JRSSEM 2026, Vol. 5, no. 12, July
E-ISSN: 2807 - 6311
P-ISSN: 2807 - 6494

System Shift Risk, Measurement Error, and the Limits of Non-Experimental Marketing Measurement: An Exploratory Validation Using RCT Benchmarks

Raymond Rubianto Tjandrawinata
School of Bioscience, Technology and Innovation Atma Jaya Catholic University of Indonesia Jakarta, Indonesia
Email: Raytjan@yahoo.com

Keywords:
advertising measurement;
causal inference;
randomized controlled trials;
double machine learning;
propensity score matching;
measurement error;
System Shift;
feedback maturity;
marketing analytics;
exploratory validation

Abstract

The increasing complexity of digital advertising environments has created significant challenges in ensuring the accuracy and reliability of marketing measurement. Although randomized controlled trials (RCTs) are considered the benchmark for causal evaluation, their implementation is often constrained by cost, operational limitations, and managerial requirements, leading organizations to rely on non-experimental measurement approaches that may generate substantial estimation errors. This study aims to explore the role of system-level factors in explaining measurement error and to validate the System Shift Framework as a diagnostic approach for understanding the reliability of non-experimental marketing measurement. This research employs an exploratory quantitative design using six aggregate-coded marketing cases and compares non-experimental lift estimates with RCT benchmarks. Data analysis was conducted using descriptive statistics, correlation analysis, linear regression, regularized logistic regression, model comparison, and cluster analysis. The findings indicate that Strategy Quality and Feedback Maturity demonstrate the strongest associations with lower measurement error, higher accuracy, and improved measurement reliability, whereas the aggregate System Shift Risk Score shows limited explanatory power for continuous measurement error. The results suggest that measurement reliability is influenced not only by statistical methods but also by an organization's capability to interpret, evaluate, and correct analytical outcomes. This study concludes that the System Shift Framework provides a complementary diagnostic perspective for improving marketing measurement practices by integrating causal inference with systems thinking.

INTRODUCTION

Marketing measurement exists within a familiar yet unresolved paradox. Firms increasingly demand precise evidence of advertising effectiveness; however, the environments in which such effects are estimated are noisy, targeted, dynamic, and structurally biased (Doucette, 2026; Iashchenko et al., 2025; Ngo & Vu, 2026; Srinivas, 2025; Theodorakopoulos et al., 2025; Woodside & Kumar, 2023). Randomized controlled trials (RCTs) have become an important benchmark in advertising effectiveness research, but they are not always feasible in routine managerial practice (Ambrosi et al., 2025; Braga et al., 2025; R. Han et al., 2024; Moschonis et al., 2023; Siopis et al., 2023; Tankha et al., 2024). Therefore, non-experimental methods remain necessary, particularly when experiments are costly, impractical, or operationally constrained. The question is not whether such methods should be used—they

must be used. The more challenging question is when their outputs should be considered reliable (T. Han & Li, 2022; Kasneci et al., 2023; Liu et al., 2023; Wang et al., 2024; Zhou et al., 2024).

Prior advertising measurement research has demonstrated that non-experimental estimates may diverge substantially from randomized benchmarks. Gordon, Zettelmeyer, Bhargava, and Chapsky (2019), using large-scale field experiments conducted on Facebook, demonstrate that observational estimates may substantially overstate advertising effects. Gordon, Moakler, and Zettelmeyer (2023) extend this concern through a large-scale evaluation of non-experimental approaches, showing that even sophisticated observational methods require careful validation against experimental evidence. Lewis and Rao (2015) further highlight the challenge by demonstrating that measuring returns on advertising is statistically difficult, even within experimental contexts, because advertising effects may be small relative to background noise and because effect heterogeneity can make precise inference challenging.

This paper advances a complementary argument. Measurement error in advertising effectiveness should not be interpreted solely as a technical limitation of an estimator. Instead, it may also represent a signal of system-level measurement failure. A marketing measurement system is not merely a statistical pipeline; it is a configuration of funnel stage, data structure, targeting logic, estimation method, decision pressure, feedback maturity, strategic clarity, organizational adaptation, and structural constraints. In other words, the estimator operates within a broader system. Sometimes the system supports valid inference; sometimes it subtly undermines it.

The global advertising industry has experienced substantial growth, increasing the demand for reliable measurement systems. According to industry reports, global digital advertising expenditure has surpassed hundreds of billions of dollars annually and continues to represent the dominant share of total advertising investment. This expansion has increased managerial dependence on attribution models, observational analytics, and machine-learning-based estimation techniques to evaluate marketing returns. Nevertheless, previous evidence indicates that these approaches may generate misleading conclusions when they fail to adequately address selection bias, consumer heterogeneity, targeting mechanisms, and unobserved confounding factors. The challenge is particularly significant because organizations frequently make high-value strategic decisions based on estimated advertising impacts that may not accurately represent incremental consumer responses. Therefore, improving the reliability of marketing measurement systems has become a critical concern for both academic research and managerial practice.

A central challenge in contemporary marketing analytics is the divergence between non-experimental measurement outcomes and experimental benchmarks. Randomized controlled trials are widely recognized as the strongest approach for establishing causal relationships because random assignment reduces systematic differences between treatment and control groups. However, implementing RCTs in real-world marketing environments is often constrained by financial costs, operational limitations, ethical considerations, and organizational complexity. Consequently, firms increasingly rely on observational approaches such as propensity score matching, synthetic control methods, attribution models, and machine-learning-based estimation techniques. Although these approaches provide practical alternatives, previous research has demonstrated that their estimates may substantially differ

from experimental benchmarks. Gordon et al. (2019), for example, found that various advertising measurement approaches evaluated through large-scale field experiments could produce inconsistent estimates of advertising effectiveness. Similarly, Gordon et al. (2023) showed that non-experimental methods require careful validation because even advanced analytical approaches may generate substantial measurement errors when compared with experimental evidence.

Previous studies have extensively examined methodological challenges in advertising effectiveness measurement. Lewis and Rao (2015) highlighted that measuring advertising returns is statistically difficult because advertising effects are often relatively small compared with existing market noise, resulting in substantial uncertainty in estimation. Gordon et al. (2019) investigated alternative advertising measurement approaches and emphasized the importance of experimental validation for assessing estimation accuracy. Gordon et al. (2023) further examined the performance of non-experimental approaches and demonstrated that sophisticated observational methods do not automatically guarantee reliable causal estimates. Other studies have contributed methodological solutions through causal inference approaches, including propensity score methods (Rosenbaum & Rubin, 1983), synthetic control techniques (Abadie et al., 2010), and double/debiased machine learning frameworks (Chernozhukov et al., 2018). Although these studies have significantly advanced statistical methodologies, they primarily focus on improving estimation procedures rather than examining the broader organizational and systemic conditions that influence measurement reliability.

Despite extensive methodological development, an important research gap remains regarding why similar analytical methods may produce different levels of reliability across organizational contexts. Existing marketing measurement literature predominantly evaluates the performance of specific estimation techniques, assuming that measurement errors primarily originate from statistical limitations or model selection issues. However, marketing measurement systems operate within broader organizational structures involving strategic decision-making processes, data quality management, feedback mechanisms, organizational learning capabilities, and managerial interpretation. A technically sophisticated model may still produce unreliable outcomes when the surrounding measurement system lacks strategic clarity or adaptive capability. Therefore, measurement error should not only be understood as a statistical problem but also as a potential indicator of system-level weaknesses affecting how data are generated, analyzed, interpreted, and applied.

Addressing this gap is increasingly urgent because inaccurate marketing measurement can lead organizations to allocate resources inefficiently, scale ineffective campaigns, and underestimate strategic risks. In highly competitive digital markets, firms must make rapid investment decisions based on analytical insights, while inaccurate measurement outcomes may create significant financial and strategic consequences. The challenge becomes more complex as marketing technologies continue to evolve, increasing the volume and complexity of consumer data while simultaneously introducing new sources of bias. Consequently, organizations require a more comprehensive perspective that integrates causal inference methods with system-level diagnostics to determine when measurement outputs can be trusted. Understanding the conditions that enhance or reduce measurement reliability is therefore essential for developing more responsible and effective marketing analytics practices.

To address this research gap, this study introduces the System Shift Framework as a

diagnostic perspective for explaining why marketing measurement systems produce reliable estimates in certain contexts but substantial errors in others. Unlike previous studies that primarily evaluate individual analytical methods, this framework conceptualizes measurement reliability as an interaction between methodological choices and systemic conditions, including strategy quality, feedback maturity, system structure, decision pressure, and organizational adaptation. The novelty of this research lies in reframing measurement error as a system-level outcome rather than merely a technical failure of statistical estimation. By integrating causal inference perspectives with systems thinking, this study provides a broader conceptual approach for understanding how organizations can develop adaptive measurement environments capable of detecting, interpreting, and correcting measurement inaccuracies. Empirically, this study aims to explore the relationship between System Shift variables, measurement error, and randomized controlled trial benchmarks using aggregate-coded marketing cases. Specifically, the research investigates three main questions: whether System Shift variables explain variations in measurement error compared with RCT benchmarks; which system components are most strongly associated with reliable measurement outcomes; and whether a systems-based diagnostic framework provides additional interpretive value beyond conventional marketing measurement variables, such as funnel stage and estimation method. The study employs descriptive analysis, correlation analysis, regression modeling, regularized logistic regression, model comparison, and cluster analysis to examine patterns between non-experimental estimates and experimental benchmarks. Although exploratory in nature, this approach provides initial empirical evidence regarding the potential role of organizational and systemic factors in shaping measurement reliability.

The contribution of this research is threefold. First, theoretically, the study extends marketing measurement literature by introducing a systems-oriented perspective that complements existing causal inference approaches. Rather than positioning experimental and non-experimental methods as competing alternatives, the study argues that measurement reliability depends on the interaction between analytical methods and the organizational systems in which they operate. Second, methodologically, the research provides an exploratory framework for diagnosing measurement regimes by identifying adaptive, transitional, and high-bias measurement conditions. Third, practically, the findings offer guidance for marketing managers by emphasizing that selecting advanced analytical tools alone is insufficient; organizations must also develop strong strategic planning processes and mature feedback mechanisms to ensure responsible interpretation of measurement outputs.

Ultimately, this research seeks to provide both academic and practical benefits by improving understanding of how marketing measurement systems can become more reliable, adaptive, and decision-oriented. For scholars, the study contributes a new conceptual lens that connects causal inference, organizational learning, and systems thinking in marketing analytics research. For practitioners, the findings highlight the importance of evaluating measurement environments before relying on analytical results for strategic decisions. By identifying the role of strategy quality and feedback maturity in reducing measurement error, this study encourages organizations to move beyond a purely technology-driven approach toward a more holistic measurement capability. Future research can build upon this exploratory validation by applying the framework to larger datasets, diverse industries, and broader marketing contexts to further examine its predictive validity and practical applicability.

METHOD

Data and case structure

The dataset consisted of six aggregate-coded marketing cases labeled MKT-001 to MKT-006. Each case included System Shift variables, a System Shift Risk Score, non-experimental lift estimates, RCT benchmarks, Measurement Error, Bias Level, Accuracy, and Progression status. The dataset size was limited; therefore, the analysis was interpreted as an exploratory validation or proof-of-concept rather than a confirmatory statistical test.

Outcome construction

Measurement Error is defined as the absolute difference between a non-experimental estimate and the RCT benchmark. Bias Level is defined as the non-experimental estimate minus the RCT benchmark. Because all six non-experimental estimates exceed their RCT benchmarks, all bias in this dataset reflects overestimation. Consequently, Bias Level and Measurement Error are identical in magnitude. Accuracy is coded as 1 when Measurement Error is at or below the median error threshold of 56.5 percentage points. Progression is coded as 1 when the Decision Risk Index is at or below the median threshold of 87.66.

Table 1. Operational definitions used in the exploratory validation.

Construct	Operational definition
Measurement Error	Absolute difference between non-experimental lift estimate and RCT benchmark.
Bias Level	Non-experimental lift estimate minus RCT benchmark; empirically identical to Measurement Error because all six cases overestimate the RCT benchmark.
Accuracy	Binary indicator coded 1 when Measurement Error is at or below the median threshold of 56.5 percentage points.
Progression	Binary indicator coded 1 when Decision Risk Index is at or below the median threshold of 87.66.
System Shift Risk Score	Composite index summarizing the coded risk condition of the measurement system.

Analytical strategy

The analysis proceeds in six steps. First, descriptive statistics profile measurement error and System Shift variables. Second, a correlation matrix evaluates the direction and strength of relationships among System Shift variables and measurement outcomes. Third, simple linear regressions estimate the association between Chokepoint Pressure and Measurement Error, and between System Shift Risk Score and Measurement Error. Fourth, regularized logistic regression evaluates the relationship between adaptive variables and binary reliability outcomes. Fifth, in-sample model comparison evaluates the explanatory power of Risk Score, Funnel Stage, Method Type, Funnel Stage plus Method Type, and Risk plus Status variables. Sixth, cluster analysis identifies measurement regimes based on risk score and measurement error.

RESULTS AND DISCUSSION

Descriptive profile

The dataset shows substantial measurement error. The mean Measurement Error is 79.0 percentage points, with a minimum of 19 and a maximum of 158. This is not a marginal

deviation from the RCT benchmark. It indicates that, in these cases, non-experimental estimates can materially overstate incremental marketing effects. The mean System Shift Risk Score is 10.83, with a minimum of 7 and a maximum of 14. Domain Lock and Actor Configuration are constant at the maximum value of 5 across all six cases, making them unusable for correlation-based explanation. Chokepoint Pressure has a mean of 4.17, System Condition has a mean of 4.00, while Position Quality averages 2.33, Strategy Quality 2.50, and Feedback Maturity 2.50.

Table 2. Descriptive profile of key variables.

Variable	Mean	Range / note	Interpretive relevance
Measurement Error	79.0	19 to 158	Substantive deviation from RCT benchmark
System Shift Risk Score	10.83	7 to 14	Most cases operate under elevated systemic risk
Chokepoint Pressure	4.17	Variable	Potential structural contributor to error
Strategy Quality	2.50	Variable	Adaptive capacity indicator
Feedback Maturity	2.50	Variable	Adaptive correction indicator
Domain Lock / Actor Configuration	5.00	Constant	Cannot explain cross-case variation

This descriptive pattern already suggests a central interpretation: the cases are structurally constrained, but their adaptive capacities differ. That difference becomes important in the correlation, model comparison, and cluster results.

Correlation analysis

The correlation matrix shows that Chokepoint Pressure is positively associated with Measurement Error ($r = 0.387$). The direction supports H2, although the association is not statistically reliable in such a small sample. System Shift Risk Score is also positively associated with Measurement Error, but weakly ($r = 0.240$), offering only limited support for H1.

The strongest results appear in the adaptive variables. Strategy Quality correlates negatively with Measurement Error ($r = -0.786$) and Bias Level ($r = -0.786$). Feedback Maturity shows the same negative relationship with Measurement Error and Bias Level ($r = -0.786$). Both Strategy Quality and Feedback Maturity correlate positively and perfectly with Accuracy and Progression in this dataset. This supports H3 and H4 at the exploratory level. More importantly, it suggests that measurement reliability may depend less on structural risk alone and more on whether the system has enough strategic and feedback capacity to absorb, correct, and interpret that risk.

Regression and classification models

The first linear regression estimates Measurement Error as a function of Chokepoint Pressure: Measurement Error = -44.53 + 29.65(CP). Each one-point increase in Chokepoint Pressure is associated with an increase of approximately 29.65 percentage points in Measurement Error. The model explains 15.0% of the variation in Measurement Error ($R^2 = 0.150$). The second model estimates Measurement Error as a function of System Shift Risk Score: Measurement Error = 18.57 + 5.58(System Shift Risk Score). The coefficient is positive, as expected, but the model explains only 5.8% of the variation in Measurement Error ($R^2 =$

0.058).

Regularized logistic regression is used because the dataset is small and some predictors nearly separate the outcomes. In the model predicting Accuracy, Strategy Quality and Feedback Maturity have positive coefficients. The in-sample AUC is 1.000, although classification accuracy at a 0.5 threshold is only 0.50 because predicted probabilities are compressed under regularization and small-sample conditions. The same pattern appears for Progression. When System Shift Risk Score is used to predict Accuracy and Progression, the coefficient is negative, indicating that higher risk is associated with lower probability of reliability. The AUC is 0.833 for both outcomes.

Model comparison

The model comparison provides one of the strongest cautionary findings. For Measurement Error, System Shift Risk Score alone explains only 5.8% of in-sample variation, while Funnel Stage explains 28.9% and Method Type explains 61.7%. The combined model of Funnel Stage and Method Type reaches $R^2 = 0.906$, as does the Risk plus Status model. This finding does not invalidate the System Shift Framework. Rather, it clarifies its proper role. The framework should not be positioned as a replacement for method-level causal evaluation. It should be positioned as a diagnostic layer that explains why certain measurement contexts become vulnerable to error and why certain methods may perform differently across system configurations.

Visual and Theoretical Interpretation

The findings support the System Shift Framework partially, not conclusively. The strongest evidence appears in the adaptive components of the framework, especially Strategy Quality and Feedback Maturity. These variables are consistently associated with lower Measurement Error, lower Bias Level, higher Accuracy, and higher Progression. This matters because it shifts the interpretation of marketing measurement error. The problem is not simply that one method is better than another. The deeper issue is whether the measurement system has the strategic and feedback capacity to discipline the use of any method.

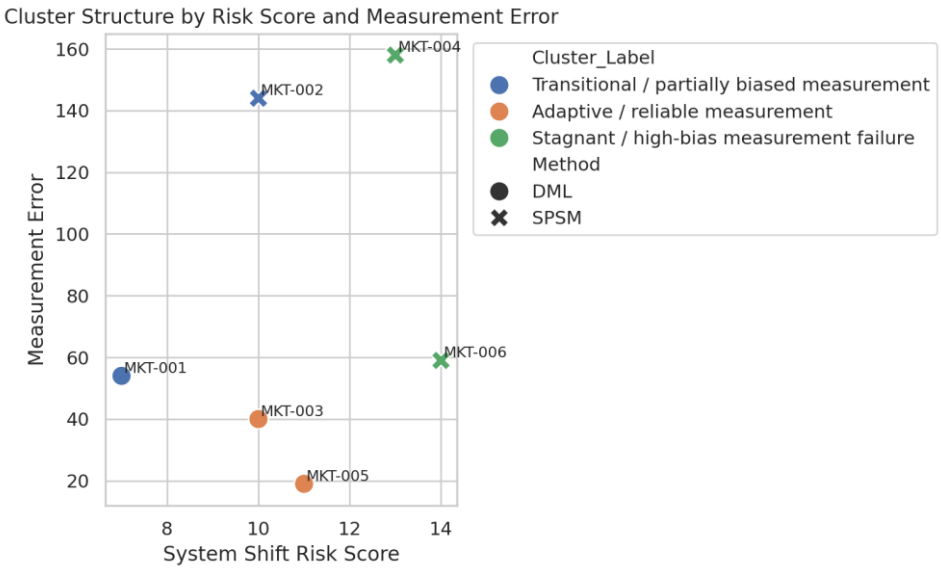

Figure 1. Cluster structure by System Shift Risk Score and Measurement Error.

Figure 1 makes the diagnostic contribution of the framework visible. The adaptive/reliable measurement cluster contains cases with lower measurement error, while the stagnant/high-bias failure cluster contains cases with high risk scores and larger error. The transitional cluster is especially important because it prevents a simplistic linear reading of risk. A case may share a funnel stage or broad risk condition with another case yet differ materially in measurement reliability. This suggests that measurement systems should be interpreted as regimes rather than as isolated model outputs.

The figure also shows that Method Type is not incidental. The high-bias failure cluster is populated by SPSM cases, whereas the adaptive cluster contains DML cases. With six observations, this cannot be generalized as a universal methodological ranking. Still, it supports a narrower and more useful claim: method type interacts with system condition. In weak or stagnant measurement systems, a method may amplify rather than resolve structural bias. In more adaptive systems, the same measurement process may become more interpretable because strategic design and feedback loops are stronger.

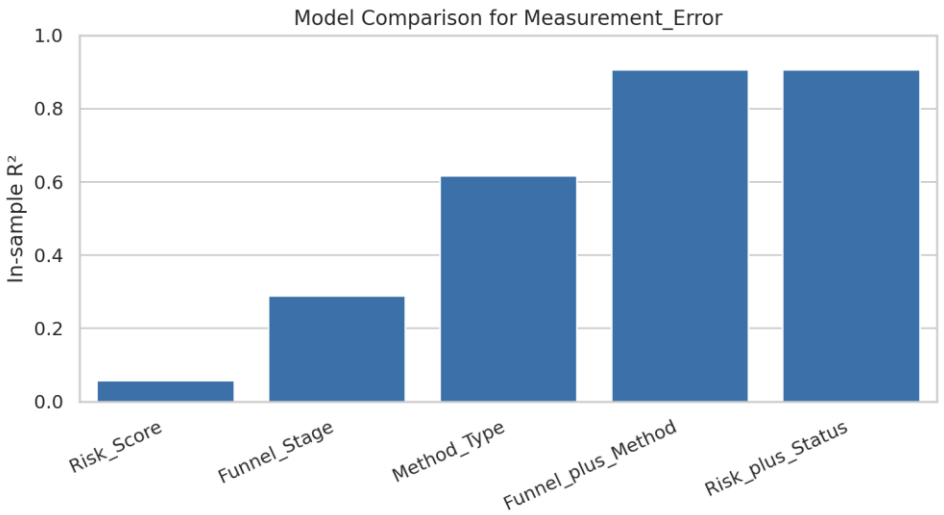

Figure 2. Model comparison for Measurement Error.

Figure 2 is a methodological warning. The aggregate System Shift Risk Score has limited explanatory power for continuous Measurement Error, whereas Method Type and the combined Funnel Stage plus Method Type model explain substantially more in-sample variation. This means that the System Shift Framework should not be presented as a substitute for causal identification or estimator-level validation. It should be presented as a diagnostic architecture that complements them. The figure also suggests that future validation must include Method Type as a core control variable. Without this control, the framework risks attributing to systemic risk what may partly arise from estimation design.

At the same time, Figure 2 does not make the framework redundant. It clarifies where its value lies. A method-level model may explain more variance, but it does not by itself explain why a measurement system enters a high-bias regime or why strategic and feedback capacities appear to separate reliable from unreliable cases. For managerial and theoretical purposes, the System Shift Framework adds a language for diagnosing the conditions under which methods become reliable or fragile.

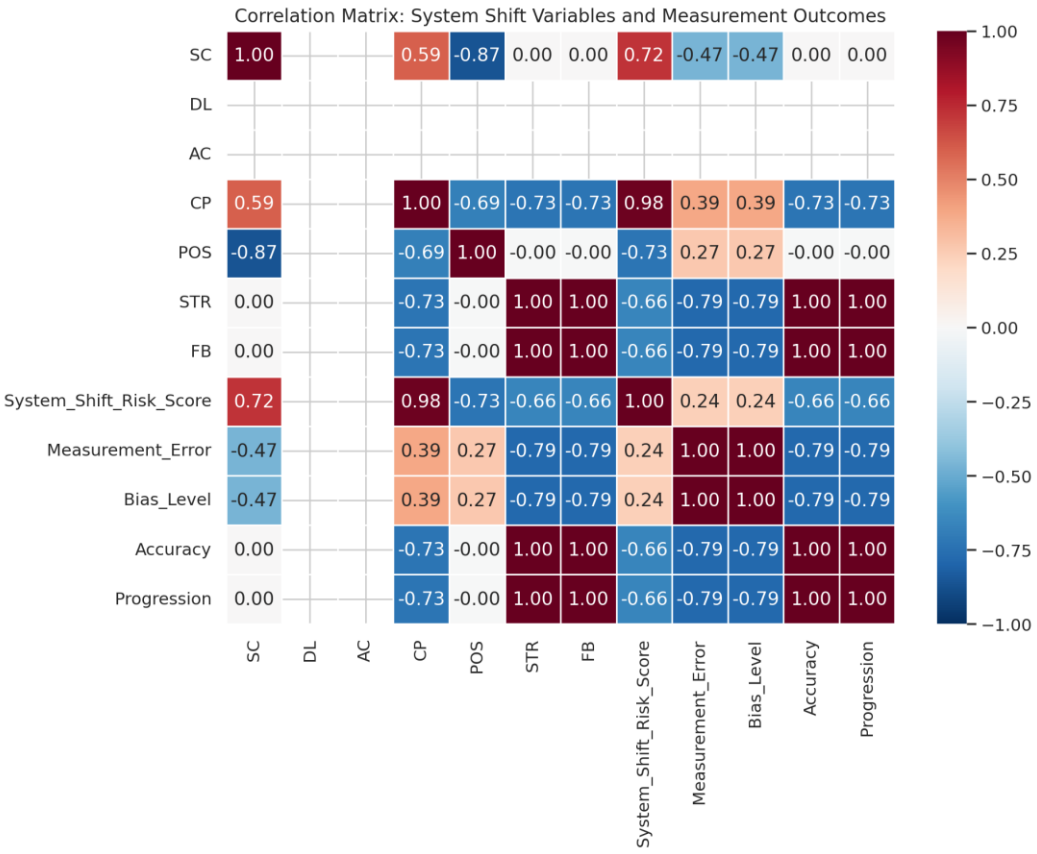

Figure 3. Correlation matrix of System Shift variables and measurement outcomes.

Figure 3 shows the internal logic of the empirical pattern. Strategy Quality and Feedback Maturity have strong negative associations with Measurement Error and Bias Level, while showing perfect positive associations with Accuracy and Progression in this small dataset. This does not prove causality. It does, however, identify the most theoretically promising part of the framework. Reliability appears to arise less from the absence of structural risk than from the presence of adaptive correction. In plain terms, the system does not become reliable simply because it is simple. It becomes more reliable when it can learn.

The correlation matrix also reveals why the aggregate risk score remains weak for continuous error. Domain Lock and Actor Configuration are constant, and Position Quality does not behave as expected. Aggregating these components into a single index may dilute the explanatory role of Strategy Quality and Feedback Maturity. A future version of the framework may therefore need to distinguish structural exposure from adaptive capacity more explicitly, rather than collapsing them into a single risk score.

Placed together, the three figures suggest a more nuanced interpretation. The System Shift Framework is not yet a predictive model of measurement error. It is, at this stage, a diagnostic framework for classifying measurement systems into adaptive, transitional, and stagnant regimes. The central theoretical insight is that measurement error emerges from a system-method interaction. Method Type matters. Funnel Stage matters. But Strategy Quality and Feedback Maturity may determine whether the organization can use, interpret, and correct measurement outputs responsibly.

Theoretical Contribution

This paper contributes to marketing measurement scholarship in three ways. First, it reframes measurement error as a system-level diagnostic outcome. Prior research has rightly focused on the accuracy of experimental and non-experimental methods. This paper adds that method performance may depend on systemic conditions, particularly strategic design and feedback maturity. Second, it introduces the System Shift Framework as an exploratory diagnostic structure for marketing measurement. The framework is not yet a validated predictive model. Its value at this stage is conceptual and interpretive: it helps classify measurement systems into adaptive, transitional, and stagnant regimes. Third, it bridges causal inference and systems thinking. Causal inference asks whether an estimate can be identified. Systems diagnosis asks whether the measurement environment is capable of producing, interpreting, and correcting that estimate. Both are needed.

Managerial Implications

For marketing leaders, the findings imply that measurement reliability should not be reduced to tool selection. A firm may adopt DML, propensity score matching, synthetic control, or another counterfactual method, yet still produce unreliable decisions if the measurement system is structurally constrained and adaptively weak. Measurement audits should therefore include feedback maturity. The question is not only whether the model was correct, but whether the organization has mechanisms to detect when the model is wrong. Strategy quality should also be treated as part of measurement validity. A poorly framed measurement question can produce technically sophisticated but decisionally misleading estimates.

High-risk measurement contexts should be classified before estimation. If a case belongs to a stagnant or high-bias regime, managers should avoid treating non-experimental estimates as decision-ready without stronger validation against experiments or quasi-experimental benchmarks. This is not a rejection of non-experimental measurement. It is a call for diagnostic humility. The system should be examined before the estimate is believed.

Limitations and Future Research

The limitations are substantial and must be stated clearly. First, the dataset contains only six aggregate-coded cases. The results cannot support strong statistical inference and should be read as proof-of-concept evidence. Second, several variables have limited or no variation. Domain Lock and Actor Configuration are constant at the maximum value across all cases, preventing meaningful correlation or regression analysis. Third, all six cases overestimate the RCT benchmark. This prevents the analysis from distinguishing overestimation from underestimation. Bias Level and Measurement Error therefore become empirically identical in this dataset. Fourth, the in-sample model comparison may overstate explanatory power because of the extremely small sample. R^2 values such as 0.906 should be treated as descriptive pattern indicators, not as externally valid predictive evidence.

Future research should validate the framework using larger datasets, campaign-level raw observations, broader variation in funnel stage and method type, and multiple categories of non-experimental estimators. A stronger design would test whether Strategy Quality and Feedback Maturity remain significant after controlling for Method Type, Funnel Stage, campaign size, baseline conversion rate, targeting intensity, and pre-treatment imbalance. A larger dataset would also allow the framework to distinguish structural exposure from adaptive capacity and to test whether the aggregate System Shift Risk Score requires reweighting.

CONCLUSION

This exploratory study concluded that measurement error in marketing analytics should not be interpreted solely as a technical limitation of statistical estimators but also as an indicator of broader system-level conditions that influence the reliability of advertising measurement outcomes. The findings demonstrated that Strategy Quality and Feedback Maturity were the factors most consistently associated with lower measurement error, higher accuracy, and improved measurement progression, whereas the aggregated System Shift Risk Score showed limited explanatory capability for continuous measurement error. The results also confirmed that Method Type remained an important determinant of measurement performance, indicating that the reliability of non-experimental marketing measurement depended on the interaction between analytical approaches and the organizational systems supporting their implementation. Therefore, the System Shift Framework provided value as a diagnostic architecture for identifying whether measurement environments operated within adaptive, transitional, or high-bias regimes. Rather than replacing causal inference methods, this framework complemented existing approaches by emphasizing that reliable marketing decisions required not only appropriate estimation techniques but also strategic clarity, organizational learning, and effective feedback mechanisms.

Future research should strengthen the empirical validation of the System Shift Framework by applying it to larger and more diverse datasets involving campaign-level observations, multiple industries, broader advertising contexts, and various non-experimental estimation methods. Subsequent studies should incorporate additional control variables, such as campaign scale, targeting intensity, baseline conversion rates, customer characteristics, data quality, and organizational analytics maturity, to examine whether Strategy Quality and Feedback Maturity remain significant predictors after accounting for methodological differences. Furthermore, future investigations may refine the measurement structure of the framework by distinguishing structural risks from adaptive capabilities rather than combining them into a single risk index. Longitudinal research designs could also provide deeper insights into how organizations improve measurement reliability over time through learning processes, technological adaptation, and feedback-driven decision systems. Such developments would contribute to establishing a stronger theoretical and practical foundation for advancing responsible, accurate, and adaptive marketing measurement practices.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490), 493-505. https://doi.org/10.1198/jasa.2009.ap08746

Ambrosi, E., Mezzalira, E., Canzan, F., Leardini, C., Vita, G., Marini, G., & Longhini, J. (2025). Effectiveness of digital health interventions for chronic conditions management in European primary care settings: Systematic review and meta-analysis. *International Journal of Medical Informatics*, *196*, 105820.

Braga, L. H., Farrokhyar, F., Dönmez, M. İ., Nelson, C. P., Haid, B., Herbst, K., Garriboli, M., Cascio, S., Nieuwhof-Leppink, A., & Kaefer, M. (2025). Randomized controlled trials–The what, when, how and why. *Journal of Pediatric Urology*, *21*(2), 397–404.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1-C68. https://doi.org/10.1111/ectj.12097

Doucette, D. (2026). *Controllable Coherence in Marketing Systems: A Poisson–Algebraic Theory of Dissipation, Resonance, and Chaos*. unpublished manuscript.

Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. Marketing Science, 42(4), 768-793. https://doi.org/10.1287/mksc.2022.1413

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193-225. https://doi.org/10.1287/mksc.2018.1135

Han, R., Acosta, J. N., Shakeri, Z., Ioannidis, J. P. A., Topol, E. J., & Rajpurkar, P. (2024). Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health*, *6*(5), e367–e373.

Han, T., & Li, Y.-F. (2022). Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles. *Reliability Engineering & System Safety*, *226*, 108648.

Iashchenko, O., Chupryna, O., Maksymov, S., Zemlianska, N., Afanasieva, O., & Popova, Y. (2025). *Information Noise in Marketing Analytics: Implications for Financial Reporting and Economic Decision-Making*.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics, 130(4), 1941-1973. https://doi.org/10.1093/qje/qjv023

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *ArXiv Preprint ArXiv:2308.05374*.

Moschonis, G., Siopis, G., Jung, J., Eweka, E., Willems, R., Kwasnicka, D., Asare, B. Y.-A., Kodithuwakku, V., Verhaeghe, N., & Vedanthan, R. (2023). Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with type 2 diabetes: a systematic review and meta-analysis of randomised controlled trials. *The Lancet Digital Health*, *5*(3), e125–e143.

Ngo, V. M., & Vu, T. M. (2026). Generative AI in marketing: mapping evolving ethical concerns through social network analysis. *International Journal of Ethics and Systems*, 1–36.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

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

Siopis, G., Moschonis, G., Eweka, E., Jung, J., Kwasnicka, D., Asare, B. Y.-A., Kodithuwakku, V., Willems, R., Verhaeghe, N., & Annemans, L. (2023). Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with hypertension: a systematic review and meta-analysis of randomised controlled trials. *The Lancet Digital Health*, *5*(3), e144–e159.

Srinivas, S. (2025). *The Personalization Paradox: A Bias-Enabled Guidance Trap*.

Tankha, H., Gaskins, D., Shallcross, A., Rothberg, M., Hu, B., Guo, N., Roseen, E. J., Dombrowski, S., Bar, J., & Warren, R. (2024). Effectiveness of virtual yoga for chronic low back pain: a randomized clinical trial. *JAMA Network Open*, *7*(11), e2442339.

Theodorakopoulos, L., Theodoropoulou, A., & Halkiopoulos, C. (2025). Cognitive bias mitigation in executive decision-making: A data-driven approach integrating big data analytics, AI, and explainable systems. *Electronics*, *14*(19), 3930.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., & Jiang, Z. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, *37*, 95266–95290.

Woodside, A., & Kumar, V. (2023). Toward a General Theory of Anomalies in Strategic Management, Marketing, Psychology, and Consumer Research: Antecedents, Processes, and Outcomes. *Marketing, Psychology, and Consumer Research: Antecedents, Processes, and Outcomes*.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*(8032), 61–68.