANAGER

Metrics and Experiment Plan

Product question

User task: [ ]

Hypothesis: If we [change], then [segment] will [behavior] because [evidence-based explanation].

Baseline and source: [ ]

Smallest useful comparison: [ ]

Metric contract

Field	Definition
Primary metric	[Name and reason]
Numerator	[Exact qualifying event or outcome]
Denominator	[Eligible population or task attempts]
Window	[Start, end, cohort rules]
Exclusions	[Bots, internal accounts, invalid events, etc.]
Retries	[Same task or new task, and why]
Quality guardrail	[Critical failure, correction, support, or trust measure]
Operating guardrail	[Latency, cost, or human review burden]
Segments	[Roles, languages, input types, or other meaningful groups]

Event sketch

- task_started: task ID, user/cohort ID where appropriate, timestamp, version

- output_presented: task ID, version, latency, source availability

- output_corrected: task ID, material correction flag, reason if available

- task_completed: task ID, defined success signal

- task_failed_or_abandoned: task ID, reason if observable

Collect only properties needed for the purpose. Avoid putting raw sensitive input into analytics events. Define what can and cannot be observed reliably.

Experiment design

Comparison: [randomized / staged pilot / qualitative / observational]

Assignment unit and contamination risk: [ ]

Sample-size/duration rationale: [consult analytics; do not invent a universal number]

Predefined decision thresholds and guardrails: [ ]

Owner who can pause: [ ]

Five stopping rules - complete before starting

Decision	Exact rule	Owner / response
Harm	[Critical failure that triggers immediate pause]	[Who pauses and investigates]
Data validity	[Missing events, assignment imbalance or measurement failure]	[Who checks; whether interpretation must pause]
End of learning	[Calendar deadline AND planned cases; how incomplete work is classified]	[Who closes the readout]
Continue or revise	[Useful outcome, required quality and acceptable total effort]	[Which bounded next step is justified]
Futility	[What makes more work on this version unhelpful]	[Stop, narrow or redesign; rationale]

For a randomized experiment, have an appropriate analyst specify sample size, duration and any sequential stopping approach. A convenient pilot count is not a power calculation. Repeatedly checking a conventional significance threshold and stopping when it becomes favorable can mislead.

Bounded pilot example - hypothetical, not results

Question: Can an assistant produce source-supported meeting-action drafts that a person can review efficiently?

- Inputs: 12 approved, sanitized note sets across three meeting types; no external sending or automatic assignments.

- Baseline: manually prepare the same cases; record full preparation time and quality under the same rubric.

- Success definition: every listed action, owner and date is supported; unknown information is marked unknown; reviewer approves before use.

- Planned end: two weeks or all 12 cases, whichever comes first. Fewer than 12 cases at the deadline means incomplete against the planned progression rule.

- Illustrative progression rule: at least 10 of 12 cases meet the rubric, median total time improves against the manual baseline, and no critical guardrail is breached. This is this fictional team's provisional decision threshold, not proof of uplift or general safety.

- Immediate pause: unauthorized disclosure or an unreviewed external action. Investigate before any restart.

- Next decision: a larger, newly scoped test, revision or stop. Never infer revenue or demand from this exercise.

Replace every example choice with a reasoned choice for your own setting. Keep a dated copy of the agreed plan; record later changes as amendments instead of rewriting the original.

Readout and decision record

Plan version / agreed date: [ ]

Actual dates / product or model version: [ ]

Planned cases / attempted / completed / missing: [ ]

Observed primary metric, denominator and uncertainty: [ ]

Guardrail breaches and incident references: [ ]

Review effort and full operating-cost scope: [ ]

Results by meaningful segment: [ ]

Data-quality problems / exclusions / deviations: [ ]

What the evidence supports: [ ]

What it does not establish: [ ]

Decision and accountable owner: [continue a bounded test / revise / stop / inconclusive]

Cheapest next evidence-producing step: [ ]

Next review date: [ ]

Worked arithmetic - hypothetical

1,000 attempted tasks; 700 successful; $300 defined variable operating cost across all attempts.

Task success = 700/1,000 = 70%.

Cost per successful task = $300/700 ≈ $0.43.

Cost per attempt = $300/1,000 = $0.30.

State which expenses the cost includes. A before-and-after difference alone does not establish causation. Record what you learned and the strength of the evidence.

Guidance

See the product metrics guide (https://aiproductmanager.site/blog/product-metrics-for-ai/) and Microsoft's pre-experiment (https://www.microsoft.com/en-us/research/articles/patterns-of-trustworthy-experimentation-pre-experiment-stage/) and during-experiment (https://www.microsoft.com/en-us/research/articles/patterns-of-trustworthy-experimentation-during-experiment-stage/) guidance. Updated October 19, 2026 by Ashley Kays. No signup required; no results implied.