Comprehensive Experiment Testing Guide for Marketing Analytics

A Complete Framework Using Affirm as a Case Study

---

1. Pre-Experiment Planning Phase

1.1 Business Objective Definition

Component	Options/Approaches	Affirm Example
Goal Types	• Revenue optimization • Conversion rate improvement • Customer acquisition cost reduction • Lifetime value increase • Engagement metrics • Retention improvement • Brand awareness	Increase merchant checkout conversion rate from 3.2% to 3.8%
Success Metrics	• Primary metrics (North Star) • Secondary metrics • Guardrail metrics • Counter metrics • Leading indicators • Lagging indicators	Primary: Checkout conversion Secondary: Average order value Guardrail: Customer complaints
Hypothesis Frameworks	• If-Then-Because format • Problem-Solution-Result • Jobs-to-be-Done • Lean Hypothesis • Scientific Method	"If we reduce the number of form fields from 8 to 4, then checkout conversion will increase by 20%, because users abandon due to friction"

1.2 Experiment Type Selection

Experiment Type	Use Cases	Statistical Complexity	Affirm Example
A/B Testing	• Single variable change • Clear binary choices • Simple implementation	Low	Testing two checkout button colors
A/B/n Testing	• Multiple variants • Exploring optimization curve • Finding local maxima	Medium	Testing 5 different payment plan displays
Multivariate Testing (MVT)	• Multiple variables simultaneously • Interaction effects • Complex UX changes	High	Testing button color × copy × position
Multi-Armed Bandit	• Dynamic allocation • Minimize opportunity cost • Continuous	High	Optimizing interest rate offers in real-time

Experiment Type	Use Cases	Statistical Complexity	Affirm Example
	optimization		
Factorial Design	• Full factorial (all combinations) • Fractional factorial • Systematic interaction study	Very High	Testing all combinations of 4 factors
Sequential Testing	• Early stopping rules • Adaptive designs • Resource optimization	High	Stopping test early if clear winner emerges
Switchback Testing	• Time-based allocation • Network effects • Market-level changes	Medium	Testing new merchant onboarding flow
Synthetic Control	• No randomization possible • Market-level rollouts • Causal inference	Very High	Launching in new geographic market

---

2. Statistical Framework & Power Analysis

2.1 Statistical Test Selection

Test Type	Use Case	Assumptions	Affirm Example
Z-Test	• Large samples (n>30) • Known population variance • Normal distribution	• Independence • Normality • Equal variance	Comparing conversion rates with 100k+ users
T-Test (Student's)	• Small samples • Unknown variance • Continuous outcomes	• Independence • Normality • Equal variance	Comparing average order values
Welch's T-Test	• Unequal variances • Different sample sizes	• Independence • Normality	Comparing segments with different sizes
Mann-Whitney U	• Non-parametric • Ordinal data • Non-normal distribution	• Independence • Ordinal scale	Comparing user satisfaction scores
Chi-Square Test	• Categorical outcomes • Independence testing	• Expected frequency >5 • Independence	Testing payment method preferences
Fisher's Exact Test	• Small sample sizes • 2x2 contingency tables	• Independence	Rare event analysis (fraud rates)
Kolmogorov-Smirnov	• Distribution comparison • Continuous data	• Independence • Continuous scale	Comparing entire distributions of loan amounts

Test Type	Use Case	Assumptions	Affirm Example
Bootstrap Methods	• No distribution assumptions • Complex statistics	• Independence	Confidence intervals for median time to repayment
Bayesian Methods	• Prior information • Sequential analysis • Posterior probabilities	• Prior specification	Updating conversion rate beliefs with new data
CUPED/CUPAC	• Variance reduction • Pre-experiment data	• Covariate stability	Using historical purchase data to reduce variance

2.2 Sample Size Calculation Methods

Method	Formula/Approach	When to Use	Affirm Example
Classical Power Analysis	$n = 2\sigma^2(Z_{\alpha/2} + Z_{\beta})^2/\delta^2$	Standard A/B tests	15,000 users per variant for 80% power
Minimum Detectable Effect	$MDE = (Z_{\alpha/2} + Z_{\beta})\sqrt{(2\sigma^2/n)}$	Fixed sample constraint	Can detect 0.5% lift with current traffic
Sequential Analysis	Alpha spending functions	Continuous monitoring	Check results daily with adjusted p-values
Bayesian Sample Size	Posterior probability thresholds	Prior information available	95% probability variant is 2% better
Simulation-Based	Monte Carlo methods	Complex designs	Simulate 10,000 experiments
Adaptive Designs	Information-based stopping	Efficient resource use	Stop when confidence interval < 0.1%

---

3. Experiment Design & Implementation

3.1 Randomization Strategies

Strategy	Implementation	Pros/Cons	Affirm Use Case
Simple Random	Random number generator	Pro: Unbiased Con: Imbalanced groups	Basic feature tests
Block Randomization	Randomize within blocks	Pro: Balance covariates Con: Complex	Balance by merchant type

Strategy	Implementation	Pros/Cons	Affirm Use Case
Stratified Random	Random within strata	Pro: Representative Con: Need strata info	Ensure all credit tiers represented
Cluster Randomization	Randomize groups	Pro: Avoid contamination Con: Less power	Randomize by merchant
Matched Pairs	Match then randomize	Pro: Increased power Con: Matching complexity	Match users by purchase history
Hash-Based	Deterministic hashing	Pro: Consistent Con: Predictable	User ID % 100 < 50
Time-Based	Temporal allocation	Pro: Simple Con: Time confounds	Day-of-week randomization
Geo-Based	Geographic clustering	Pro: Market-level Con: Spillover effects	Test in specific states

3.2 Implementation Frameworks

Framework	Architecture	Use Case	Affirm Implementation
Feature Flags	• LaunchDarkly • Optimizely • Split.io • Custom flags	Real-time control	Toggle checkout flow variants
Server-Side Testing	• Backend randomization • API-driven • Database flags	Sensitive features	Payment terms calculation
Client-Side Testing	• JavaScript-based • Google Optimize • VWO	UI/UX changes	Button color testing
Edge Computing	• CDN-level • Cloudflare Workers • Lambda@Edge	Low latency	Personalized landing pages
Mobile SDKs	• Firebase A/B • Apptimize • Custom SDKs	App experiments	Mobile app checkout flow
Full-Stack Platforms	• Amplitude Experiment • Statsig • Eppo	End-to-end solution	Integrated testing platform

---

4. Monitoring & Quality Assurance

4.1 Pre-Launch Checks

Check Type	Methods	Critical Metrics	Affirm Example
AA Testing	Run control vs control	• Type I error rate • Randomization balance	5% false positive rate confirmed

Check Type	Methods	Critical Metrics	Affirm Example
Sample Ratio Mismatch	Chi-square test on assignment	• Expected vs actual ratio • P-value < 0.001 flags issue	50.1% vs 49.9% (acceptable)
Covariate Balance	• T-tests on features • Standardized differences	• Age, income, credit score • <0.1 std dev difference	All user attributes balanced
Technical Validation	• Logging verification • Event tracking • Edge cases	• 100% event capture • No data loss	All checkout events tracked
Canary Deployment	Gradual rollout	• Error rates • Performance metrics	1% → 5% → 25% → 50%

4.2 Runtime Monitoring

Monitoring Type	Tools/Methods	Alert Thresholds	Response Plan
Statistical Significance	• Sequential testing • P-value tracking • Confidence intervals	• Alpha spending • Adjusted p-values	Daily significance checks
Practical Significance	• Effect size monitoring • Business impact	• Minimum meaningful difference • ROI thresholds	Alert if effect < 1%
Data Quality	• Missing data rates • Outlier detection • Event validation	• >5% missing data • Impossible values	Automated data quality reports
System Health	• Latency monitoring • Error rates • Resource usage	• P95 latency >2s • Error rate >0.1%	PagerDuty alerts
Sample Size Tracking	• Daily accumulation • Projection models	• Behind schedule • Imbalanced exposure	Traffic allocation adjustment

---

5. Analysis Methods & Techniques

5.1 Primary Analysis Approaches

Method	Statistical Technique	Implementation	Affirm Example
Frequentist Analysis	• Hypothesis testing • Confidence intervals • P-values	• Python statsmodels • R stats package	t-test on conversion rates

Method	Statistical Technique	Implementation	Affirm Example
Bayesian Analysis	• Posterior distributions • Credible intervals • Bayes factors	• PyMC3 • Stan • JAGS	P(variant > control) = 0.97
Regression Adjustment	• ANCOVA • Linear models • Propensity scores	• Include covariates • Reduce variance	Adjust for user credit score
Machine Learning	• Causal forests • Double ML • Targeted learning	• CausalML • EconML	Heterogeneous treatment effects
Time Series	• Interrupted time series • ARIMA models • State space models	• Seasonal adjustment • Trend analysis	Account for holiday effects
Survival Analysis	• Kaplan-Meier • Cox regression • AFT models	• Time to event • Censoring	Time to first purchase

5.2 Advanced Techniques

Technique	Purpose	Complexity	When to Use
CUPED	Variance reduction using pre-data	Medium	30-50% variance reduction possible
Difference in Differences	Control for time trends	Medium	Rollout experiments
Instrumental Variables	Handle non-compliance	High	Users don't see assigned variant
Synthetic Control Method	Create counterfactual	High	Single unit treatment (new market)
Regression Discontinuity	Exploit thresholds	High	Credit score cutoffs
Quantile Treatment Effects	Distribution impacts	Medium	Effect varies by user segment
Meta-Analysis	Combine multiple experiments	Medium	Synthesize learning across tests
Heterogeneous Effects	• Subgroup analysis • Interaction terms • ML methods	High	Personalization opportunities

---

6. Results Interpretation & Decision Making

6.1 Statistical Interpretation Framework

Aspect	Considerations	Best Practices	Common Pitfalls
Effect Size	• Practical vs statistical significance • Confidence intervals • Standardized effects	Report both absolute and relative	Focusing only on p-values
Multiple Testing	• Bonferroni correction • FDR control • Hierarchical testing	Pre-specify primary metric	P-hacking with many metrics
Segment Analysis	• Pre-specified vs exploratory • Interaction tests • Multiple comparison adjustment	Register segments before	HARKing (post-hoc storytelling)
Long-term Effects	• Novelty effects • Learning curves • Seasonality	Run for full business cycle	Stopping too early
External Validity	• Generalizability • Population differences • Context changes	Document assumptions	Over-generalizing results

6.2 Decision Frameworks

Framework	Decision Rule	Risk Consideration	Affirm Example
Hypothesis Testing	Reject if $p < \alpha$	Type I/II errors	Ship if $p < 0.05$
Bayesian Decision	Max expected utility	Loss function	Ship if P(lift>0) > 0.95
Risk-Adjusted	Consider downside	Worst-case scenario	Ship if worst case > -0.5%
Portfolio Approach	Optimize across tests	Resource allocation	Fund top 20% of ideas
Sequential Decision	Continue/stop rules	Opportunity cost	Stop if futility boundary crossed
Multi-Stakeholder	Weighted objectives	Stakeholder alignment	Product + Risk + Legal approval

---

7. Post-Experiment Actions

7.1 Implementation Strategies

Strategy	Approach	Risk Mitigation	Timeline
Full Rollout	100% immediate	Monitor closely	1 day
Gradual Rollout	50% → 75% → 100%	Staged validation	2 weeks
Holdout Groups	Keep 5-10% control	Long-term measurement	Ongoing
Regional Rollout	Geography-based	Market differences	1 month
Segment-Based	High-value users first	Protect core base	3 weeks

Strategy	Approach	Risk Mitigation	Timeline
Feature Flags	Instant rollback capability	Quick reversion	Immediate

7.2 Knowledge Management

Component	Methods	Tools	Deliverables
Documentation	• Experiment briefs • Analysis notebooks • Decision logs	• Confluence • GitHub • Notion	Experiment report template
Knowledge Sharing	• Review meetings • Wiki updates • Lunch & learns	• Slack channels • Email digests	Weekly experiment review
Meta-Learning	• Cross-experiment analysis • Pattern recognition • Failure analysis	• DataBricks • Tableau • Python notebooks	Quarterly insights report
Process Improvement	• Retrospectives • Efficiency metrics • Tool evaluation	• JIRA • Asana • Linear	Monthly process review

---

8. Advanced Topics & Considerations

8.1 Network Effects & Interference

Type	Detection Method	Mitigation	Affirm Example
Direct Network Effects	• Cluster randomization • SUTVA violations	• Market-level randomization • Synthetic control	Merchant referral programs
Indirect Effects	• Spillover analysis • Spatial models	• Buffer zones • Cluster designs	User word-of-mouth
Competitive Effects	• Game theory models • Market share analysis	• Simultaneous moves • Strategic timing	Competitor response to pricing
Platform Effects	• Two-sided markets • Feedback loops	• Careful measurement • Dynamic models	Merchant-consumer interactions

8.2 Ethical Considerations

Aspect	Guidelines	Implementation	Example
Fairness	• Equal treatment • No discrimination	• Bias testing • Fairness metrics	Equal approval rates across demographics

Aspect	Guidelines	Implementation	Example
Transparency	• User notification • Opt-out options	• Privacy policy • Clear communication	"You're seeing a new checkout experience"
Risk Management	• Minimize harm • Protect vulnerable users	• Guardrail metrics • Safety protocols	Don't test on high-risk loan applicants
Data Privacy	• GDPR compliance • Data minimization	• Anonymization • Retention policies	Hash PII, delete after analysis

---

9. Tools & Technology Stack

9.1 Experimentation Platforms

Platform Type	Options	Key Features	Best For
Commercial Platforms	• Optimizely • Adobe Target • Google Optimize • VWO	• Visual editors • Built-in stats • Integrations	Marketing teams, quick starts
Developer-Focused	• LaunchDarkly • Split.io • Statsig • Eppo	• Feature flags • SDKs • API-first	Engineering-led experimentation
Analytics-Integrated	• Amplitude Experiment • Mixpanel • Heap	• Unified data • User journey • Cohort analysis	Product analytics teams
Open Source	• GrowthBook • Wasabi • PlanOut • Unleash	• Customizable • Self-hosted • Cost-effective	Teams with engineering resources
Custom Built	• Internal platforms • Microservices	• Full control • Tailored needs	Large organizations

9.2 Analysis Tools

Tool Category	Options	Use Case	Integration
Statistical Software	• R + tidyverse • Python (scipy, statsmodels) • SAS • SPSS	Complex analysis	Data pipelines
Notebooks	• Jupyter • RStudio • Databricks • Google Colab	Exploratory analysis	Version control
BI Tools	• Tableau • Looker • PowerBI • Sisense	Dashboards	Real-time monitoring
Data Processing	• Spark • Presto • BigQuery • Snowflake	Large-scale processing	ETL pipelines

Tool Category	Options	Use Case	Integration
Workflow Management	• Airflow • Prefect • Dagster • Luigi	Automation	Scheduled analysis

---

10. Example End-to-End Experiment: Affirm Checkout Optimization

Stage 1: Planning

- **Hypothesis**: Reducing checkout steps from 5 to 3 will increase conversion by 25%
- **Metrics**: Primary: Conversion rate, Secondary: Time to complete, Guardrail: Error rate
- **Sample Size**: 50,000 users per variant (α=0.05, β=0.20, MDE=2%)

Stage 2: Design

- **Type**: A/B test with 50/50 split
- **Randomization**: Hash-based on user ID
- **Duration**: 3 weeks (full purchase cycle)

Stage 3: Implementation

- **Platform**: Internal feature flag system
- **Monitoring**: Real-time dashboard in Looker
- **Quality Checks**: AA test passed, SRM check daily

Stage 4: Analysis

- **Method**: Two-proportion z-test with CUPED
- **Results**: 23% lift (95% CI: 18%-28%), $p < 0.001$
- **Segments**: Larger effect for mobile users (28% vs 19%)

Stage 5: Decision

- **Recommendation**: Ship to 100% with 1-week staged rollout
- **Holdout**: 5% control for long-term monitoring
- **Next Steps**: Explore further optimization for desktop experience

Stage 6: Learning

- **Insight**: Form simplification has diminishing returns after 3 fields
- **Future Tests**: Focus on mobile-first design
- **Process Improvement**: Implement automatic SRM alerts

This comprehensive guide provides a complete framework for running experiments in a marketing analytics context, with specific examples from Affirm throughout. Each section includes multiple options and approaches, allowing teams to select the most appropriate methods for their specific needs and constraints.