УДК 004.8:330.46 **JEL Classification C81, C93, L86**

Hlinenko Larysa
ORCID: http://orcid.org/0000-0002-3105-4568
PhD (Engineering), Ass. Professor, Department of Electronics and Information Technology
National University «Lvivska Polytechnika»
(Lviv, Ukraine)
Daynovskyy Yuriy
ORCID ID: http://orcid.org/0000-0001-8421-2104
Doctor of Economics, Professor, Department of Marketing
Lviv University of Trade and Economics
(Lviv, Ukraine)
Zanichkovskyy Yuriy
ORCID ID: http://orcid.org/0009-0002-4188-3383
PhD Student, Department of Electronics and Information Technology
National University «Lvivska Polytechnika»
(Lviv, Ukraine)

CONTEXT-AWARE CHOICE OF METRICS IN A/B TESTING

The study is aimed at improving the methodology of selecting metrics for A/B testing by introducing an extended set of metric attributes that can serve as the main search criteria for determining the appropriate metrics for a particular experimental context. Along with such parameters as sensitivity, latency and noise, structural characteristics (e.g., data type and metric aggregation scheme, associated business process) are proposed to be included in the evaluation system. By summarising the published results of numerous studies, the values of key metric parameters such as sensitivity, noise and latency are ordered, a hierarchical model of interdependencies between metrics of different levels is built, and the procedure for forming a set of specific metrics for A/B testing using a metric tree for the correct identification of surrogates is determined.

Keywords: A/B testing, metric, metric tree, sensitivity, noise.

DOI: 10.15276/mdt.9.4.2025.1

Statement of the problem in general form and it's connection with important scientific or practical tasks. A/B testing is a type of controlled online experiments currently regarded as the "gold standard" for evaluating improvements to websites, mobile applications and services in e-commerce and search engines [56]. Large companies such as Amazon, Facebook, Google, LinkedIn, and Microsoft are investing in their own experimentation systems supporting A/B testing in many other companies. E.g., Microsoft's platform [27] supports experiments in Bing, MSN, Cortana, Skype, Office, xBox, Edge, Visual Studio, etc., running thousands of experiments per year. At the same time while numerous start-ups (e.g., Apptimize, LeanPlum, Optimizely, Taplytics) offer A/B testing solutions for smaller sites. A/B tests discussed in this article are online controlled experiments conducted on live traffic to support data-driven decisions in online businesses, such as e-marketplaces Amazon and eBay, portal sites like Yahoo and MSN [57], e-aggregators like Booking.com [44], search engines like Microsoft Bing and Google [24, 58].

ISSN 2522-9087 (Print)
ISSN 2523-434X (Online)

In the simplest controlled experiment (A/B test), users are randomly assigned to one of two options: control (A) or test (B). Usually, the control is an existing system, and the test is an existing system with a new property X added. The user's interaction with the system is registered and metrics regarded to be the indicators of treatment effect are calculated. If the experiment has been designed and executed correctly, the only thing that is consistently different between the two variants is the feature X. External factors, such as seasonality, the impact of launching other features, competitors' actions, etc., are distributed evenly between the control and treated variants and are supposed not to affect the results of the experiment. Therefore, any difference in performance between the two groups must be attributed to X. This confirms the causal relationship between changes made to the product and changes in user behaviour. Accordingly, the availability and selection of the right metrics is crucial for the successful conduct and subsequent evaluation of the experiment results [23, 28, 56].

Analysis of the latest research and publications, which initiated the solution of this problem and on which the author relies. The importance of metrics as a mechanism for providing correct assessment of the AB testing results has been recognised and studied in the management field since at least 1956 [56]. It is described by such statements as “What gets measured gets managed”, “What gets measured, gets managed”, “What you measure is what you get” and “You are what you measure” [27]. In recent years, several papers have been published on defining metrics for online controlled experiments [21, 23, 39, 61, 76].

K.Rodden et al. [76] proposed a user-centred framework for defining and interpreting metrics in A/B testing and, basing on Google's experimental infrastructure, presented a taxonomy of metrics tied to stages of the user journey, categorised as behavioural, attitudinal and business metrics.

L.Rachitsky in [74] substantiated the necessity of aligning the metric selection with business objectives. Based on their experience at Microsoft, W.Machmouchi and G.Buscher [61] presented a set of guiding principles for building reliable, interpretable, and decision-aligned metrics in online A/B testing: clarity and interpretability (metrics should have unambiguous definitions and business meanings); stability and sensitivity (metrics should balance reliability against random noise and the ability to detect meaningful changes); hierarchical relevance (each metric should be associated with a clearly defined purpose and place in the metric hierarchy). They propose a hierarchical organization of metrics, distinguishing between Objective Evaluation Criteria (OECs), which serves as the primary variables for decision-making, and supporting or diagnostic metrics, which are intended to facilitate root cause analysis but not to make treatment decisions on their own.

W.Machmuchi and G.Boucher emphasise that the effectiveness of A/B testing depends not only on statistical methods, but also on the quality and structure of the metrics themselves.

R.Kohavi and A.Gupta [37, 39] developed these principles of selecting “right” metrics in the form of the mandatory possession by metrics of at least seven basic properties STEDIL: Sensitivity (S), Trustworthiness (T), Efficiency (E), Debuggability (D), Interpretability and Actionability (I), Inclusivity and Fairness (I). Approaches to creating such metrics: division, decomposition and segmentation, are thoroughly outlined in [65].

In [23] and [61], several principles for developing good metrics for overall evaluation criteria are outlined. An overall evaluation criterion (OEC) is an indicator or a set of indicators $Y = \{Y_i\}$, by the change of which the effectiveness of implementing change X is evaluated. OCE metrics are key metrics for a given type of business, which can usually be measured easily during a short experiment. It is recommended to set only one metric as the OCE, but a composite indicator combining various metrics is also possible [27].

The development of a proper IES is a subject of active research [6, 13, 23, 27, 28, 56, 61]. In [28], a data-driven approach to assessing various indicators of OCE is described, while [6] considers the feasibility of using short-term and cumulative long-term metrics as OCE. A.Deng and X.Shi [23] introduced metric aggregation as a weighted linear combination, which was later adopted and extended by E.Kharitonov et al. [49].

K.Liou [60] et al. proposed a more sensitive variance-weighted treatment effect estimator relying on the individual variances of experimental users according to pre-experiment data. In [27], a wider set of problems related to the interpretation and selection of metrics is analysed. Many works also focus on the problems of correct interpretation of changes in metrics values [6, 27, 28, 56], the most common of which include ignoring or misinterpreting the statistical significance of a metric change; insufficient sample size, neglecting secondary effects (impact on other important metrics); incorrect assessment of metric changes during early test evaluation, using inappropriate statistical tests, incorrect interpretation of so-called ratio metrics, i.e. metrics calculated as the ratio of two other metrics; rejection of the presence of metric changes in case of large variance with a small sample size; failure to take into account user segmentation, novelty effect, etc. A lot of studies address the behaviour of metrics and describe methods for defining them.

R.Kohavi et al. [54] provide several examples of non-intuitive metric behaviour and show the importance of identifying skewed metrics with further truncating them to improve sensitivity.

P.Dmitriev and X.Wu in [28] present results of thorough examination of metric quality and state metric parameterization to be crucial for correct designing the A/B tests. They thoroughly characterise metrics in terms of parameters such as variance, bias, and sensitivity, emphasising that the choice of metric fundamentally affects the power and reliability of the experiment. They distinguish metric types - binary, continuous, and counting - and emphasise how the statistical properties of each type affect the ability to detect treatment effects (MDE values). They show that metrics with high variance or skewness can reduce statistical power, thus requiring the use of variance reduction techniques or alternative aggregations. A.Drutsa et al. [30] examine the sensitivity of user engagement metrics, especially those used as primary or guardrails in content and UX experiments.

Problems of metric interpretability and failure modes are reviewed by P.Dmitriev et al. [27], who substantiate that even sophisticated statistical analysis, can mislead decision makers if metrics are poorly aligned with causal or business goals. These cautions are reflected in R.Kohavi, D.Tang, and Y.Xu fundamental research [58], which offers a comprehensive roadmap for conducting reliable A/B testing in an industrial setting. Authors identify the functions of metrics in experimentation, classifying metrics as: primary (catching the intended effect); guardrail (safety constraints, e.g., latency, NPS); auxiliary, supporting metrics (used for diagnostics, debugging, e.g., failure rate, scroll depth) and exploratory metrics (not specified in advance; used to generate hypotheses). They emphasize such properties of metrics as directionality (moving in valid direction); stability (variance and seasonality); sensitivity (MDE and CV), and actionability (producing relevant decision). A key point is highlighting the role of guardrails as practical checkpoint metrics which should not deteriorate even if the primary success metric improves.

A.Deng [17] investigates statistical properties of metrics. Deng categorizes metrics by data type and aggregation schemes, highlighting the influence of metric choice and metric calculation method on statistical power, experimental sensitivity and interpretability. In [18] A.Deng proposes to define the sensitivity of a metric as the probability that the metric will experience a statistically significant change in response to the true effect of a treatment

evaluating its significance by statistical power (the probability of catching a treatment effect) and probability of change (movement) in desired direction. It is stated that improvement of sensitivity may be achieved by increasing the power through variance reduction (e.g., CUPED, covariance adjustment), by increasing the sample size or by increasing the effect size by changing transforming the definition of the metric (e.g., changing the type of aggregation).

Improving the sensitivity of A/B tests makes experiments more efficient due to the fact that, for fixed error probabilities, a more sensitive test is better at detecting smaller effects with greater precision or requires less data, which drives the large number of studies focusing on ways to increase sensitivity. Converting the resulting metric to a metric with lower variance is a common approach to improve sensitivity [14, 21, 24, 59]. A.Deng, Y.Xu, R.Kohavi and T.Walker [24] introduced CUPED (Controlled-experiment Using Pre-Experiment Data) method reducing the variance (variability) of a metric in A/B tests by using previous user data (covariates) to predict current values. The method is based on the application of covariance correction of the original metric Y by a value determined on the basis of historical (pre-experimental) data concerning the metric behaviour. Method proved to provide 20–50% variance reduction when appropriate covariates are available and metric is aggregated as average one. Further improvement of the method by introducing augmentation and metric decomposition [19] due to the application of ANA (Approximate Null Augmentation) method made it possible to extend the technique to ratio and percentile metrics, improving signal-to-noise ratio and allowing for variance reduction even when pre-period data is unavailable or weakly correlated. Problems of CUPED application under weak correlation of covariate with the response to change [59] could be probably overcome by synthetic metrics using for modelling the response of the main metric in the population under study a similar population that has not been exposed to the change [59, 88]. Various other studies report that historical data can be used to improve sensitivity by reducing the variance [24] or directly to optimise sensitivity [49]. Kohavi et al. [54] describe several examples of metric misbehaviour, illustrating the benefits of detecting distorted metrics in a timely manner and cutting them off (truncating them) to improve sensitivity. Other transformations to improve sensitivity include binarisation of quantity and revenue metrics [59]. Some studies have shown that sensitivity can be improved by adapting the method of evaluating test effectiveness [26]. D.Dimmery, E.Bakshy, and G.Sekhon [26] developed consistent, small bias, empirical shrinkage Bayes Stein-type estimators for detection of causal effects in large-scale online experiments applying multi-arm bandit algorithms. By using shrinkage techniques, they reduce variance while controlling bias, leading to more reliable effect size estimates for metrics that are otherwise highly volatile.

In [21] authors define the metric decomposition as a sensitivity optimization procedure. The authors propose a concept of component-based sensitivity metric, breaking down the metric components into submetrics (e.g., CVR × AOV for Revenue) and calculating the components of sensitivity – variance and level of motion sensitivity – separately for each component. They show that that decomposed metrics, such as weighted aggregate metrics for each user, can significantly increase sensitivity by reducing variance and focusing on components that are responsive to treatment and present a simulation-based diagnostic procedure for comparing candidate metric formulations.

In case long-term business or strategic metrics are used in the AB test, directionality is added to these characteristics [13, 27, 41, 53]. Directionality identifies how well the metric aligns with long-term business results [23, 27, 53]. For example, P.Dmitriev and X.Wu [28] discuss the possibility of engaging subject-oriented experts to develop true directionality labels for a set of experiments. H.Hohnhold et al. [41] develops directionally consistent

metrics using a set of long-term experiments. In most cases, correlation and regression analysis methods based on "historical" data are used to assess directionality, while in [13] survival modelling is applied to validate directionality and surrogacy for determining a particular set of short-term metrics capable of assessing the long-term metric. A.Deng and X.Shi [23] treat directionality (stable metric movement in one direction for positive treatment effects and in the opposite direction for negative effects) as one of the important characteristics in metric selection, suggesting that preliminary experiments should be used to collect a corpus of good metrics and estimate sensitivity and directionality using Bayesian a priori data. They provide seven empirically based lessons on developing, validating and deploying metrics on such large-scale platforms as Bing and Skype and propose a practical guidance on developing a metric system, emphasizing the necessity of: alignment with the business (metrics should reflect what really matters); back-testability (showing a consistent difference between treatment and control in previous experiments); reducing variance by employing CUPED and other sensitivity increasing methods and correct choice of aggregation and randomization (per user, per session, per population). This is complemented by recommendations of M.Schulzberg et al. [79] focusing on the use of multiple metrics in risk-aware decision rules.

Considering directionality as constituent of the metric sensitivity is further developed by O.Jeunenet al [46] and Wen Qin et al [72]. In addition to common error types – Type I (false positives) and Type II (false negatives) – O.Jeunen et al. [46] introduce the Type III errors (errors in effect sign connected with opposite movement direction of North Star and supporting metric) and proposed methodology for validating proxies or secondary metrics based on their error characteristics and statistical power. It allows ranking proxy indicators by reliability and effectiveness thus providing a data-driven validation pipeline for evaluating proxy indicators in any production environment which facilitates quick and trustworthy decision-making when North Star indicators are inconclusive.

Microsoft researchers Wen Qin et al. [72] continue investigation of metric nature started by A.Deng et al. [23]. In analysis of metric actionability, A.Deng and X.Shi [23] decompose sensitivity into two components: statistical power and movement probability both being important for successful application of metrics. Following that authors [72] consider Metric sensitivity as movement probability × statistical power and propose to assess both components of metrics using power analysis to verify that the minimum DTE (Detectable Treatment Effect (DTE) or MDE) is attainable and movement analysis performed on historical A/B tests to estimate movement probability and to study if the metric moves as expected. Movement analysis engaged a historical test corpus, both labelled (known to have effects) and unlabelled to compute number N_i of true/false positives and negatives within the corpus. The obtained confusion matrix of true/false positives and negatives (fig. 1) is used to identify the movement component of sensitivity by summarizing the behaviour of the metric on the labeled corpus. Authors expect a sensitive to have a large $N_1/(N_1+N_2)$, as close to 1 as possible, where N_1 and N_2 are the numbers of H_1 (treatment effect exists) true positives and negatives respectively. A robust metric (i.e., one that is less susceptible to noisy movements) will have $N_3/(N_3+N_4)$ where N_3 and N_4 are the numbers of H_0 (treatment effect doesn't exist) true positives and negatives respectively. The authors state, that the metric of interest has enough sensitivity if it fulfils the following qualities: a) it is assessed that the minimum DTE of the metric can be achieved; b) the movement direction is always explainable for A/B tests covered by N_1 in the confusion matrix; confusion matrix shows for the metric high $N_1/(N_1+N_2)$ and $N_3/(N_3+N_4)$ close to the significance level; the metric has large observed movement probability compared with other candidates. If no metric from the developed fulfils all of

them, the one that meets most of them is to be chosen. Furthermore, the metric sensitivity could be improved by: reducing the impact of outliers by capping the value of the metric at a fixed maximum value [58, 61]; by variance reduction by CUPED or shrinkage [26], by substituting the metric value by the Log function on the metric value: $x \rightarrow \log(1+x)$ or by changing the level of metric aggregation: Single Avg → Double Avg or Weighted Avg [47]); by applying more short-term proxy metric found by predictive models [39].

The recent N.Larsen et al. paper [59] provides a thorough review of A/B testing methodology, addressing the key issues of statistical inference, sequential testing, metric correlation, and heterogeneity. This review emphasises the correlation, confusion and multiplicity of different metric types and reveals their special features: highly correlated metrics (e.g., time on site and scroll depth) can overstate the first type error; low-power metrics (e.g., churn rate) require longer test durations or larger samples; composite metrics multiple outcomes (e.g., engagement score), but are likely to obscure trade-offs. The authors state the necessity of multimetric correction and pre-experimental metric sensitivity analysis.

A.Fabijan et al. [32] present consolidated checklists designed to increase the reliability and scalability of online controlled experiments. The checklists cover metric design at high-level and advocate for designing the metric set instead of single metric reflecting the specific user behaviour the tested feature is supposed to induce (feature metric). This set is to contain in addition to feature metric such metrics as: data quality metrics, guardrail metrics proving other metric movements are not crossing important business constraints, and success metrics revealing whether the tested feature promotes achieving long-term company goals. Metrics are to be examined for choosing a winning variant specified for each kind of metrics. Authors show that most of the experiment analysis is to be conducted before the experiment is started, and clarify ways of ensuring the process quality though their list of items does not include checks on the ability of the selected metrics to provide the desired experiment parameters (duration, power, etc.) due to the metrics' specific characteristics.

A lot of research addressing to A/B testing is related to the study of so-called proxies or surrogate metrics. These metrics substitute the goal metric in the A/B test due to less latency, bigger sensitivity and simplicity of measuring. Duan et al. [31] specify the requirements for proxy metrics: sensitivity to treatment effect, casual connection, robustness, small latency, interpretability, actionability and consistency with the target metric, confirmed experimentally by strong correlation with the primary metric. The applicability of surrogate metrics to increase the efficiency of AB testing is investigated and confirmed on several real business cases in [31]. At the same time, A.Deng, N.Larsen and others [14, 21, 24, 59], studying the properties of surrogate metrics, prove that their use is limited by the problems of minimum detectable treatment effect (MDE). To eliminate these problems and increase the sensitivity and reliability of A/B testing in an online environment, authors propose to decompose the metric of interest into components with different signal-to-noise ratios, with allocating components with high signal and low noise.

S.Athey et al. [7] present a typology of cases of successful or unsuccessful use of surrogate metrics, and outline the conditions under which proxies can be meaningful indicators for long-term impact. [13] provides practical guidance on the development and validation of surrogacy metrics and proposes a method for testing statistical assumptions about surrogacy to ensure that directionality is consistent with long-term outcomes without any additional expert judgement or long-term experiments.

The latest studies have proved the efficiency of application valid combinations of multiple surrogates [7]. V.Zhang et al. [90] implement the developed in [7] surrogate index approach of estimating longer-term impact from shorter-term outcomes in A/B tests. A.Bibaut

et al. [11] propose to solve the problem of reasonable selecting the directional and reliable short-term surrogate metrics (S) able to stand in for a long term primary metric (Y) basing on the correlation between S and Y. The true correlation is suggested to be estimated via the covariance matrix of true average treatment effects (ATEs) on Y and Ss in previous experiments with modifying the naïve estimator of S/Y correlation to overcome problems of evaluating surrogacy in A/B testing with low noise / signal ratio. Developed linear models of relationships between metrics proved to be good in managing metric trade-offs, especially in case of contradictory impact of different proxies.

A.Zito et al. [92] introduces a new approach to the evaluation and selection of proxy (surrogate) metrics based on the concept of Pareto optimality. Admitting the contradictory character of two main requirements to the good proxy metric: short term sensitivity and long term correlation with North Star metric, – the authors show that the choice of appropriate proxy always requires a trade-off between these features. To solve the problem of optimization the choice, the authors, instead of aggregating these parameters into a single scalar score, propose identifying Pareto-optimal proxy metrics – metrics for which no other combination or single metric offers strictly better sensitivity and directionality defined as a linear combination (weighted sum) of their recalculated values; all proxy metrics in combination are to be instrumental and causally connected to outcomes of interest. This approach does not give a single “best” proxy but reveals a Pareto front of decisions, enabling to choose one of them according to acceptable risk and response time.

M.Schultzberg, S.Ankargren, and M.Frånberg [79] suggest forming decision rules for A/B testing employing several heterogeneous metrics to be based not only on statistical hypothesis testing but also on the consideration of Bayesian risk thresholds for different metrics. They distinguish success metrics (primary metrics which are wished to improve, e.g., Revenue, CR etc.), guardrail metrics (protective metrics for ensuring the indicators of the key business processes (e.g., latency, churn, UX) do not deteriorate), deterioration (metrics that allow for deterioration within certain limits, but require deterioration to be avoided beyond a certain threshold (e.g., increase in support ticket, bounce rate)) and quality metrics (monitoring the experiment integrity or data pipeline status (e.g., sample ratio mismatch, population balance)) that are integrated into the decision-making procedure by productive rules (e.g., “approve if success is achieved and performance does not deteriorate”), enabling choice of decision alternatives robust to metric uncertainty and multidimensional trade-offs. The new feature is recommended for adoption if a) at least one success metric is statistically superior; b) all guardrail metrics do not show statistically significant deterioration (are not lower); c) none of deterioration metrics shows degradation exceeding the stated threshold and d) all quality checks are passed. So, the concept of composite multi metric testing [58], is expanded by risk-aware decision making algorithm.

Highlighting the previously unresolved parts of the general problem to which the article is devoted. Despite the growing body of literature on evaluating A/B testing results using various types of metrics, several important gaps remain. Numerous studies have proposed a lot of metric taxonomies that classify metrics into different categories – such as: primary, secondary/ surrogate metrics, key performance indicators (KPIs) and strategical, tactical and operational metrics; performance and guardrail metrics etc. – and explore methods for determining effective combinations of metrics. A wide range of metric properties have been studied and recommendations have been made on which properties (e.g., sensitivity, variance) should be considered when selecting metrics and what criteria “good” metric has to satisfy [7, 27, 28, 36, 46, 53, 58, 72]. Common errors in metric interpretation

and sensitivity of engagement metrics, driving to the practical risks of unbalanced or unstable metrics, were studied by P.Dmitriev et al. [27], A.Deng et al. [23] and A.Drutsa et al. [30].

Recent advances in the methodology of online controlled experiments have increasingly focused on the development, evaluation and selection of metrics that help to make decisions under uncertainty. M.Schultzberg et al. [79] presented a methodology for risk-aware A/B tests design, whereby metrics are divided into primary, guardrail and evaluation metrics, and are judged both by statistical evidence and business risk preferences. Alongside this, S.Athey et al. [7] and A.Zito et al. [92] have highlighted the role of surrogate or proxy metrics in accelerating reliable conclusions as they developed formal tools, such as the surrogate index and Pareto optimality principles, to evaluate the quality of proxy metrics in the context of long-term impact estimation. Extending these conceptual contributions, empirical methodologies, such as those proposed by W.Duan et al. [31] and O.Jeunen et al. [46], evaluate the performance of metrics based on observed errors in historical experiments. In particular, O.Jeunen et al. present a thorough pipeline for evaluating the error performance of types I-III secondary metrics used in large-scale A/B tests, thereby overcoming the gap between theoretical desirability and operational reliability. This is in line with the findings of R.Kohavi et al. [53, 54, 58] and A.Deng et al. [21, 23] who also encourage empirical validation and robustness analysis of metric choice. Zito et al. [92] suggest the complex proxies design and Pareto efficiency rather than point diagnosis based on a single metric. Taken together, the works demonstrate a significant shift in A/B testing practice from manual KPI development to data-driven and multi-objective optimised metrics with main emphasis on sensitivity, variance control, alignment with business decisions, and causal interpretation.

Despite this progress, the recent literature has not fully addressed several important issues. First, existing studies often do not consider important contextual attributes of metrics, such as the business process or object under evaluation or the types of data/aggregations of metrics. Second, few studies quantify key characteristics that affect the metric's usefulness, such as sensitivity, variance (noise), and latency which defines minimal required test duration. Third, while the literature often refers to the causal relationship between surrogate and key business metrics, the structure of interdependencies between metrics of different levels (strategic, tactical, operational) remains underdeveloped, and no explicit models or diagrams of interdependencies between metrics are provided. The descriptive hierarchies of indicators proposed in [23, 61] or [17, 21, 59] do not provide a metric hierarchical tree that reflects the interconnections between indicators and casual relationships. Thus, they cannot capture the hierarchical dependencies and pathways of influence between metrics of different ranks, which could provide a basis for selecting adequate surrogate metrics with better temporal, sensitivity and noise characteristics to be used in the design of an optimal metric set.

Formulation of the purpose of the article (statement of the problem). This research aims to fill these gaps by expanding the typology of metrics characteristics that can serve as primary gates in the selection of adequate metrics, providing quantitative or at least ordinal estimates of their sensitivity, latency and variance (noise) and identifying hierarchical relationships and interactions between metrics of different ranks, thus creating a basis for displaying dependencies between metrics in the context of A/B testing of marketplaces. Ultimately, this work aims to support the formalisation of A/B testing planning at an early stage by enhancing the metrics selection process.

Statement of the main material of the research with full justification of the scientific results obtained. A metric is usually a quantitative indicator for evaluating the state of business by determining a characteristic of a population that reflects this state. In A/B

testing, metrics are used to compare the tested variants: the values of the indicator achieved during the experiment for each of the variants define the efficiency of the tested changes.

A general list of metrics used in AB testing in B2C e-commerce is given in [78]. Based on [5, 8, 12, 34], the list of the main marketplace metrics is given in the first three columns of Table 1. The ranks of metrics and their classification features (columns 4-6 of Table 1) were developed by the authors.

The metrics themselves (Table 1) are not specific to 2-sided marketplaces, but unlike other types of e-commerce platforms, they are collected separately for sellers and buyers. For example, conversion to the target action (CR) is evaluated for buyers by such indicators as: conversion to registration; conversion to sending a message to the seller; conversion to adding a product to the cart (Add to Cart Rate); conversion to purchase (Checkout Rate, Purchase Rate), while for sellers through conversion to registration; conversion to placing an ad; conversion to selling on the platform.

Table 1 – Marketplace metrics used in AB testing

Metric (Name, Abbreviation)	Interpretation	Calculation formula	Rank	Role type Studied object / process	Funnel stage
Gross Merchandise Value, GMV	Total value of goods sold through marketplace	Sum of order values per period, AOV × Number of orders	0	Primary, KPI Monetization Scaling	Action
Net Revenue	Revenue after deductions (returns, cancellations)	GMV×Take Rate /100% or GMV - Returns - Discounts	0	Primary, KPI Monetization	Action
Customer Lifetime Value, CLTV	Predicted total revenue from user over lifetime	ARPU × Avg Lifetime; Sum of discounted future revenues	0	Primary. KPI Monetization / Retention	Retention, Action
Take Rate	Percentage commission fee taken per transaction	Value is defined by business model (Net Revenue / GMV)×100%	1	Primary / Exploration Monetization / Business model	Action
Average Revenue per User, APRU	Average revenue generated per user, ties user engagement to revenue	Net Revenue / Active Users (Buyers + Sellers)	1	Primary / Proxy Monetization	Action
Revenue Per Visitor, RPV	Average revenue generated per site visitor	Total revenue / Number of visitors	1	Primary / Secondary Monetization Acquisition / UX	Action
Net Promoter Score, NPS	Loyalty sentiment score	%Promoters - %Detractors	1	Primary / Secondary Retention, UX	Intent, Action Retention
Conversion Rate, CR	Percentage of visitors who complete a purchase	(Purchasers / Visitors) × 100%	1	Primary /secondary Conversion	Action
Retention Rate (Day 7, 14, 21, 30, 60)	Percentage of users retained over time	[(Customers at the End of Period -New Customers)/ Customers at the Start of Period] × 100%	1	Primary/Secondary/ Proxy (For Day 7-14)	Retention
Buyer-to-Seller Ratio	Marketplace balance of demand and supply	(Unique Buyers / Unique Sellers) × 100%	1	Secondary/guardrail Liquidity, health	Intent Awareness
Customer Acquisition Cost, CAC	Average cost of acquiring a new customer	Total marketing spend / New customers	1	Primary/Secondary Monetization	Action Retention
Return on Ad Spend (ROAS)	Revenue generated by each $ spent on advertising	Revenue attributed to ads / Ad spend	1	Primary/Secondary Monetization Acquisition	Awareness Action
Buyer | Seller Churn Rate	Share of users (buyers or sellers) who stopped using the service	(Lost users / total users at the start of the period) × 100%	1	Secondary/guardrail Performance, Marketplace	Retention

	throughout the period			liquidity, health	
Average Order Value, AOV	Average amount spent per order (average cheque)	GMV / Number of Orders	2	Proxy Monetization	Action
Returning Users Rate	Share of users returning to the platform after first visit	(Returning users / Total users) ×100%		Proxy Retention Engagement, UI	Retention
Repeat Purchase Rate	Share of users who made more than one purchase	(Returning Buyers / All Buyers) × 100%	2	Proxy Retention	Retention
Click-Through Rate, CTR	Share of users who clicked after seeing content	(Clicks / Impressions) × 100%	3	Proxy Acquisition, Engagement	Awareness Consideration
Customer satisfaction score, CSAT	Average customer satisfaction	Sum Scores / Responses	3	Proxy Retention	Retention Awareness Consideration
Add to Cart Rate	Share of sessions with items added to cart	(Add to Cart Events / Sessions) × 100%	2	Secondary / Proxy Conversion Monetization	Intent
Checkout Rate	Percentage of users who proceed from cart to checkout	(Checkouts / Sessions) × 100%	2	Primary / Secondary / Proxy Conversion UX	Action
Cart Abandonment rate	Percentage of tasks (carts) users start but don't complete	(Abandoned carts / Created carts) × 100%	2	Secondary / Proxy Conversion Monetization	Consideration
Bounce Rate	Percentage of visitors leaving after viewing only one page	(Visitors with single page session / visitors) × 100%	3	Secondary / Proxy Acquisition /UX	Consideration
Engagement Rate	Share of users who interacted with content or functionality	(Engaged users / Visitors) × 100%	3	Primary / Secondary Engagement	Consideration
DAU/WAU/MAU	Active users in daily/weekly/monthly buckets	Unique Active Users / Period	2	Proxy Acquisition	Awareness Intent
Match Rate	Share of searches with satisfactory listings	(Searches with Match / All Searches) × 100%	2	Secondary Liquidity, health	Intent
Fill Rate	Share of search or demand requests fulfilled by inventory	(Fulfilled / Total Requests) × 100%	2	Secondary Performance, Liquidity, health	Action
Supply Coverage	Share of demand covered by available supply	(Available Stock Keeping Units (SKUs) / Demand SKUs) × 100%	3	Secondary Performance, Liquidity, health	Intent
Fulfilment Rate	Share of paid orders fulfilled successfully	(Fulfilled Orders / Paid Orders) × 100%	2	Secondary Performance, Liquidity, health	Action
Utilization Rate	% of inventory or capacity actually used	(Used Inventory Capacity / Total Available) ×100%	2	Secondary Performance, Liquidity, health	Action
Scroll Depth	Average percentage of page scrolled by visitors	(Total scrolled pixels / Total page height) × 100%	3	Proxy Engagement	Consideration
Time on Site	Average duration of visitor sessions	Sum of session durations / number of visits	3	Proxy Engagement	Consideration

Monetisation metrics depend on the business model, and there are marketplace-specific metrics that describe the marketplace's liquidity through the balance of supply and demand, of buyers and sellers (Match rate, Fill Rate etc.).

Specifying metrics for A/B testing is an initial step determining the test result for the business. When choosing metrics, one should first of all consider the business parameters they reflect, the tasks they solve directly (metric level), the components of the business system and the business processes they characterise, etc. For e-marketplaces, there are specific metrics

characterising the effectiveness of a marketplace website or application: overall website popularity (DAU, Sessions), dynamics of attracting new users (Unique Visitors); engagement with content and ways of its presentation, compliance with expectations (Time on Site, Bounce Rate, Page per Session, Error Rate); balance of sellers and buyers: efficiency in achieving business goals (conversions, churn rate, return rate); traffic sources and efficiency of traffic management (traffic from search engines; position in search results); key search words; social traffic indicators (traffic from networks and the number of mentions, likes, etc.).

Different sources offer different classifications of metrics according to the identified classification features: by the degree of direct impact on the achievement of business goals [28, 58]; by the role in A/B testing (primary, secondary, guardrail, data quality); by the place in the metrics hierarchy; by the object of evaluation (business status in general, liquidity, monetisation, user attitude, user behaviour, data quality, etc.); by randomisation (granular) level (session, user, subgroup, cluster (cities, categories, or geographic regions, device, [84], etc.); by the method of determination; by the purpose of the evaluation (evaluation of test success, evaluation of metrics suitability, evaluation of test correctness, evaluation of population homogeneity, etc.) by the function performed in the selection and interpretation of metrics (success metrics, data quality metrics, protective and diagnostic metrics [27, 36, 50]).

Classification of metrics into 3 types according to their role in AB testing (by functionality, or how they operate in the experiment): primary, secondary and guardrail, was first proposed by R. Kohavi et al. in [53, 54, 58]. Recently the additional category of metrics referred as exploratory was identified in practical guides for AB testing [1, 2].

Primary Metrics, also known as Core Metrics, Goal Metrics, North Star Metrics, or Success Metrics, are usually represented by the top-level business metrics and serve as Key Performance Indicators (KPIs) for evaluating an A/B test [70]. They reflect the main objective of the experiment and are closely aligned with the strategic long-term business goals and, accordingly, usually require a long time to evaluate their changes. Such indicators for the marketplace can be as follows GMV, Net Revenue, CLTV, Take Rate etc. Such Top Level business indicators are rarely used as primary indicators in the A/B test due to their latency and aggregation. In practice, the next level of the hierarchy, which directly defines these roadmaps, is used as primary metrics, namely: CR Match Rate, Fill Rate, Retention Rate, Utilization Rate etc. A primary metric shows whether the change being tested is achieving the desired result. It is used to plan the experiment for determining the minimum detectable effect (MDE) and sample size (to ensure sufficient power to detect meaningful change) [2].

Secondary Metrics (Driver Metrics, Leading Metrics) are indirect or predictive indicators that are commonly used to measure short-term goals, e.g., AOV, Cart Abandonment Rate, Add To Cart Rate. They are aligned with the key metrics, but are more sensitive and actionable, and better adapted for online experiments. Secondary indicators are not decision-making criteria, but they help to comprehensively understand the impact of the experiment and to identify unforeseen side effects. For example, if we aim to attract new users, the number of new users registered per day may be one of the driving metrics. Within this group, a subgroup of so-called proxy metrics or proxies [2] can be distinguished. Proxy indicators are indirect indicators used to evaluate an outcome that cannot be measured directly (e.g., in case of long-term metrics). In the experiment, proxy indicators can be used as a substitute (surrogate) for the actual goal metric. The surrogate indicator should have a strong correlation with the goal metric which should be checked regularly, e.g., Time on Cite as a proxy for engagement; Click (On_Buy_Button) or Checkout Rate as proxies for conversion.

Guardrail metrics (e.g., Page Load Time, Payment Failure Rate, and Customer Support Tickets) are safeguards for monitoring and ensuring the health, stability, and overall integrity

of a system during an experiment. They do not measure the success of the experiment, but are vital to identify unintended negative impacts on business, user experience and/or operational performance and to catch early potential risks of degraded performance, increased error rates, or adverse user behaviour (e.g., a new feature may increase latency, leading to a higher bounce rate or more subscription cancellations). This category of metrics also includes the so-called sanity metrics [38], which track the correctness of the plan, procedure, and implementation of the A/B test, for example, sample match and homogeneity of the control and test groups, etc. At the same time, the driver metrics of one team may be the guardrail metrics for another team/teams.

Exploratory metrics are the metrics of interest used not for decision-making or experiment evaluation but for detecting changes in behaviour, unintended consequences or hypothesis generation (e.g., Interactions per Session or Time on Site (as proxy for Engagement Rate)). They are often used in post-analysis to reveal new directions of research.

Composite metrics are a combination of several separate metrics into one to capture a diverse view of success. They are often used as strategic metrics, but can lead to reduced sensitivity. Examples: Overall Evaluation Criterion (OEC) as a weighted combinations of metrics of engagement retention rate, engagement rate etc., revenue (RPV, APRU, CR), and satisfaction (NPS; CSAT; Utility as difference between an aggregated reward function for each documented interaction c_i (e.g., long click, position, dwell time) and function Penalty (Effort), penalizing quantity of effort (e.g., multiple queries, time spent) [62]).

According to the function performed in the selection and interpretation of metrics, they are distinguished as success metrics, data quality metrics, guardrail, debugging and diagnostic / local function metrics [23, 27, 36, 50].

The success metrics are the metrics that evaluate the efficiency of tested changing. These are the primary metrics according to R. Kohavi's classification [58].

Data Quality Metrics include metrics that allow one to assess the correctness of the planning and execution of the experiment, ensuring the qualitative reliability of the experiment results, identifying and eliminating the inconsistency of the measurement method with respect to a new property (too little time, untimely stopping of the test) or other problems. An example of such a metric is the ratio of the number of users in the treatment group to the number of users in the control group: sample ratio mismatch (SRM) generally means a strong sampling bias sufficient to invalidate the results of the experiment [91].

Local Feature metrics [27] and Diagnostic Metrics [36] are used to measure the usage and functionality of specific features/components of a product (e.g., CTR on specific elements of a web page). These metrics are often used to identify the source of changes in higher-level metrics and to diagnose Overall Evaluation Criteria (OEC) metrics, helping to understand where the change (or lack of change) in OEC is occurring. At the same time, the assessment of changes in local metrics can be ambiguous: it is easy to increase the CTR on a page element at the expense of other elements. Improvements in local metrics may also be due to undesirable side effects of the introduction of new functionality.

Debugging metrics are usually used to help us understand why some important metrics, especially the goal ones, move or do not move, and how to interpret their movements [23]. For key metrics whose models are based on several indicators of user behaviour, it is important to understand how these individual indicators affect the overall dynamics of the metric and whether it is worth decomposing the target metric into simpler components.

Definition of Guardrail Metrics by [23, 27, 36, 50] as which, not being indicators of the feature under test, should not be degraded while deciding whether to implement the feature under test, is standard for metrics of this type [53, 54, 58]. For example, on websites

such as Bing or MSN, Page Load Time (PLT) is usually a guardrail metric for which only a slight degradation is allowed.

The results of A/B testing are evaluated by the treatment effect, i.e. the change in a certain indicator as a result of the implementation of a certain function. Since this indicator characterises a certain component of the marketplace's business or business process to be improved, A/B testing metrics can be broken down into such groups:

– Business Outcome Metrics, such as GMV, Net Revenue, CLTV. They operate at the highest level of the metrics hierarchy, measure the impact of treatment on business KPIs; may be primary and/or guardrail;

– metrics for tracking the growth and sustainability of the marketplace, e.g., Total Number of Sellers, Number of New Sellers, Listing Count, Activation Rate;

– liquidity (health) and balance metrics, which qualitatively show the probability of transactions on the marketplace, e.g., Match Rate, Buyer to Seller Ratio, Fill Rate;

– metrics of providing sustainable financial state, i.e. monetization metrics, e.g., AOV, Take Rate, Purchase CR, CAC, CPA. These are shorter-term metrics that are much more often directly applicable as metrics for the success of AB tests;

– metrics of attracting/activating new users, e.g., CAC, CR, Number of New Sellers / Buyers / Suppliers; DAU/WAU/MAU;

– metrics for assessing changes in the effectiveness of targeted user actions – conversion metrics at different levels and stages: CR; CTR of elements (banners, buttons, links); Add-to-Cart Rate; Checkout Conversion Rate. Sometimes they are considered as indicators of user management effectiveness;

– metrics for evaluating marketplace participants specific to the evaluation of 2 marketplace sides, sellers and buyers. For example, for consumer side (buyers) such metrics are Success (Transaction Rate), AOV, Bounce Rate, Buyer Activation Rate, CTR, Purchase Frequency; for supply side (sellers) – Listing Count, Request Acceptance Rate, Seller Activation Rate, Unsubscription Rate etc. Such metrics as Match Rate and CAC can be considered as metrics of marketplace management quality;

– metrics for displaying actual user behaviour when interacting with the platform (Behavioural Metrics [82], e.g., CTR, Time on task, Task success rate, Time on Element, Scroll Depth), i.e. indicators that capture real user actions during their interaction with a product or platform;

– metrics for assessing user interest i by the nature and intensity of interaction with the product on the platform, i.e. Engagement Metrics [82], e.g., DAU/MAU, Time on Page, CTR, Bounce Rate);

– metrics of the actual attitude of users to the platform, the degree of satisfaction from cooperation with it, i.e. Attitudinal Metrics, e.g., NPS, CSAT, SUS etc. These are mostly secondary metrics that are assessed on a nominal, category or ordinal scale;

– metrics of sustainability of the platform's use by users – Retention Metrics [82], e.g., Retention Rate DX, Churn Rate, Repeat Purchase Rate influenced by NPS;

– content engagement metrics used to evaluate the effectiveness of design and content changes: Time on Page, Scroll Depth, Engagement Rate;

– metrics for tracking the efficiency of marketing campaigns through the evaluation of advertising and email campaigns, e.g., percentage of emails opened; percentage of unsubscribes, ROAS, CPA;

– Feature Adoption metrics [45], which evaluate the utility of introducing the new functionality by tracking the frequency of user interaction with it. For example, if the new

feature offered to marketplace sellers is the automatic creation of a basic store, the primary metrics for assessing the feasibility of its implementation in the A/B test may be Time to First Listing – the average time from registration to publication of the first product and First-Week Listing Rate – the share of sellers who added ≥1 product in the first 7 days, secondary metrics – the average time in the store setup wizard and the number of views/sales of products after auto-configuration, the retention rate of sellers for 7, 14 or 20 days, etc.;

– Data Quality / Sanity metrics [27], which allow to assess the correctness of the experiment setup and conduct, the reliability of the experiment results, e.g., too early test termination or misinterpretation of the results. There is a group of Sanity metrics [38], the constancy of which is an invariant of any A/B test, such as the invariance of the population during testing, the same size and homogeneity of the control and test groups, etc.;

– performance metrics – metrics of the platform's technical state and its impact on user experience, e.g., Error rate, % Broken Images or Links, Page Load Time. These metrics are operational; mainly session level, primarily used as guardrail metrics, rarely as primary or secondary metrics under certain experiment goals.

As it is shown in [73] after the analysis of publications on A/B testing metrics in recent years, the most common group of A/B metrics is Engagement Metrics (half of cases), which relate to the number of conversions, number of user sessions, time spent on the website, and metrics displaying the use of the app or website (e.g., number of posts rated, number of bookings made). The second largest group is click-through metrics (1/6 of cases: number of clicks, number of clicks per query, and good click-through rate). This is followed by monetisation-related metrics, i.e. revenue and costs (number of purchases, order value, revenue per email opened, and cost per ad) and group of technical performance metrics (app response time, bandwidth used, total latency, or audio playback latency). The remaining groups are metrics that track undesirable effects in A/B tests, such as bounce rate or number of cancelled subscriptions), views such as number of page /product views), and user feedback, such as number of customer complaints or literal feedback.

In case of A/B testing in marketplaces metrics may be classified by marketplace funnel categories (acquisition and marketing metrics → retention and engagement metrics → operational (performance) / quality / guardrail / sanity metrics → conversion metrics → monetisation metrics. Acquisition and marketing metrics (CAC, ROAS, CTR) reflect the expenses for involving and activating a user to the marketplace. Retention and engagement metrics (DAU/WAU/MAU, Retention Rate (7/14/30/60, Returning Users Rate, % Buyers with Repeat Purchase (2^{nd} purchase in another category), Engagement rate, Scroll Depth, Session Duration) determine customer long term value and cost of acquisition new users during significant periods of time. Operational (performance) / quality / guardrail / health /sanity metrics (Page Load Time, Support Tickets per Order, Return Rate, Fraud Rate and specific for marketplace liquidity/performance/health metrics such as Fill Rate, Buyer to Seller Ratio, Match rate, % Active Buyers/Sellers, Fulfilment Rate, Listing count / relevance) are not directly aligned to business goals and work on different funnel stages: these metrics support the reliability and convenience of user/platform experience. Conversion metrics (CR (for user level), Purchase rate (for item or session level), Add to Cart Rate, Cart Abandonment Rate, Checkout Rate measure the efficiency of transactional performance and movement from funnel top to funnel bottom. Monetisation metrics (GMV, Net Revenue, Take Rate, ARPU, and RPV) show the efficiency of marketplace business and management. Some metrics, such as CLTV or RPV can be considered as reflecting synergy of metrics of different categories (e.g., CLTV as function of ARPU (Monetization metric) and Retention (Retention and Engagement metric)).

As [8] suggests, conversion and monetization (revenue) metrics are the most essential in A/B testing of marketplaces. These metrics directly correlate with business goals and revenue generation. A higher CR indicates an increase in the number of users completing actions that contribute to business success, such as making a purchase or subscribing to the platform. This boosts revenue, customer acquisition and overall business growth. The most popular conversion metrics are: total CR (or purchase rate); Form Submission Rate, Sign-up Rate ad funnel conversion rates (Add-to-Cart Rate, Checkout Rate etc.,) or inversely metrics such as Cart Abandonment Rate. RPV is a key monetization metric reflecting the average amount of money generated by each visitor. J.Gallant [35] complements this list of metrics with the following: CTR, Bounce Rate (combined with Retention rate and Page Views), Scroll depth, Abandonment rate (rather combined with Funnel CRs), Session duration or Time on Page (in conjunction with other UX metrics), AOV, Churn rate and Net Revenue.

Metrics also are classified by the time needed for obtaining statistically proved results. As to experiment duration short and long terms are distinguished. Short-term metrics are metrics that provide immediate or near-term results (usually during or shortly after an experiment, mainly from 1 to 2 weeks). They can usually be measured directly during the experiment and allow for a quick assessment of the treatment effect (e.g., CTR, session level CR). Long-term metrics (21+) measure deferred outcomes that are difficult or impossible to measure during the experiment (e.g., CLTV, Net Revenue). They directly reflect the desired business outcomes and require the experiment design considering long-term impacts [2]. Metrics evaluating of which needs 3-4 weeks are assessed as medium term ones (e.g., CR).

If we treat marketplace metrics as indicators of the quality of the workflow elements aimed at achieving business goals, we can view them as branches of a tree of activities (functions) and define the hierarchy of these indicators according to the degree of their direct involvement in achieving the business results. According to the place of the metric node in this tree, a classification of metrics by functional significance can be proposed: highest, zero (0) rank metrics corresponding to key business parameters; first rank metrics – the main strategic metrics that directly support the achievement of the metrics of the previous level; second rank metrics that ensure the desired state of the metrics of the previous level, directly affecting their values; third rank metrics that support the metrics of the previous ranks by identifying factors that directly affect or may affect them; fourth rank metrics – operational metrics of direct measurement during the experiment, which are measured every day during the experiment and reflect the efficiency of individual operations on the marketplace.

0-Rank (top-level) metrics reflect the main business results and liquidity of the marketplace: GMV, Net Revenue, CLTV. These indicators are aggregated, their assessment requires a long accumulation of data, so they are practically not used as direct indicators of success (primary metrics); their change as a result of the implementation of the tested changes is assessed by extrapolating the results relating to the metrics of the following levels.

Rank 1 metrics are strategic metrics that actually and predictably support the achievement of the main results of the business, ensure sufficient monetisation, liquidity and user retention, such as CAC, ROAS, Buyers to Seller Ratio, Take Rate, APRU, RPV, CR, Retention/Churn Rate. These indicators are often used as success KPIs, some of them are long-term, and when used in the A/B test, are also recommended to be replaced by shorter-term correlated indicators of lower ranks.

Rank 2 metrics are supportive driver metrics that are directly related to the final results, i.e. to the processes that ensure the achievement of the required values of strategic metrics, and reflect intermediate results in achieving the final goals under the tested changes in input data. Examples of metrics at this level include: CR, AOV, Repeat Purchase rate,

ISSN 2522-9087 (Print)
ISSN 2523-434X (Online)

Match Rate, Retention D7 Rate, DAU/WAU/MAU, Fulfilment Rate, Utilisation Rate, etc.

Rank 3 metrics are intermediate (proxy) metrics that do not directly shape the final results, but provide guidance on what may affect these results. Examples of such metrics are Page Views per Session, CTR, Product Listing Rate, Activation Rate, Scroll Depth, etc.

Rank 4 metrics are indicators of the efficiency of individual transactions obtained by direct measuring the parameters of these transactions. These metrics enable monitoring the quality of service, delivery, order processing, interaction between users and sellers, as well as the operational management of these activities. These metrics are often used as secondary metrics (when selected as primary 2nd rank supporting metrics) to understand how specific operations (like delivery or seller interaction) impact overall business results, and as guardrail metrics. Examples: Request Processing Time, Seller Response Time, Failed Listings Rate, Error Rate per Item, Bounce Rate, CTR, etc.

Regardless of the function performed, to ensure a reliable assessment of A/B test results, a metric must meet a number of requirements first clearly formulated in [61]: directionality (interpretability), which means that metric improvement always results in improved user experience or business outcomes and is understandable and explainable to stakeholders; sensitivity, i.e. the ability to detect small changes if they indicate true differences; alignment, i.e. the relevance to the long-term goals of business and user and reliability, i.e. low variance and robustness to noise.

These metric parameters are strongly influenced by metric data and aggregation types, level of randomization and appropriate to them methods of evaluation and statistical confirmation. Accordingly, these parameters are to be considered when forming a set of metrics for OCE.

There are two main types of data, qualitative and quantitative [43]. Among qualitative data, there is discrete nominal, discrete ordinal (rank), discrete binary and ordinal binary data; these data are evaluated by verbal features. A discrete metric takes a finite (at least 2) number of values that do not have a natural order (website design; position of the call to action (CTA) button; colour, introduction of a new option). A discrete ordinal metric accepts a finite number of values (integers, categories), and the values have a natural order, but the interval between categories is not defined (user ratings (satisfaction level). A categorical binary metric can have only two possible category values: 0/1, yes/no, better/worse, black/white, or any other two options (clicks – whether a button was clicked or not). Categorical data is usually summarised by estimating shares (percentages) or, as they are called, proportions, or ratios (shares of users who completed a purchase or trip after registration, CR, CTR).

Quantitative data can be discrete or continuous; they can be quantified by counting (discrete) or measuring (continuous). Continuous variables can take on an infinite number of real values, while discrete variables can take on a finite number of real values within a given interval. These data can be evaluated on interval scales and scales of ratios (proportions). Quantitative measurement data are usually generalised using the statistical characteristics of the centre (mean, mode, median) and spread (variance, standard deviation) of a series of values (Table 2).

The type of data is important for the proper application of statistical methods for processing test results [67]. For nominal variables, mode is often the only meaningful measure of centre and trend; for ordinal variables, mode or median can be used, but not mean; for continuous and discrete variables, all three measures: mean, median and mode – can be used, and the variability and “noise” can be measured by variance [29].

Table 2 – Types of metrics data (built by the authors according to [43])

Features	Qualitative data		Quantitative data	
	Nominal / binary (nominal scale)	Ordinal (scale of categories)	Continuous (interval/ratio scale)	Discrete (interval/ratio scale)
Orderliness	None inherent order	Has a meaningful order, by certain feature or intensity of manifestation	Available, data takes arbitrary values within the range	Available, data takes discrete values within the range
Comparability, arithmetic operations	Identity comparison only (same/different)	Comparing for identity, ratios (bigger, smaller) and rank	Comparing for identity, establishing ratios (bigger than, less than, by how much), adding and subtracting	Comparison for identity, establishing a ratio (more, less, by how much), all arithmetic operations
Example	Clicks (yes/no), region	NPS	Retention/churn rate	Number of visitors Number of views
Distribution centre measures	Mode	Mode, median	Mode, median, mean	Mode, median, mean
Measures of dispersion	Quartiles (1st, 4th), percentiles	Quartiles, percentiles, minimum, maximum	Variance, standard deviation, range	Variance, standard deviation, range

The suitability and effectiveness of statistical methods for hypothesis testing in A/B testing also depend on the type of data in the metric scales [29, 85]. When choosing statistical characteristics of the centre of distribution, one should take into account their advantages and disadvantages:

– the average is more stable and understandable, more clearly defined and easy to calculate, but is dependent on outliers and may yield a value that is not in the distribution;

– the median is also easy to determine, does not depend on extreme values, can be used for categorical data, but does not take into account all observed values, requires ordering of data and intervals, and is not suitable for further mathematical transformations, which limits its use as a component of a composite metric;

– the mode is extremely fast to be determined from the distribution graph, shows the most frequent value in the distribution, even for qualitative data, does not depend on outliers, but is difficult to be determined for some distributions. It takes into account not the values from the observations, but their frequency, and is also unsuitable for further mathematical transformations, which also limits its use as a component of the composite metric.

When determining the metric data types for A/B test, it appeared to be insufficient to evaluate them as only qualitative/quantitative and discrete/continuous, since the metric parameters important for use in A/B testing, such as MDE, sensitivity, noise (variance), and time horizon, will be affected not only by the data categories themselves, but also by the way they are obtained. Thorough studying of marketplace metrics proved binary (Yes/No, e.g. click) not to be the only discrete ordinal metric applied in A/B testing. It is accompanied by discrete nominal (colour, CTA button design / placement), ordinal, called sometimes as rank (CSAT, 1-5) and count (number of clicks or page views). The continuous data type will be inherent in both metrics where the result on a continuous scale will be obtained by direct measurement or by adding continuous values (e.g., GMV), and as a result of recalculating binary or discrete measurement results, mainly as a result of adding and subsequently dividing homogeneous metrics. The metrics obtained in this case are considered to be of proportion, or

ISSN 2522-9087 (Print)
ISSN 2523-434X (Online)

ratio type, these values are dimensionless or estimated in percentages and reflect the share of the numerator in the denominator. Moreover, the proportion metrics display different essential for A/B testing features depending on their denominator and numerator: continuous-derived (ratio of continuous metrics or continuous and count ones); count-derived ratio of two count metrics, e.g., Match Rate, CTR, and binary-derived (ratio of sums binary outcome to sums of binary or count metrics, mainly total users, e.g., CR, Bounce Rate, Retention Rate). It's worth to note that continuous derived proportion metrics are mainly considered to be continuous by data type with aggregation type "Multiplicative" (Mult) if estimated in the numerator's units of measurement (e.g., ROAS, APRU, RPV) and proportion if measured in %, e.g., Take Rate. Some authors [16] call proportional the ratio of two count metrics while refer to binary-derived metrics as to ratios. Further, rather popular are so called composite metrics (scores combining several metrics, often weighted, e.g., Engagement index, UX Success Score, NPS). Metric data type constraints the range of further applied statistical methods and affect such metric parameters as noise and sensitivity [58]. For example, for continuous metrics the test significance is recommended to be evaluated by T-test (Student or Welch) while binary-derived ratio are tested by Welch's T-Test in case the metric is not heavily sparse (Table 3).

Table 3 – Metric data type impact on A/B testing attributes (summarized by authors based on data from [9, 16, 17, 18, 21, 22, 23, 24, 28, 58, 63, 85])

Metric data type	Way of receiving	Sensitivity (MDE)	Noise (variance)	Time horizon	Sample size	Statistical test	Example
Binary	Values 0/1 aggregated for all users	Medium	Low – Medium	Medium	Medium	Z-test for proportions χ^2 test	CTA clicks, Colour clicks
Discrete count based	Integer values aggregated for users	Medium	Low – Medium	Medium	Medium	Wilcoxon rank sum Poisson regression	DAU/MAU/WAU, Number of visits / Items purchased /New customers
Conti nuous (monetary)	Real number indicators for each / all users	Low – Medium	High	Long – very long	Large – Very large	Student's / Welch T-test, linear regression, bootstrap	GMV, RPU, AOV, APRU
Conti nuous (temporal)	Real number indicators for each / all users	Medium	Medium –High	Medium	Large	Welch T-test, Mann-Whitney U-test	Time on Site, Dwell Time, AVG Delivery Time / Deviation
Proportion (ratio) binary derived	Division of sums of binaries and counts	High (for big initial values)	Medium	Short – medium	Medium – Large	Z-test for proportions χ^2test	CTR, CR, Checkout rate, Add to Cart Rate
Proportion (ratio) count derived	Division of sums of counts	Medium	Medium-High	Short – Medium – Long	Medium – Large	Delta + T-Test Bootstrap Permutation	Retention Rate (Day 7, 14, 30), Buyer/Seller ratio, Fill Rate, Churn
Ordinal quantita tive	Aggregated rankings	Low – Medium	High	Long – Very long	Large-Very large	Mann-Whitney U-test	NPS

Besides the data type it is important to consider the method of aggregation of the measured set of values used to calculate the metric value (treatment effect). According to A. Deng [17, 21], for A/B tests, the aggregate metric sccheme can be defined as:

– Single average (Single Avg) [17], referred sometimes as Simple Average (Simple Avg) in which the metric serves as an estimate of the average value (mathematical expectation) of the general population and is calculated by averaging the measured indicator across users, sessions, or events (RPV, APRU, CTR etc.). The calculation formula shows that Simple Avg can be strongly distorted by outliers or heavy users, especially when distribution appears to be skewed;

– Double average (Double Avg) [17] – an arithmetic mean determined in relation to another average indicator due to preliminary averaging the within subgroups (e.g., per user, per product or per session) with further averaging across these groups. For example, if click-through rate is defined as the number of clicks per page view, it is a single average, but if it is calculated for each user and then taken as an arithmetic average over all users, it is a double average. Double Avg reduces bias from uneven user activity although may introduce additional variance due to the difference in subgroup sizes. The choice of the calculation method between single (simple) average and double average depends on the purpose of testing and the approach to evaluating its results;

– Weighted Average (Weighted Avg) [17] is an extended type of Double Avg, in which different weights are assigned to different experimental units when calculating the metric. For example, ROAS can be weighted by individual advertising campaigns, CR, AOV, CTR – by traffic volumes, % of active consumers – by regions, etc.;

– Conditional Average metric (Conditional Avg) is similar to average metric, except that the unit base is under a condition: aggregation is conditional on subgroup membership, e.g., RPV per Active User, AOV for converters, CSAT for return users [72];

– Percentile and, less often, Median [17]. These metrics manage especially well with measuring skewed distributions, such as Page Load Time or Delivery Delay, for which means are not representative. The percentile is better suited for assessing the tails of the distribution, median – for its centre; both of these metrics are popular for assessing performance/speed. Page load times of 75%, 95%, and even 99% are common percentile metrics.

This set of metric types is complemented by so called Ratio metrics [21], referred sometimes as proportions [16]. These metrics are defined as ratios of aggregated numerator and denominator sums (e.g., CR = total quantity of conversions/total visitors) and tend to be biased when the denominator varies significantly across groups. The data type of ratio metrics always is continuous though numerator and/or denominator data type can be binary (i.e. representing a binary outcome for each visitor in the experiment) which sometimes mislead to assessing them as binary [16]. However, in case both numerator and/or denominator data type are binary these metrics are easier to interpret and communicate; they follow a normal distribution; so some authors call them Binomial [2].

In our opinion, this classification does not take into account a number of metrics obtained by simply adding or multiplying / dividing measured indicators, which is inherent in, for example, such indicators as GMV or net income. Therefore, it is advisable to add two more categories to this list:

Simple Sum, when the metric is calculated as the sum of the values of a series (GMV = sum of transactions), data type is inherited from series;

Multiplicative (Mult), obtained by multiplying or dividing two metrics (Net Revenue = GMV×Take Rate). Obviously, depending on the calculation method, the same metric can be aggregated according to a different rule (Net Revenue = GMV – (expenses + taxes +

deductions)) and be of the Sum type. If metric value is obtained by dividing or multiplying several metrics it sometimes is suggested to have Index/Score type (e. g. Engagement score, NPS, UX index). The special type of ratio metrics is presented by metrics combining aggregations of two different events within a single metric [15], which in fact are the composite ratio metrics, e.g., Revenue per Add-to-Cart Click (Revenue / Add-to-Cart Clicks) instead of simple metric RPU. In case of such metrics the numerator and the denominator as separate simple metrics are to be applied in the same experiment.

The last two aggregation types are observed mainly for KPI metrics, which are low sensitive, strongly lagged, and prone to long-time effect. Actually, it limits usage of these North Star metrics in real A/B tests with substituting them with appropriate surrogate metrics.

In general, each aggregation type has its pros and cons, less or more significant for different test aims and conditions, business and organizational context (Table 4).

To our opinion, the choice of method of aggregation is to be a trade-off between the levels of satisfaction, possible duration and sample size of different to A/B test and characteristics of test metrics in every specific experiment. Thus, high-variance metrics, often arising from skewed or sparse data, reduce experiment sensitivity [18, 58], thus needing more test duration and sample sizes. Techniques like shrinkage estimation [26] and variance reduction [17, 28] improve detectability of treatment effects. This trade-off can be made by selection the metric with characteristics better for the experiment design.

The metric variance can be significantly reduced with accompanying reduction of noise, MDE (increasing the sensitivity), sample size and test duration if logarithmic transformation of mean, CUPED or shrinkage Adjustment are applied, e.g., CUPED removing variants caused by covariates provides 40-60% reduction in sample size and test duration [3, 24, 44]. CUPED performance is highly dependent on the metric; to perform well it needs pre-experiment data our users and a metric with high variance across the user population with covariate evenly distributed between treatment and control [10]. Due to the fact that these methods of variance reduction are commonly used in blogs and industry reports the terms Log-transformed Mean (averaging after Log(x) transforming the per-user values) and CUPED adjusted mean (mean after removing variance explained by covariates) sometimes occur.

The choice of the metric preferential aggregation is heavily influenced by metric data type (Table 5).

The choice of aggregation method significantly affects the bias, sensitivity, noise, variance, and interpretability of experimental results, which directly impacts statistical power, experimental duration, and business relevance [21]. E.g., simple averages have less variance but can be biased if heavy users predominate, while double averages can reduce bias via increasing variance. Thus, choosing an appropriate aggregation method requires an understanding of the underlying data distribution and the goals of the experiment. Choosing the right aggregation scheme, aligned with user behaviour patterns and business goals, increases the robustness of the metric, reduces bias, and increases the sensitivity. The use of additional variance reduction methods optimises the experimental design [17, 21, 26].

Moreover, changing the aggregation method can improve metric parameters, e.g., sensitivity. Usually A/B metrics are averaged across the randomization units. As [72] prove, using alternative methods to aggregate the metric can help increase its sensitivity. Thus, a ratio (proportion) metric showing the share of the units satisfying some condition is the easiest to move compared to the other types.

The simple average metric for measuring checkout conversion rate in a highly heterogeneous user population disproportionately weighs heavy users, thus potentially distorting the expected treatment effect. In contrast, the double average calculates the

conversion rate for each user and then averages it over the entire user population, reducing the bias from uneven distribution of user activity and outliers [21].

Table 4 –Impact of aggregation methods on metric characteristics (built by authors based on data from [3, 4, 9, 17, 18, 24, 26, 52, 56, 57, 58, 77])

Metric parameter	Aggregation type			
	Single Average	Double Average	Ratio (of Sums)	Percentile / Median
Metric variance	Low due to overall aggregation (in case of large sample size), but influenced by heavy users	Higher due to user-level noise	Low to medium, depending on the skewness: lower if dominated by heavy users	Low variance in the centre; unstable in the tails (e.g. 95th percentile)
Statistical power	High (due to low variance), but can cause misleading results	Lower if no variance reduction is applied (e.g., CUPED)	Often high, depending on the distribution of the denominator, does not always reflect impact at the user level	Less power to detect effects; more samples needed
Bias	Biased towards heavy users or high-frequency events	Low bias; better reflects average user-level behaviour	Can be biased if denominator is affected by treatment	Minimal bias for median; percentiles are robust but may ignore subtle effects
Noise/outliers susceptibility	Highly vulnerable, sensitive to large volumes or outliers	Less sensitive; individual effects are averaged	Moderate sensitivity; vulnerable if numerator or denominator has skewness/outliers	Robust to outliers (especially medians); extreme percentiles may be noisy
Stability and robustness	Unstable in in heavy-tailed distributions (if there is heterogeneity in usage)	More stable with heterogeneous user populations More reliable; less sensitive to changes in usage	Unstable if denominator is small or zero	Highly stable near median; less stable in extreme percentiles
Metric Noise	From moderate to high	High	Very high	Low for skewed distributions
Metric sensitivity (evaluated by MDE)	Moderate	Low	Moderate if base value is big enough, otherwise very low	High if outlier prone
Needed sample size	Baseline	Much bigger than baseline	Much more bigger than baseline	Smaller than baseline
Needed test duration	Baseline	Much bigger than baseline	Much more bigger than baseline	Baseline, or less
Alignment with business metrics	Good for general totals (e.g., GMV)	Good for user-level KPIs (RPV, APRU, CSAT)	Valid if both components reflect the business context	Good for user experience (e.g., delivery delay)
Ease of interpretation	Easy to calculate and explain but hides user variability and heterogeneity	Intuitive; reflects the impact on each user	Often confusing; risk of misinterpretation	Easy to understand though more difficult for statistical interpretation
Common cases of usage	Revenue metrics, RPV, total CR, Time on Site Averaging (by users	APRU, CR, Retention rate Averaging (by users, session etc.) all	ROAS, CTR, CTA Other ratio metrics Sum of every unit numerator values /	Page load time, delivery delay (median, percentile), user experience

	etc.) computed sum of all numerator values	computed metrics per unit (usually per user)	Sum of every unit denominator value;	metrics

Table 5 – Correspondence between metric data and aggregation types (built by authors based on data from [3, 4, 9, 17, 18, 23, 24, 26, 52, 56, 57, 58])

Data Type	Aggregate Metric	Examples
Binary	Sum of 0/1 values	Clicked CTA button (yes/no)
Discrete nominal (qualitative / quantitative)	Percentile, median	Website design; position of the CTA button; colour; Search algorithm
Discrete ordinal	Sum, percentile	Ratings; Degree of satisfaction (stars)
Discrete interval (Count)	Average count per user/session	Clicks, Searches, Orders, Page views
Continuous	Average of continuous variable (Simple Avg, Double Avg, Weighted Avg)	Revenue per user (RPU), Time on site, Scroll depth, Delivery time
Proportion (continuous derived)	Proportion of aggregated outcomes	Take Rate
Proportion (Binary-derived)	Proportion (e.g., CR = Conversioned Users / Users)	Conversion rate (CR), Bounce rate, Retention Rate
Ratio (Count-derived)	Proportion of aggregated outcomes: Sum(numerator)/Sum(denominator)	CTR = Clicks / Impressions, Fill rate, Match rate
Index / Composite	Avg score across users; Mult or proportion of aggregated outcomes	Engagement index, UX Success Score, NPS
Aggregated ordinal	Proportional or categorical summary	CSAT score (1–5), Likert scale responses

The statistical power of Conditional Avg metric is improved if the numerator and denominator are correlated. If the extreme case is to be monitored percentile metric (e.g., 99 Percentile of Time in App) is better choice for sensitivity. The application of transformation of metric aggregation method and variance reduction techniques is investigated in [84], authors of which report encouraging results by taking logarithm of the response metric and leveraging pre-experiment data as covariates in the regression.

Choosing the right metric provides a balance between statistical validity, business relevance, and test duration. This balance is supported by metric characteristics such as inherent noise, sensitivity, and minimum test duration. Metrics with low noise, high sensitivity, and alignment with objectives typically lead to faster and more efficient test execution; they converge to the mean faster and achieve statistical confidence more quickly with smaller sample sizes.

In A/B testing, the sensitivity of a metric is the ability of the metric to detect even small real changes (effects) that occur under the test. In [24, 59] sensitivity is mostly assessed by the Minimum Detectable Effect(MDE) – the smallest difference between the control and treatment variants the test can detect with sufficient statistical power (usually ≥80%) and significance. Sensitivity is influenced by three core components: uplift size; variance size σ^2 and sample size N [58]. As MDE $\sim\sigma/N^{1/2}$, smaller differences between treatment and control require higher sensitivity; high metric variance dilutes sensitivity; larger samples increase sensitivity, but at operational cost (longer tests). Calculation MDE = $f(\sigma/N^{1/2})$ was grounded on historical data or based on the desired statistical confidence and test power and sample size N. Metrics are considered sensitive if they detect small changes across population: the smaller the changes are, the more sensitive the metric is (Table 6).

Most sensitive metrics (Table 6) provide a low MDE (1-3%), that is, the test can detect even very small improvements. Sensitivity <2% is hardly achieved in practice and needs extremely high traffic and very large samples. Less sensitive metrics have a higher MDE, i.e.

a larger effect is needed to prove the results; sensitivity of 3-5% is most achievable in real A/B tests. Moreover, though binary-derived ratio metrics of conversion expose the highest sensitivity the true treatment effect is difficult to detect at session level if initial values are small. That's why they are advised to be used at user and higher randomization levels [51].

Table 6 – Qualitative and quantitative assessment of the metrics' sensitivity (based on data from [55, 58, 66, 80])

Sensitivity	Statistical power	MDE, % change from the base value	Metric functionality for AB test assessment
Very high sensitivity, almost unachievable	≥95 %	~1%	Enables detecting even small changes, often due to low variance and large amounts of data
High sensitivity	≥90 %	~1-3%	
Medium sensitivity	~80%	~3-5%	Detects moderate changes, requires a moderate sample size or test time for being statistically significant
Low sensitivity	≤75%	~5-10%	Responds only to large changes, often due to high variability or challenges in data acquisition
Very low sensitivity	<70%	>10%	

Online controlled experiments require a sufficient number of users to identify a statistically significant result. If the number of users involved in the experiment is measured in thousands or tens of thousands (which is typical for Ukrainian marketplaces) rather than in hundreds of thousands or millions (as it natural for A/B testing in Amazon, eBay, Etsy, Microsoft, Netflix etc.), larger effects easier to detect are needed, versus the small effects the large sites are concerned about. For example, to increase the sensitivity of an experiment (i.e., to reduce the size of the detectable effect) by a factor of 10, e.g., from 5% to 0,5%, the number of users needs to be 10^2=100 times bigger [58].

As small treatment effects (uplift of 1-5%) are common [58], the metrics are the better the less ATE they measure, thus they are to be capable of maximizing the statistical power. A common technique for improving sensitivity is transforming the metric Y into Y* of lower variance [59], which, with other factors being equal, leads to an estimate of τ with lower variance. The metrics in AB tests are numerous and differ in their properties, making it almost impossible to define a single universal transformation. Popular transformations for metric sensitivity improvement consist in binarizing count and continuous metrics as well as in changing the way aggregation is performed [23, 49]. If metric tends to be skewed and prone to outliers, trimming (capping, i.e. excluding extreme metrics over the threshold before calculating the mean) and winzorization (replacement of the excluded values by the most extreme retained ones or by using percentiles of data) techniques [83, 86] can be applied. Popular approaches to increase the sensitivity include different variants of pre- [19, 24], in- [20] and postexperimentation processing such as sample stratification and variance reduction techniques, e.g., CUPED [19, 24], triggering [53], their combination [25, 75] and shrinkage [26], widely applied in Microsoft, Netflix [87], Booking [44], Meta [60, 88], Linkedin [48], Uber [40], Airbnb [25], DoorDash [89], Faire [75]. Ratio metrics and per-user aggregation are also proposed [58] as such that can reduce the variance arising from skewed distribution.

The noise level determines the robustness of a metric to random influences. It correlates with the metric's variance across the sample, which is caused by the nature of the metric [17]. Variance (dispersion) shows the degree to which a data point deviates from the mean value in the dataset, thereby reflecting the spread of the data relative to the centre (sample span). A metric is highly dispersed if the data from individual users vary widely, such

ISSN 2522-9087 (Print)
ISSN 2523-434X (Online)

as RPV or AOV. A large dispersion makes it difficult to detect small changes trustworthily, requiring a larger sample size or longer test duration to detect meaningful differences. The test result turns out to be more "noisy".

The noise (level) of a metric is the degree of its stability when measured on different samples of users, i.e. a kind of estimate of the probability of random fluctuations of the metric regardless of the changes under test, under the influence of random factors. The noise is estimated by the variance D of the metric values for individual users (or other indicators taken as the unit of the experiment) compared to the average value m for the sample: $CV = D^{1/2}/m = \sigma/m$, where CV – coefficient of variation (estimated noise level of the metric); m – estimated average value of the metric for the sample; D, σ – variance and standard deviation, respectively. A high variance is typical for metrics with a large spread of values across the sample, such as RPV or AOV. The larger the CV is, the greater the noise level, the less stable and suitable the metric is for testing, and the longer the minimum acceptable test duration to detect a true treatment effect (Table 7).

Table 7 – Estimating the noise level of a metric by CV values [24, 58, 81]

CV	Noise level	Interpretation (metric stability, risks of using in a test)
< 0,3	Low	Metric is stable, recommended for AV testing
0,3 – 1,0	Medium	Acceptable for tests, stability decreases, predicted test duration increases with increasing CV
> 1,0	High	Hazard: test is long or unstable. If the time of the experiment is limited, or if there is a high probability of long-term effects of factors outside the experiment, replacement with a surrogate metric is recommended

Low-noise metrics mainly include metrics that are calculated as the ratio of the sum of the results of actions to the total number of actions, for example, conversion rates CR, CTR, Onboarding completion rate (Number of users who completed onboarding / Total number of users who started onboarding). Metrics with a high level of noise and variance include mainly metrics with high variability across users and random outliers (e.g., RPV, LTV, Session duration, NPS). These metrics converge to the mean more slowly and generally reach statistical significance with larger sample sizes and longer terms of testing.

For different categories of metrics the characteristics differ in significance. Primary metrics are intended to measure the main outcome or expected impact of an experiment and are usually taken as criteria for launching a feature. They are be aligned with long-term business goals or user satisfaction as well as with the proposed change, be statistically sensitive (have small minimum detectable effect (MDE) under needed statistical power), stable, interpretable and reproducible. Proxy metrics are to agree with primary, reflect the funnel stage, be more short term and sensible. Guardrail metrics ensure non-deteriorating of product, platform or UX value under treatment effect. These metrics should not worsen beyond acceptable thresholds (e.g. "should not increase by more than X% or cross the value Y") regardless the treatment result for OCE. They are to be understandable, truthworthy, business result sensitive, measure negative side effects such as latency, reliability, or churn. The possibility of using metrics as surrogates should be confirmed by analytical and correlation analysis or historical A/B tests, as well as by the reaction of guardrails and sanity checks to their changes.

Thus, the metric is selected in accordance with the purpose of testing from a set of metrics that reflect changes in the business indicators aimed at improving. Based on the existing work of many researchers, we can assume that the main parameters that are to be taken into account when selecting metrics from the set of metrics that can assess the success of the experiment are the following parameters: sensitivity, estimated through the MDE, noise

and latency. The more the sensitivity and the less the noise are, the more short duration of experiment and less sample size may be enough for A/B test. Metrics are considered as short-term, if the needed for MDE achievement test duration does not exceed 1 week, medium-term ones – 1-2,5 weeks and long term if are longer than 3 weeks. These limits are to some extend floatable because of minimum recommended test duration (1-2 weeks) and metrics with test valid term 30 days and more are considered very long. Proposed in [61] principles of A/B metric design, though qualitative, allow formulating several practical tips for metric selection, namely: metric's movement should clearly reflects the desired change; as metric sensitivity increases while getting closer to the reflected user action, it is advisable to favour intermediate metrics to detect smaller, meaningful changes in user behaviour; for noisy metrics stabilization should be applied (e.g., by CUPED or capping outliers); for keeping the directionality short-term surrogates (proxies) are to be verified for correlation with long-term top-level or primary metrics with casual interpretation of the dependencies observed.

The actual sensitivity of the metric also depends on the consistency of the metric's change in the same direction as useful ATE, so the concept of relative sensitivity is introduced. Statistical power reflects the likelihood that a movement can be detected. The movement probability shows how likely an indicator movement in needed direction is to occur. The relative sensitivity is expertly evaluated by a combination of metric's ability to detect a change in the feature under test and the probability of the metric's movement in the direction of the desired change in the business indicators that the experiment is aimed at improving.

Satisfying the directionality demand helps to exclude from the metric set so-called "vanity metrics". These are metrics that, while often visually striking, do not strictly align with business goals. They can create a false sense of success and potentially confuse decision makers [8]. Examples of such metrics are: Page Views (needs to have specified the reasons of user engagement); Time on Site (does not always correlate with a better user experience, as an increase in session duration can be the result of more effort to find the relevant information). The application of these metrics requires identifying the context of user behaviour and satisfaction and even such popular metric as Bounce Rate does not provide understanding the reasons of leaving the page visited as well as insights in user behaviour or engagement beyond this bounced page. These metrics can be used only if applied in one set other metrics in which specify the real sense of vanity metrics change. While designing the metric portfolio the role of the metric in the experiment should be considered and revised in case the metric's characteristics don't correspond to the chosen role in the experiment (Table 8). When forming a set of metrics for A/B testing, it is necessary to ensure its functional completeness and consistency, i.e. the set should include metrics that perform different roles in the experiment (are key indicators or primary metrics or their surrogates, auxiliary explanatory metrics and security metrics) and do not contradict with each other in terms of reflecting the direction of change of the primary metrics; furthermore, the roles of the metrics should not be duplicated (i.e. the set of metrics should not be redundant) and the metrics should not conflict in achieving the result (i.e. it is unacceptable to deteriorate one metric at the expense of another).

Accordingly, metrics with a significant or strong pairwise correlation of any sign should not be included in the set. This suggests that the role of the metric and its rank should be determined before including it in the set for evaluating the results of an AB experiment, since evaluation by surrogate metrics of lower ranks is usually more short-term.

Table 8 – Properties of metrics with different role in AB testing (compiled by authors based on data from [1, 2, 33, 35, 53, 54, 58])

Metric's role	Functions	Properties	Consequences of significant change	Evaluation approach
Primary (KPI)	Defining the success criteria Detecting the results of experiment (achieving business goals), providing the determining of MDE, experiment duration and sample size	Aligned to the long-term goals of business or user Sensitive (small minimum detectable effect (MDE)) Have small variance, be stable, interpretable and repeatable Be measurable and actionable	Ship / not ship the tested feature or continue the experiment iterating its plan	P-value based hypothesis testing (frequenciest or Bayesian)
Driver (secondary, surrogate)	Ensuring a comprehensive understanding of the experiment's impact Predicting the experiment's results in shorter time and with less sample sizes Revealing probable false positive (FP) in primary metric assessment Substitute the long-term KPI	Short-term, more sensitive / less noisy than primary ones Aligned to the funnel stage Correlate with primary metrics, be directional to them and surrogate Significant impact on primary metric without changes in secondary suggest FP for primary	Stop or continue the experiment iterating its plan Reveal a reason of treatment effect and predict experiment's results Reconsider evaluation of experiment result	P-value based hypothesis testing (frequenciest or Bayesian)
Guardrail	Monitoring the health, stability, and overall integrity of the system Detecting unintended negative impacts Used as stop signals, trigger alarms or filters for feature adoption	Directional and interpretable Comprehensive, measure negative effects (e.g., churn, latency, errors, fraud rate, unsubscribe rate) Sensitive, interpretable and visible; protective and non-negotiable	Stop the experiment or revert the best point achieved if degrades Reconsider new feature Simultaneously implement new feature and protective changes	Binary Threshold based (must not change in any direction or change not more than A %)
Exploratory	Generating insights or hypotheses Supporting hypotheses built on results obtained for primary / secondary metrics, study consequences and discover new ideas Redesigning metric portfolio, layer or role in experiment Debugging	Not enough sensitive and directional for the decision making or evaluating experiment results, but providing new hypotheses and experiment goals Mainly long-term, noisy and ambiguous, may require additional validation May be quantitative or qualitative, proxy-based May be proxy, but become primary / secondary in further experiments	Investigate for new features or metrics Redesign metric portfolio or change their design or method of evaluating depending on the direction of change Evaluate the correlation with other metrics from portfolio, implement or change method of its optimization Change the metric layer	Descriptive Qualitative or quantitative Assessed after experiment or on historical data Regression modelled Survive modelled
Composite	Combining several separate metrics into one to capture a diverse view of success	Reflect complex treatment effect on different significant features of the product or process Often have smaller sensitivity	Ship / not ship the tested feature Redesign the metric if too long experiment time	P-value based hypothesis testing Modelled

For the correct, fast and reasonable selection of metrics for evaluating a particular experiment with further specifying the parameters of the latter, it is desirable to form a database of metrics described through: their possible roles (functions) in the experiment; characteristics affecting the parameters of the experiment (data type; sensitivity (estimated through statistical parameters, MDE and direction of movement); noise, time horizon (latency); connection with business indicators; possible surrogates; correlated metrics, etc. The quantitative assessment of the metric's sensitivity depends on many factors, some of which are uncontrollable, so to justify the choice, we propose the "Relative Sensitivity Score"

parameter, which takes into account both MDE, variance and probability of moving in the correct direction and is assessed in points, from 1 (very low sensitivity) to 5 (very high). Rating on an ordinal scale allows one to choose the best metric from several ones in terms of sensitivity. The examples of ordinal estimates of these parameters for some metrics are shown in the Table 9.

Table 9 – Parameters of A/B testing metrics (built by authors based on data from [3, 4, 20, 21, 23, 42, 44, 52, 58, 64, 68, 72, 87])

Metric Abbr	Data type	Aggregation type	Sensitivity (assessed by CV, MDE, stat. power)	Relative Sensitivity Score	Noise	Time horizon (Latency, Test Duration), Days
GMV	Continuous (monetary)	Aggregated (Sum)	Low	Medium – Low	High	Long – Very Long (30+)
Net Revenue	Continuous (monetary)	Composite (Sum)	Low	Medium – Low	High	Long – Very Long (30 – 30+)
CLTV	Continuous (monetary)	Composite (Sum), Single Avg	Medium – Low	Low	High – Medium	Long – Very Long(30 – 30+)
Net Promoter Score (NPS)	Ordinal (0–10)	Composite, ratio (count based)	Low	Low – Medium	High	Long (21 – 30) – Very Long 31+)
Revenue per User (RPU)	Continuous (monetary)	Single Avg, Weighted Avg	Medium	Low – Medium	Medium – High	Medium (14 – 21)
Average Order Value (AOV)	Continuous (monetary)	Single Avg, Weighted Avg	Medium	Low – Medium	Medium – High	Medium (7 – 21)
Conversion Rate (CR)	Proportion Ratio of sums, binary derived	Single /Double Avg,	High	Medium – High	Low – Medium	Short – Medium (7 – 14)
Add to Cart Rate	Proportion Ratio of sums, binary derived	Single /Double Avg,	High (for big initial values)	Medium – High	Low	Short (7 – 14)
Checkout rate	Proportion Ratio of sums, binary derived	Single /Double Avg,	High (for big initial values)	Medium – High	Low	Short (7 – 14)
Bounce Rate	Proportion Ratio of sums, count derived	Single /Double Avg	Low – Medium	Low – Medium	Low – Medium	Short – Medium (7 – 21)
Click-Through Rate (CTR)	Proportion Ratio of sums, binary derived	Single Avg,	High	High	Low – Medium	Short (5 – 7)
Retention Rate (Day 30)	Proportion Ratio of sums, count derived	Single Avg,	Low – Medium	Low – Medium	Medium – High	Long

Retention Rate (Day 7)	Proportion Ratio of sums, count derived	Single Avg, Double Avg	Medium	Medium	Medium – High	Short – Medium
Returning Users Rate	Proportion Ratio of sums, count derived	Single Avg, Double Avg	Medium	Low – Medium	Medium – High	Medium – Long (10 – 30)
% Buyers with Repeat Purchase	Proportion Ratio of sums, count derived	Single Avg, Double Avg	Low	Low – Medium	High	Long – Very Long (21 – 30+)
Cart Abandonment Rate	Proportion Ratio of sums, count derived	Single Avg	Medium	Low – Medium	Medium – High	Short – Medium (7 – 21)
Churn Rate	Proportion Ratio of sums, count derived	Single Avg, Double Avg	Low – Medium	Low	Medium	Long
Time on Site	Continuous (temporal)	Single Avg, Double /Weighted Avg,	Medium	Low – Medium	Medium – High	Short – Medium (14 – 21)
Session Length	Continuous (temporal)	Single Avg, Double /Weighted Avg,	Medium	Low – Medium	Medium – High	Short – Medium (14 – 21)
Scroll Depth	Continuous	Single Avg, Percentile	High	Medium	Medium	Short (7 – 14)
Task Completion Rate	Proportion Ratio of sums, count derived	Single Avg, Double Avg	High	Medium – High	Low	Short – Medium (7 – 14)
DAU/WAU/MAU	Count	Aggregated (Sum)	Medium	Low – Medium	Low – Medium	Short – Long (14 – 30+)
Page Load Time	Continuous (temporal)	Median	Medium	Medium	Low – Medium	Short – Long (14 – 30+)
Match Rate	Proportion Ratio of sums, count derived	Ratio	Medium	Medium	Medium	Medium – Long (10 – 30)
Fill Rate	Proportion Ratio of sums, count derived	Ratio	Medium	Medium	Medium	Medium – Long (10 – 30)
Supply Coverage	Proportion Ratio of sums, continuous derived	Ratio	Medium	Low – Medium	Medium	Medium (10 – 21)
Fulfilment Rate	Proportion Ratio of sums, binary derived	Ratio	Medium	Medium	Medium	Medium (10 – 21)
Utilization Rate	Proportion Ratio of	Ratio	Medium	Low – medium	Medium	Medium (10 – 21)

	sums, continuous derived					

To categorize A/B testing metrics, we use ordinal descriptors: from very low to very high, values of which are derived from experimental and theoretical assumptions. To assess the statistical behaviour of marketplace A/B testing metrics, we applied ordinal qualitative estimates – low, medium, and high – for three key properties: noise, sensitivity, and latency (i.e. expected test duration). These estimates are based on a generalisation of empirical studies [3, 4, 44, 52, 64, 68] and theoretical developments under absence of publicly available unified quantitative estimates. E.g., [20] presents quantitative results (test durations, variances, MDEs etc.) of 141 experiment on Airbnb's experimentation platform; results and parameters of online marketplace pricing A/B test ran from September 28, 2018 to October 31, 2018 on a population of 4557234 listings, are given in [42]; data from five A/B tests concerned the manipulation of user interface elements and search engine algorithm are reflected in [64]. The diversity of organisations, platforms, and methods of metrics assessment makes it difficult to obtain accurate quantitative estimates; however, aggregating the results of a number of studies allows for a qualitative ordering of metrics by key indicators [20, 21, 23, 42, 58, 72, 87, 90]. Moreover, appealing to the sensitivity dependence on experiment duration, we sometimes assessed sensitivity comparing test durations with different test durations and similar sample sizes (traffic) [42].

Noise is assessed by the normalized metric stability indicator, the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean. According to [81] and taking into account the specifics of AB testing metrics we propose treating results with a CV < 0,5 as low noise, with a CV of 0,5 – 1,5 as medium noise, and with a CV > 1,5 as high noise. This framework correlates with [58], which document typical high variance in skewed metrics. Binary or binary-derived proportion metrics (e.g., CR, CTR-clicked at the user level) tend to have low to medium noise (CV ≈ 0,2 – 0,6), while count-derived ratio or long-tail continuous metrics (e.g., RPU, GMV) typically present high noise (CV > 1,5 -2) [24, 58].

Sensitivity in terms of MDE inversely corresponds to noise and reflects the MDE at a fixed power and confidence thresholds: high CV leads to a high MDE and diminished power. According to this, we classify marketplace metrics, such as: CR (low noise, high sensitivity), CTR (medium noise), RPU (high noise, low sensitivity) and GMV (very high noise), – providing a systematic framework for metric selection in experimental design. Metrics with high variance or rare events require much larger differences in the value of the treatment effect, resulting in low sensitivity. The actual sensitivity of the metric also depends from the stability of the metric's directionality reflected via probability of metric movement and is estimated from the historical data for each experimentation platform and organisation. We have based the estimates of Relative Sensitivity Score on the data from numerous studies and our own research on the casual hierarchical relationships of the metrics.

The value of the latency (by which we mean the duration of the test required to achieve sufficient statistical power) depends on the sensitivity and noise of the metric, the initial value of the metric (in the case of ratio metrics), and the sample size which is limited by the amount of traffic on the website. We classify the expected duration as very short (up to 1 week), short (up to 1,5 weeks), medium (1,5 – 3 weeks), long (3-4 weeks), and very long (4+ weeks) based on typical sample size requirements and event frequency reported in industry case studies [3, 4, 20, 39, 44, 52, 64, 68]. For example, short to medium duration tests are mainly enough for conversion rate (CR) metrics due to limited variance and frequent

user events, while continuous revenue-based or combined metrics (e.g., GMV, ARPU) often require longer periods to stabilise noisy estimates. Considering the heterogeneity of platforms and data patterns, these ordinal mappings seem to be rather guidelines than fixed thresholds and should be tested using variance and power analysis on a historical basis in each experimental system.

As metric's data and aggregation type significantly impact on metric parameters, when selecting metrics matching the experiment goals we should consider these types as well as possibilities of metric improvement by changing the aggregation type.

As Table 9 shows, binary-derived ratio metrics expose the best combination of characteristics, but, as it is shown in [58, 69] the design and application of these metrics can generate many problems. Firstly, with a low baseline, a few percent increase may be immeasurable, so the use of these metrics at the session level is not recommended. Secondly, the change in the indicator depends on changes both in the numerator and denominator of the ratio, which makes the interpretation of the change in the metric ambiguous for all cases except those with an unchanging denominator. Therefore, it is advisable to handle the numerator and denominator as debugging metrics for a rate metric; also when choosing a rate metric as a success one, the metric relatively more stable denominator is to be chosen [23]. Moreover, the relative sensitivity and bias of the ratio metric strongly depend on the type of aggregation: Single Avg (computed as ratio of averages, i.e. as total sum of numerator values for all users divided by total sum of denominator) vs. Double Avg (computed as average of ratios, i.e. average of individual user ratios). They differ primarily in the way in which users are weighted. In the case of the average of ratios, users have the same weight in the final metric because each user has a single value for the ratio [17, 18, 23]. For the ratio of averages, the weight is proportional to the denominator value for each user, e.g., a user with more page views will have a greater impact on final metric, e.g. CTR. Because of the difference in weights, Simple Avg (average ratio) ratio appears to give more importance to heavy users, while Double Avg (average of ratios) optimisation treats all users as equally valuable.

In order to design a valid metric set the causal interdependences between metrics of different ranks, e.g., key business and proxy metric, is to be clearly outlined. To achieve that a metric hierarchical tree reflecting the hierarchical casual dependencies and pathways of influence between metrics of different ranks is to be built. Arranging the metrics structure of any business, for example, a marketplace is convenient to do by constructing a tree-like graph of metrics, which, along with hierarchical relationships, will also display functional relationships of metrics belonging to the same level of the tree. In addition, the presence of casual links of a lower-level metric with several higher-level metrics leads to the fact that a child metric may have several parent metrics. Building such a metrics tree requires:

– developing a comprehensive set of indicators reflecting business goals and the means to achieve them;

– identifying the metrics reflecting the state of these indicators, sufficiently sensitive to their change and measurable/determinable during A/B testing in a reasonable time;

– identifying the highest (zero rank) metrics directly reflecting the achievement of strategic business goals and adopting them as the top of the highest level. The peculiarity of building a hierarchical metrics tree is the presence of several key metrics (interrelated roots of the tree), the evolution of which reflects the main goal of each business – sustainable growth of the business value;

– determining the functional significance (ranks) of selected metrics and functional interdependencies between them;

– figuring out the distribution of selected metrics by tree levels;

– highlighting the hierarchical relationships between metrics of different levels, pointing out the potential surrogacy of lower level metrics in regard to influenced metrics;

– revealing interrelations derived correlation dependencies between metrics, which may result from both casual dependencies and subjection to the influence of the same factors.

The metrics tree developed on the basis of the ranking the e-marketplace metrics (Table 1) is shown in Fig. 1. It is reasonable to use this tree when forming a set of metrics for a particular A/B test by selecting a subtree with a root node corresponding to the primary North Star metric of test success to identify a set of secondary, exploration and guardrail metrics and their surrogates.

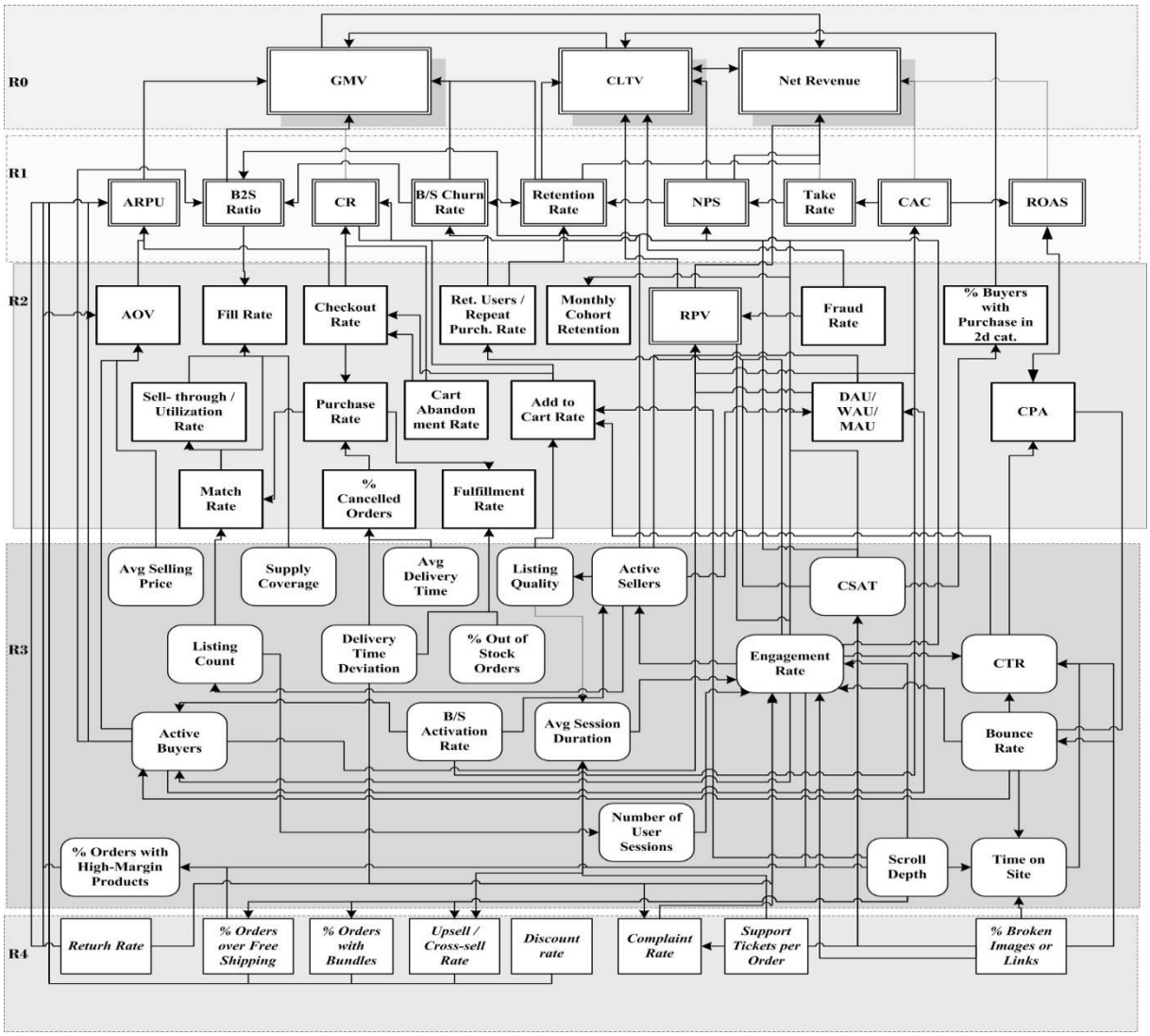

Figure 1 – Hierarchical tree of metrics for A/B testing of marketplaces

(developed by authors)

The constructed hierarchy is a sketch of a logical model of the metrics' database, which can be transformed into a real database scheme after determining the attributes of metrics involved in the search algorithm. Aggregating the Tables 1, 2 and Figure 1 provides a catalogue of metrics, which can be used for metric set design. When choosing a metric from the catalogue, a matrix of indicative assessment of the metric parameters (Table 10) is

suggested to be used for selecting the metric from a set of metrics of the same functional purpose.

In this matrix score 1 is assigned to very bad values (low for sensitivity and relative sensitivity, high for variance and noise, very long for test duration); 2 – to bad ones (medium – low l for sensitivity and relative sensitivity, medium – high for variance and noise, long for test duration); 3 – to medium, 4 – to good (medium – high for sensitivity and relative sensitivity, medium low for variance and noise, long for test duration) and 5 – to very good ones (high, low, very short), where ordinal estimates correspond to Table 9 and are figured out after data published in numerous research reports of such companies as Airbnb, Microsoft, Amazon, Netflix etc. and articles after the results of such studies [3, 4, 20, 21, 23, 42, 44, 52, 58, 68, 72, 87, 90].

Table 10 – Matrix of comparative assessment of metric parameters (developed by authors)

Parameter	Sensitivity (assessed by MDE)	Variance (CV)	Relative Sensitivity (MDE, statistical power × movement probability)	Noise	Latency (Time horizon)
Score	5	5	5	5	5
	4	4	4	4	4
	3	3	3	3	3
	2	2	2	2	2
	1	1	1	1	1

Thus, the total score serves as an average indicator of the metric's quality: metrics with all parameters equal to 1 (the bottom row of the matrix) are unacceptable (total score = 5), 5 is ideal (total score = 25), which is unattainable in practice. Estimated values of the CR metric parameters for sufficiently large baseline values and sample sizes are highlighted in grey. After filling in the matrix for each candidate metric, they can be compared and the best one selected. If there is a shortfall in a particular parameter, the possibility of applying improvement methods [19, 20, 23, 24, 25, 26, 49, 53, 75, 83, 86] should be examined.

Relying on the built tree of metrics and the assessment of their parameters, the following workflow can be proposed to build an efficient initial set of primary, secondary and guardrail metrics for a particular A/B test:

1) define the strategic objective of the experiment;

2) define the desired test duration and sample size limits;

3) define the North Star metric appropriate to experiment goal and tested feature;

4) define the business funnel stage and category of object or process the treatment effect on which is to be evaluated by metric;

5) activate the metric tree and select from chosen category the KPI metrics which can be used as primary metrics considering directionality and interpretation;

6) choose from metric catalogue a single or a multiple metric as OEC;

7) estimate category of metric test durations and compare it to the desired one;

8) select metrics with appropriate latency or look for more short-term surrogates from the metric catalogue and metric tree;

9) activate the metric catalogue and define functionally linked to KPIs' candidates for secondary driver, exploration and guardrail metrics;

10) identify composite metrics, decompose them if possible referring to calculation formula; substitute complex metrics by their constituents to avoid unambiguity of interpretation;

11) estimate the metrics' parameters: data type, aggregation type, sensitivity (MDE, statistical power), movement probability, noise etc. given in catalogue. Power estimation

should be MDE based; movement probability assessed by one of the methods proposed in [72, 92] applied to historical data. Relative sensitivity score is to be assessed by identifying the metric sensitivity parameters (MDE, statistical power ×movement probability);

12) step-by-step check if metrics' parameters match to test duration for planned statistical power and uplift; use check results for data type and aggregation type as gates;

13) select metrics most appropriate to test conditions. If latency and sensitivity / noise parameters are unacceptable, define the set of surrogate metrics;

14) repeat analysis of metric parameters for each surrogate metric taking into account possibilities and method of parameter improvement;

15) identify correlated metrics in each category (driver, exploration etc.) and compare metrics' parameters within each category; the matrix of parameter score may be used;

16) if possible, apply Pareto optimization to metric set;

17) assess possible improvement of metric parameters by different methods and revise the matrix of parameter score;

18) exclude the worst (by parameter score) strongly correlated metrics, and select one (up to two if correlation is weak) best metric for each category. Identify methods for each metric parameters improvement;

19) validate metrics set according MECE principle (mutually exclusive and collectively exhaustive) [71]. Though this principle is recommended for validating the metric decomposition we suggest it can be applied for validating the metric set considering KPIs;

20) check the interpretations of metrics movement in certain directions for contradictions concerning impact of movement directions on KPI change. If they exist, construct a combo metric, e.g., as a weighted sum if the chosen metrics;

21) examine movement probability for each selected candidate experimentally by AA test. If confirmed, define the final metric set. In case confirmation failed revise the metric set.

A slightly simplified algorithm for creating such a set of metrics is shown in Figure 2.

It is obvious that the formation of such a set requires valid data to confirm the possibility of using surrogate metrics as key indicator substitutes, which requires identification of casual and correlation dependencies between metrics derived from the hierarchical tree of metrics and historical data. In addition, it is necessary to identify metrics whose changes allow for ambiguous interpretation and formulate restrictions on their use.

Conclusions from this research and prospects for further developments in this area. By analysing theoretical studies and previously accumulated experimental data, we have expanded the field of classification features of metrics for their step-by-step selection in a particular experiment, established quantitative estimates of sensitivity and noise parameters for many metrics, and evaluated their correspondence tor the duration of the experiment. The study of functional relationships between metrics gave the possibility of ranking them by functional significance, with further building the hierarchical tree of metrics and justifying recommendations for selecting a set of possible surrogate metrics for higher-ranked metrics. Together with the results of the metrics parameter estimation, this allowed us to develop an algorithm for selecting a set of metrics for testing in a particular experiment.

ISSN 2522-9087 (Print)
ISSN 2523-434X (Online)

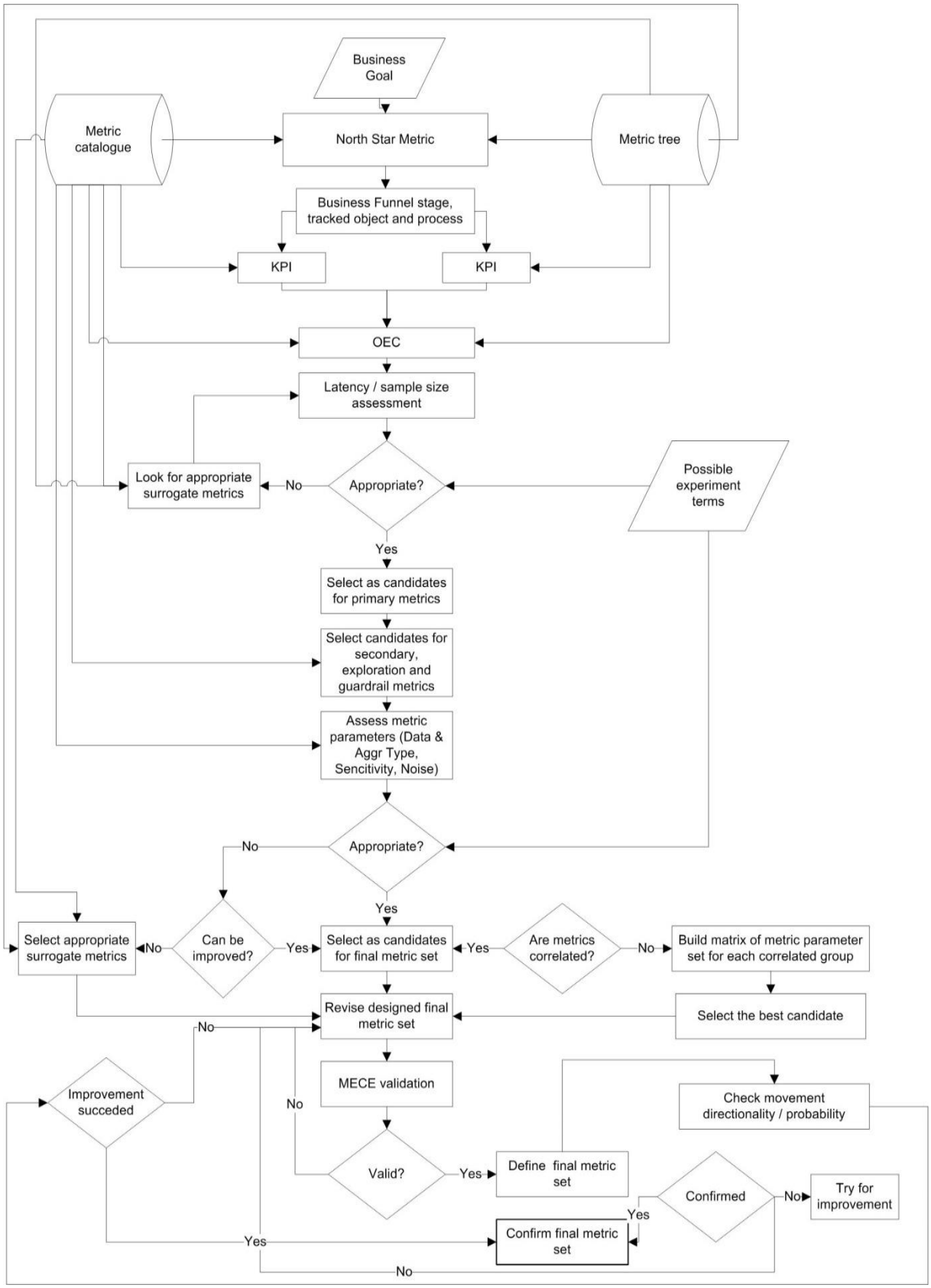

Figure 2 — A simplified algorithm for creating the initial set of A/B test metrics (developed by authors)

Further research should be devoted to refining quantitative estimates of metrics parameters, expanding the scope of research on the correlation of metrics and the interaction of their parameters, evaluating the effectiveness of using certain metrics for testing specific features, and creating a database of metrics for AB testing based on the results of these studies.

1. AB Smartly. (2024). Creating an experiment. Available at: https://docs.absmartly.com/docs/web-console-docs/creating-an-experiment/ (accessed 12.06.2025).
2. AB Smartly. (2024). Understanding experimentation metrics. Available at: https://docs.absmartly.com/docs/web-console-docs/understanding-experimentation-metrics/ (accessed 12.06.2025).
3. Airbnb Engineering. (2024). Experiments at Airbnb. Available at: https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7 (accessed 12.06.2025).
4. Airbnb Engineering. (n.d.). The Airbnb tech blog. Available at: https://medium.com/airbnb-engineering (accessed 12.06.2025).
5. Anderson, Y. (2025). How to measure eCommerce success: Key online marketplace metrics. Available at: https://www.cs-cart.com/blog/how-to-measure-your-marketplace-success-the-key-ecommerce-marketplace-metrics/ (accessed 12.06.2025).
6. Ankargren, S., Frånberg, M., & Schultzberg, M. (2024). It's about time: What A/B test metrics estimate. arXiv Preprint. DOI: https://doi.org/10.48550/arXiv.2411.06150.
7. Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. NBER Working Paper No. 26463. Available at: https://www.nber.org/papers/w26463 (accessed 12.06.2025).
8. AWA Digital. (2024). A/B testing metrics that truly matter. Available at: https://www.awa-digital.com/blog/ab-testing-metrics/ (accessed 12.06.2025).
9. Bakshy, E., Eckles, D., & Bernstein, M. (2014). Designing and deploying online field experiments. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14). DOI: https://doi.org/10.1145/2566486.2567967.
10. Berk, M. (2021, February 18). How to double A/B testing speed with CUPED. Medium. Available at: https://medium.com/data-science/how-to-double-a-b-testing-speed-with-cuped-f80460825a90 (accessed 12.06.2025).
11. Bibaut, A., Chou, W., Ejdemyr, S., & Kallus, N. (2024). Learning the covariance of treatment effects across many weak experiments. In Proceedings of the 30th ACM SIGKDD Conference on *Knowledge Discovery and Data Mining (KDD '24)*. DOI: https://doi.org/10.1145/3637528.3672034.
12. Bytom, J. (2025). A/B testing upsell strategies: Key metrics to track. Available at: https://ultralabs.digital/blogs/ecom-insights/ab-testing-upsell-strategies-key-metrics-to-track (accessed 12.06.2025).
13. Chandar, P., Thomas, B., Maystre, L., Pappu, V., Sanchis-Ojeda, R., Wu, T., Carterette, B., Lalmas, M., & Jebara, T. (2022). Using survival models to estimate user engagement in online experiments. In Proceedings of the ACM Web Conference 2022 (pp. 3186–3195). DOI: https://doi.org/10.1145/3485447.3512038.
14. Courthoud, M. (2022). Understanding CUPED. Towards Data Science. Available at: https://towardsdatascience.com/understandingcuped-a822523641af (accessed 12.06.2025).
15. Datsenko, M. (2025). Using ratio metrics for enhanced experimentation insights. Optimizely Blog. Available at: https://www.optimizely.com/insights/blog/ratio-metrics-in-experimentation/ (accessed 12.06.2025).
16. Deb, A., & Bhattacharya, S. (2018). Under the hood of Uber's experimentation platform. Uber Engineering Blog. Available at: https://www.uber.com/en-SE/blog/xp/ (accessed 12.06.2025).
17. Deng, A. (2023). Chapter 8: Statistical analysis of A/B tests. In Causal Inference for Experimentation. Available at: https://alexdeng.github.io/causal/abstats.html (accessed 12.06.2025).

18. Deng, A. (2023). Chapter 10: Improving metric sensitivity. In Causal Inference for Experimentation. Available at: https://alexdeng.github.io/causal/sensitivity.html (accessed 12.06.2025).
19. Deng, A., Bhatt, S., & Xu, Y. (2023). From augmentation to decomposition: A new look at CUPED in 2023. arXiv Preprint. Available at: https://alexdeng.github.io/public/files/CUPED_2023_Metric_Decomp.pdf (accessed 12.06.2025).
20. Deng, A., Du, M., Matlin, A., & Zhang, Q. (2023). Variance reduction using in-experiment data: Efficient and targeted online measurement for sparse and delayed outcomes. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). DOI: https://doi.org/10.1145/3580305.3599928.
21. Deng, A., Hagar, L., Stevens, N., Xifara, T., & Gandhi, A. (2024). Metric decomposition in A/B tests. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4885–4895). Available at: https://dl.acm.org/doi/10.1145/3637528.3671556 (accessed 12.06.2025).
22. Deng, A., Knoblich, U., & Lu, J. (2018). Applying the delta method in metric analytics: A practical guide with novel ideas. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 233–242). DOI: https://doi.org/10.1145/3219819.3219828.
23. Deng, A., & Shi, X. (2016). Data-driven metric development for online controlled experiments: Seven lessons learned. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 77–86). DOI: https://doi.org/10.1145/2939672.2939700.
24. Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (pp. 123–132). Available at: https://exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf (accessed 12.06.2025).
25. Deng, A., Yuan, L.-H., Kanai, N., & Salama-Manteau, A. (2023). Zero to hero: Exploiting null effects to achieve variance reduction in experiments with one-sided triggering. In Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM ’23), 823–831. DOI: https://doi.org/10.48550/arXiv.2112.13299.
26. Dimmery, D., Bakshy, E., & Sekhon, J. (2019). Shrinkage estimators in online experiments. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2914–2922). DOI: https://doi.org/10.1145/3292500.3330771.
27. Dmitriev, P., Gupta, S., Kim, D. W., & Vaz, G. (2017). A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1427–1436). DOI: https://doi.org/10.1145/3097983.3098024.
28. Dmitriev, P., & Wu, X. (2016). Measuring metrics. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (pp. 429–437). DOI: https://doi.org/10.1145/2983323.2983356.
29. Donges, N. (2024). Types of data in statistics: A guide. Built In. Available at: https://builtin.com/data-science/data-types-statistics (accessed 12.06.2025).
30. Drutsa, A., Ufliand, A., & Gusev, G. (2015). Practical aspects of sensitivity in online experimentation with user engagement metrics. In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (pp. 763–772). DOI: https://doi.org/10.1145/2806416.2806496.
31. Duan, W., Shan, B., & Zhang, C. (2021). Online experimentation with surrogate metrics: Guidelines and a case study. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM ’21) (pp. 193–201). DOI: https://doi.org/10.1145/3437963.3441737.
32. Fabijan, A., Dmitriev, P., Olsson, H. H., Bosch, J., Vermeer, L., & Lewis, D. (2019). Three key checklists and remedies for trustworthy analysis of online controlled experiments at scale. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice

(ICSE-SEIP) (pp. 1–10). Available at: https://exp-platform.com/Documents/2019%20FabijanDmitrievOlssonBoschVermeerLewis_Three-Key-Checklists_ICSE_SEIP.pdf (accessed 12.06.2025).
33. FasterCapital. (n.d.). The difference between primary and secondary metrics and how to measure them. Available at: https://fastercapital.com/topics/the-difference-between-primary-and-secondary-metrics-and-how-to-measure-them.html (accessed 12.06.2025).
34. Fleexy.dev. (2024). 20 key metrics to scale your online marketplace. Available at: https://fleexy.dev/blog/20-key-metrics-to-scale-your-online-marketplace/ (accessed 12.06.2025).
35. Gallant, J. (n.d.). 10 A/B testing metrics to analyze results and measure success. Unbounce. Available at: https://unbounce.com/a-b-testing/metrics-kpis/ (accessed 12.06.2025).
36. Goel, I. (2024). Three kinds of metrics: The success, the guardrail, and the diagnostic. Available at: https://vwo.com/stats-blog/three-kinds-of-metrics-the-success-the-guardrail-and-the-diagnostic/ (accessed 12.06.2025).
37. Gupta, A. (2025). How to choose the right metrics to evaluate experiments. Available at: https://www.news.aakashg.com/p/metrics-experiments (accessed 12.06.2025).
38. Gupta, P. (2024). Udacity A/B testing course lecture transcripts: Choosing and characterizing metrics (Chapter 3). Available at: https://medium.com/@prannerta100/udacity-a-b-testing-course-lecture-transcripts-choosing-and-characterizing-metrics-chapter-3-4affbcdcd05a (accessed 12.06.2025).
39. Gupta, S., Kohavi, R., Tang, D., & Xu, Y. (2019). Top challenges from the first Practical Online Controlled Experiments Summit. ACM SIGKDD Explorations Newsletter, 21(1), 20–35. DOI: https://doi.org/10.1145/3331651.3331655.
40. Harinen, T., & Li, B. (2019). Using causal inference to improve the Uber user experience. Uber Engineering Blog. Available at: https://www.uber.com/en-UA/blog/causal-inference-at-uber/ (accessed 12.06.2025).
41. Hohnhold, H., O'Brien, D., & Tang, D. (2015). Focusing on the long-term: It's good for users and business. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DOI: https://doi.org/10.1145/2783258.2788583.
42. Holtz, D., Lobel, R., Liskovich, I., & Aral, S. (2020). Reducing bias from interference in online marketplace pricing experiments. DOI: https://doi.org/10.48550/arXiv.2004.12489.
43. Jaberi, M. H. (2024). Understanding types of variables in data science: Numerical and categorical (nominal, ordinal, continuous, and discrete). Medium. Available at: https://medium.com/@jaberi.mohamedhabib/understanding-types-of-variables-in-data-science-numerical-and-categorical-nominal-ordinal-02d3bb60fd55 (accessed 12.06.2025)
44. Jackson, S. (2018). How Booking.com increases the power of online experiments with CUPED. Booking.com Data Science Blog. Available at: https://booking.ai/how-booking-com-increases-the-power-of-onlineexperiments-with-cuped-995d186fff1d (accessed 12.06.2025).
45. Jahin, A. (2024). How behavioral data enhances quantitative user feedback. Available at: https://www.wudpecker.io/blog/how-behavioral-data-enhances-quantitative-user-feedback (accessed 12.06.2025).
46. Jeunen, O., Baweja, S., Pokharna, N., & Ustimenko, A. (2024). Powerful A/B-testing metrics and where to find them. In Proceedings of the 18th ACM Conference on Recommender Systems. DOI: https://doi.org/10.48550/arXiv.2407.20665.
47. Jiang, L., Xia, T., Townsend, J., Jin, J., Siegel, H., Ramamurthy, V., & Qin, W. (2021). Why tenant-randomized A/B test is challenging and tenant-pairing may not work. Microsoft Research. Available at: https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/why-tenant-randomized-a-b-test-is-challenging-and-tenant-pairing-may-not-work/ (accessed 12.06.2025).
48. Jin, Y., & Ba, S. (2022). Towards optimal variance reduction in online controlled experiments. DOI: https://doi.org/10.48550/arXiv.2110.13406.

49. Kharitonov, E., Drutsa, A., & Serdyukov, P. (2017). Learning sensitive combinations of A/B test metrics. In Proceedings of the 10th ACM International Conference on Web Search and Data Mining (pp. 651–659). DOI: https://doi.org/10.1145/3018661.3018708.
50. Kim, A. (2024). Experimentation metrics: Deciding what to measure. Available at: https://www.geteppo.com/blog/experimentation-metrics-deciding-what-to-measure (accessed 12.06.2025).
51. Kohavi, R. (2025). My thoughts on A/B testing in low traffic websites. LinkedIn. Available at: https://www.linkedin.com/posts/ronnyk_5-ways-to-ab-test-on-low-traffic-websites-activity-7064415275315576833-y9EX/ (accessed 12.06.2025).
52. Kohavi, R. (n.d.). The Experimentation Platform (ExP). Available at: https://exp-platform.com/ (accessed 12.06.2025).
53. Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 786–794). DOI: https://doi.org/10.1145/2339530.2339653.
54. Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014). Seven rules of thumb for web site experimenters. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1857–1866). DOI: https://doi.org/10.1145/2623330.2623341.
55. Kohavi, R., Henne, R. M., & Sommerfield, D. (2007). Practical guide to controlled experiments on the web: Listen to your customers not to the hippo. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07) (pp. 959–967). DOI: https://doi.org/10.1145/1281192.1281295.
56. Kohavi, R., & Longbotham, R. (2023). Online controlled experiments and A/B tests. In D. Phung, G. I. Webb, & C. Sammut (Eds.), Encyclopedia of Machine Learning and Data Science (pp. 1–13). Springer. DOI: https://doi.org/10.1007/978-1-4899-7502-7_891-2.
57. Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181. Available at: http://www.exp-platform.com/Pages/hippo_long.aspx (accessed 12.06.2025).
58. Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press. DOI: https://doi.org/10.1017/9781108653985.
59. Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., & Stevens, N. T. (2024). Statistical challenges in online controlled experiments: A review of A/B testing methodology. The American Statistician, 78(2), 135–149. DOI: https://doi.org/10.1080/00031305.2023.2257237.
60. Liou, K., & Taylor, S. J. (2020). Variance-weighted estimators to improve sensitivity in online experiments. In Proceedings of the 21st ACM Conference on Economics and Computation (pp. 837–850). DOI: https://doi.org/10.1145/3391403.3399542.
61. Machmouchi, W., & Buscher, G. (2016). Principles for the design of online A/B metrics. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 589–590). DOI: https://doi.org/10.1145/2911451.292673126731.
62. Machmouchi, W., Hassan, A., Zitouni, I., & Buscher, G. (2017). Beyond success rate: Utility as a search quality metric for online experiments. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17) (pp. 757–765). ACM. DOI: https://doi.org/10.1145/3132847.3132850.
63. Mair, M., Kattwinkel, M., Jakoby, O., & Hartig, F. (2020). The MDD concept for establishing trust in non-significant results – A critical review. Environmental Toxicology and Chemistry, 39(11), 2283–2294. DOI: https://doi.org/10.1002/etc.4847.
64. Mandić, M., Gregurec, I., & Vujović, U. (2023). Measuring the effectiveness of online sales by conducting A/B testing. Market-*Tržište, 35*(2), 223–249. http://dx.doi.org/10.22598/mt/2023.35.2.223
65. Microsoft Research. (2022). STEDII properties of a good metric. Available at: https://www.microsoft.com/en-us/research/articles/stedii-properties-of-a-good-metric// (accessed 12.06.2025).

66. Miller, E. (n.d.). Evan's awesome A/B tools. Available at: https://www.evanmiller.org/ab-testing/ (accessed 12.06.2025).
67. Mishra, P., Pandey, C. M., Singh, U., & Gupta, A. (2018). Scales of measurement and presentation of statistical data. Annals of Cardiac Anaesthesia, 21(4), 419–422. DOI: https://doi.org/10.4103/aca.ACA_131_18.
68. Mueller, M. (2024). Meta-experiments: Improving experimentation through experimentation. Booking.ai. Available at: https://booking.ai/meta-experiments-improving-experimentation-through-experimentation-6bdee314c512 (accessed 12.06.2025).
69. Nie, K., Kong, Y., Yuan, T. T., & Burke, P. B. (2020). Dealing with ratio metrics in A/B testing at the presence of intra-user correlation and segments. In Web Information Systems Engineering – WISE 2020 (pp. 563–577). DOI: https://doi.org/10.1007/978-3-030-62008-0_39.
70. Prasad, N. (2024, March 21). Step-by-step walkthrough to A/B testing fundamentals. Medium. Available at: https://medium.com/learning-data/step-by-step-walkthrough-to-a-b-testing-fundamentals-0d8ba67be113 (accessed 12.06.2025).
71. Productfolio. (2025). MECE principle. Available at: https://productfolio.com/mece-principal/ (accessed 12.06.2025).
72. Qin, W., Machmouchi, W., & Martins, A. M. M. (2021). Beyond power analysis: Metric sensitivity analysis in A/B tests. Microsoft Research. Available at: https://www.microsoft.com/en-us/research/articles/beyond-power-analysis-metric-sensitivity-in-a-b-tests/ (accessed 12.06.2025).
73. Quin, F., Weyns, D., Galster, M., & Costa Silva, C. (2024). A/B testing: A systematic literature review. Journal of Systems and Software, 211. DOI: https://doi.org/10.1016/j.jss.2024.112011.
74. Rachitsky, L. (2021). Choosing your North Star Metric. Future. Available at: https://future.com/north-star-metrics/ (accessed 12.06.2025).
75. Ramirez Ritchie, E., & Wong, K. B. (2021). How to speed up your A/B test: Outlier capping and CUPED. Faire Engineering Blog. Available at: https://craft.faire.com/how-to-speed-up-your-a-b-test-outlier-capping-and-cuped-8c9df21c76b (accessed 12.06.2025).
76. Rodden, K., Hutchinson, H., & Fu, X. (2010). Measuring the user experience on a large scale: User-centered metrics for web applications. In Proceedings of the 28th International Conference on Human Factors in Computing Systems (pp. 2395–2398). DOI: https://doi.org/10.1145/1753326.1753687.
77. Saleh, K. (2024). How long should you run an A/B test for and how to calculate duration of A/B test. Invesp CRO Blog. Available at: https://www.invespcro.com/blog/how-long-should-you-run-an-ab-test-for/ (accessed 12.06.2025).
78. Salesforce. (2025). A/B testing metrics for B2C commerce. Available at: https://help.salesforce.com/s/articleView?id=cc.b2c_ab_testing_metrics.htm&type=5 (accessed 12.06.2025).
79. Schultzberg, M., Ankargren, S., & Frånberg, M. (2024). Risk-aware product decisions in A/B tests with multiple metrics. arXiv Preprint. DOI: https://doi.org/10.48550/arXiv.2402.11609.
80. Siroker, D., & Koomen, P. (2013). A/B Testing: The Most Powerful Way to Turn Clicks Into Customers. Wiley.
81. Six Sigma US. (2024). Coefficient of variation: Mastering relative variability in statistics. Available at: https://www.6sigma.us/six-sigma-in-focus/coefficient-of-variation/ (accessed 12.06.2025).
82. Sullivan, K. (2023). Mastering UX metrics: A comprehensive guide to understanding user behavior and design impact. Available at: https://medium.com/design-bootcamp/mastering-ux-metrics-a-comprehensive-guide-to-understanding-user-behavior-and-design-impact-4f7a9063ad6e (accessed 12.06.2025).
83. Trenque, J. (2024). Handling outliers with Winsorization in A/B testing. Kameleoon Help Center. Available at: https://help.kameleoon.com/handling-outliers-with-winsorization-in-ab-testing/ (accessed 12.06.2025).
84. Wang, D., Goldberg, P., Burke, P., & Bhoite, D. (2018). Designing and analyzing A/B tests in an online marketplace. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1447–1452). DOI: https://doi.org/10.1109/ICDMW.2018.00206.

85. Wei, M. (2022). How to select the right statistical tests for different A/B metrics. Towards Data Science. Available at: https://towardsdatascience.com/how-to-select-the-right-statistical-tests-for-different-a-b-metrics-c8a1865851e (accessed 12.06.2025).
86. Wilcox, R. (2005). Trimming and Winsorization. In Encyclopedia of Biostatistics. John Wiley & Sons, Ltd. DOI: https://doi.org/10.1002/0470011815.b2a15165.
87. Xie, H., & Aurisset, J. (2016). Improving the sensitivity of online controlled experiments: Case studies at Netflix. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 645–654. DOI: https://doi.org/10.1145/2939672.2939733.
88. Zhang, C., Coey, D., Goldman, M., & Karrer, B. (2021). Regression adjustment with synthetic controls in online experiments. Meta Research. Available at: https://research.facebook.com/publications/regression-adjustment-with-synthetic-controls-in-online-experiments/ (accessed 12.06.2025).
89. Zhang, J., & Tang, Y. (2020). Improving online experiment capacity by 4X with parallelization and increased sensitivity. DoorDash Engineering Blog. Available at: https://careersatdoordash.com/blog/improving-experiment-capacity-by-4x/ (accessed 12.06.2025).
90. Zhang, V., Zhao, M., Dimakopoulou, M., Le, A., & Kallus, N. (2023). Evaluating the surrogate index as a decision-making tool using 200 A/B tests at Netflix. DOI: https://doi.org/10.48550/arXiv.2311.11922.
91. Zhao, Z., Chen, M., Matheson, D., & Stone, M. (2016). Online experimentation diagnosis and troubleshooting beyond AA validation. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). (pp. 498–507). DOI: https://doi.org/10.1109/DSAA.2016.61.
92. Zito, A., Greaves, D., Soriano, J., & Richardson, L. (2025). Pareto optimal proxy metrics. Applied Stochastic Models in Business and Industry, 41(2). DOI: https://doi.org/10.1002/asmb.70003.

Гліненко Л.К., *канд. техн. наук, доцент, доцент кафедри ЕЗІКТ, Національний університет «Львівська політехніка» (Львів, Україна)*

Дайновський Ю.А., *д-р екон. наук, професор, професор кафедри маркетингу, Львівський торговельно-економічний університет (Львів, Україна)*

Занічковський Ю.В., *аспірант кафедри ЕЗІКТ, Національний університет «Львівська політехніка» (Львів, Україна)*

Контекстно-орієнтований вибір метрик в A/B тестуванні.

Робота присвячена вдосконаленню методологічної основи вибору метрик для A/B-тестування за рахунок розширення набору атрибутів метрик, що застосовуються у якості критеріїв вибору метрик для A/B тесту. Досліджено вплив типу даних та схеми агрегації метрики на її ефективність для A/B тестування, запропоновано використовувати ці параметри при первинному відборі метрик. За результатами узагальнення результатів експериментальних досліджень визначені порядкові значення ключових для A/B тестування параметрів метрик, таких як чутливість, зашумленість і необхідна тривалість експерименту. Побудовано ієрархічну модель взаємозалежностей між метриками і визначено процедуру формування набору конкретних метрик для A/B тесту на основі цієї моделі. Кінцевим результатом дослідження є формалізація ранніх етапів планування A/B-тестування з побудовою відповідного алгоритму.

Ключові слова: A/B *тестування*, *метрика*, *дерево метрик*, *чутливість*, *зашумленість*.

Received: 25 May 2025
Accepted: 30 July 2025
Published: 28 Desember 2025