European Journal of AI, Computing & Informatics

Article

Research on Measuring User Behavior Response Differences Supported by Propensity Scoring Method

Zhimeng Liu [1,*]

[1] Harvard T.H. Chan School of Public Health, Harvard University, Boston, United States
* Correspondence: Zhimeng Liu, Harvard T.H. Chan School of Public Health, Harvard University, Boston, United States

Abstract: Measuring the differences in user behavior responses is increasingly crucial for identifying targeted intervention effects and improving overall digital platform operation in the era of big data. However, observational studies often face significant methodological challenges. To overcome the profound influence of user characteristic differences, inherent sample selection bias, and unobserved confounding factors, the propensity score method is systematically utilized to construct a robust measurement system. This study meticulously classifies users based on the specific type of behavior response and deeply explores the empirical possibility of applying this advanced statistical approach in complex groups that were not randomly sampled. A comprehensive and complete model is constructed around critical methodological issues, including the precise selection of dependent variables, the rigorous elimination of covariates, the accurate estimation of propensity scores, and the implementation of advanced sample matching and weighting techniques. Furthermore, analytical methods such as inter-group comparison, heterogeneity testing, and extensive robustness testing are adopted to significantly enhance the accuracy, reliability, and persuasiveness of the measurement results. Ultimately, this research provides vital technical support and actionable insights for platform administrators and marketers in identifying nuanced behavioral characteristics, rigorously evaluating marketing intervention effects, and strategically optimizing enterprise operations for sustainable growth and improved user engagement.

Keywords: propensity score; user behavior; sample matching; intervention evaluation; data analytics

1. Introduction

The cumulative data derived from user clicks, page views, purchase behaviors, stay durations, and interaction patterns on digital platforms serve as a robust foundation for conducting response variance analysis. Each user exhibits unique characteristics, preferences, activity levels, and payment capacities, which can significantly influence their behavior. When comparisons are limited to pre- and post-intervention scenarios or between distinct groups, there is a risk of introducing sampling bias and confounding variables. To address these challenges, the propensity score method can be employed to predict the likelihood of users accepting an intervention or being assigned to a specific group [1]. This approach transforms multiple related variables into a scoring system that facilitates matching and weighting, thereby enhancing sample homogeneity. By analyzing changes in user responsiveness through this method, researchers can more precisely assess the effects of interventions and identify group-specific traits. Such insights are invaluable for optimizing platform operations, refining product offerings, and implementing effective customer segmentation strategies, ultimately driving improved outcomes in digital environments.

Received: 07 March 2026
Revised: 27 April 2026
Accepted: 09 May 2026
Published: 12 May 2026

2. Measurement of User Behavior Response Differences Based on Propensity Score Method

2.1. The Connotation of Behavioral Response and Classification of Types

User behavior response refers to the variations in user behavior that arise due to factors such as platform push notifications, promotional campaigns, page redesigns, product price changes, and service interactions. This concept encompasses not only whether a user engages in a specific action but also the intensity, frequency, duration, and overall effectiveness of that action. For instance, user responses can manifest as clicking on a link, reading content, saving items, making purchases, engaging in repeat purchases, storing information, or interacting with the platform. From an analytical perspective, these responses can be categorized into immediate responses, delayed responses, persistent responses, and attrition responses. Each type of response provides valuable insights into the degree of user acceptance of intervention factors and the pathways of behavioral transformation. Furthermore, these classifications serve as critical benchmarks for assessing behavioral differences. By clearly defining the concept and types of user behavior responses, researchers can more accurately identify the specific effects of interventions [2]. This clarity helps to avoid misinterpreting general behavioral changes as direct outcomes of targeted interventions. Such distinctions are essential for ensuring the precision and reliability of behavioral analysis, as they allow for a more nuanced understanding of user engagement and the factors influencing it.

2.2. The Measurement Application Logic of Propensity Score Method

The propensity score method serves as a robust analytical tool for measuring differences in user behavior responses within non-randomized experimental settings. In online environments, users are typically not exposed to recommendations, activity notifications, or feature displays in a random manner. Instead, user characteristics such as age, geographic location, purchasing power, historical activity levels, and personal preferences often vary significantly. Directly comparing outcomes without accounting for these differences can lead to selection bias, which undermines the validity of the analysis. The propensity score method addresses this issue by estimating the probability of an individual receiving a specific intervention based on observable covariates. This score is then used to implement matching, stratification, or weighting techniques, ensuring that the treatment and control groups share more comparable characteristics. By minimizing the impact of confounding variables, this method enhances the reliability of causal inferences [3]. Employing the propensity score method shifts the analytical focus from surface-level comparisons to a deeper exploration of causal relationships, thereby providing a more nuanced understanding of user behavior dynamics in complex systems.

3. Construction of a Measurement Model for User Behavior Response Differences Based on Propensity Score Method

3.1. Identification of Response Variables and Setting of Indicators

An essential component of constructing a model to measure differences in user behavioral responses is the accurate identification of user response variables. This process depends on the nature of the study intervention and the specific behavioral transformation objectives. For instance, if the research focuses on the effects of marketing exposure, core response signals such as click rate, browsing depth, and stay time should be prioritized as dominant indicators. Conversely, for studies centered on sales incentives, metrics like order conversion rate, average customer spending, frequency of repurchases, and the number of coupons redeemed can be utilized. When the emphasis is on long-term customer engagement, additional indicators such as membership retention rates, event participation levels, and user lifetime value may be introduced. It is critical to differentiate between binary variables, continuous variables, and count variables, as the choice of estimation method must align with the data distribution characteristics. To minimize the impact of random disturbances, it is necessary to clearly define the observation period,

feedback criteria, and statistical scope. These parameters ensure that the feedback results accurately reflect the magnitude of behavioral changes among users who have been influenced by the intervention. Furthermore, the selection and calibration of these indicators must be methodical to ensure the reliability and validity of the measurement model, thereby enabling a more precise analysis of user behavior dynamics [4, 5].

3.2. Covariate Selection and Confounding Control

Covariate selection plays a pivotal role in ensuring the precision of propensity assessment and the effectiveness of measuring response differences. User responses are influenced by a multitude of factors, including demographic characteristics, device types, geographical location, historical consumption patterns, frequency of visits, content preferences, account levels, and prior conversion records. These variables not only shape the probability of users accepting an intervention but also significantly impact the resulting response outcomes. To minimize biases, the selection process should integrate theoretical relevance with data availability, avoiding an over-reliance on significance tests that may lead to default biases. For variables exhibiting strong behavioral tendencies, dimensionality reduction techniques such as correlation analysis, analysis of variance, Lasso regression, and variable importance ranking derived from tree models can be employed [6, 7]. Effective confounding control is essential to preserve the relationship between intervention allocation and behavioral outcomes while reducing the influence of selection bias on the estimation of response heterogeneity. By addressing these aspects, researchers can enhance the robustness of their findings and ensure that the observed effects are attributable to the intervention rather than extraneous factors.

3.3. Propensity Score Estimation and Sample Matching

The estimation of propensity scores represents a pivotal step in simplifying the multidimensional characteristics of customers into a single probability value. This probability reflects the likelihood of users being selected into the treatment group and can be calculated using various methods, including Logistic regression, Probit models, Random Forest, and Gradient Boosting Decision Trees (GBDT). The choice of a specific model should be guided by considerations such as interpretability, predictive accuracy, and the ability to achieve covariate balance. Once the propensity scores are estimated, comparable samples can be constructed through techniques such as nearest neighbor matching, radius matching, kernel matching, or stratified matching. During the matching process, it is essential to define an appropriate distance range and exclude data instances that fall outside the support domain to prevent distortion of the estimated values caused by excessively high propensity scores. To ensure the effectiveness of the matching procedure, post-matching tests should be conducted to evaluate the consistency of standard deviation, covariate histograms, and variance ratios. These tests help verify that the treatment and control groups are balanced in their fundamental characteristics, thereby establishing a robust foundation for analyzing differences in outcomes [7, 8]. Such methodological rigor is critical for deriving reliable and meaningful insights from the data.

3.4. Construction of Weight Structure and Sample Correction

The construction of the weight function primarily addresses the issue of sample imbalance and enhances the representativeness of response difference estimation. By employing propensity scores, techniques such as inverse probability weighting, balanced weights, and overlap coefficients can be utilized. These methods ensure that the treated group and the control group achieve a quasi-random allocation effect after weighting, thereby improving the reliability of the analysis. For samples with a propensity to exhibit ratings close to 0 or 1, appropriate weight truncation or pruning should be implemented. This step is crucial to prevent excessive weights from inflating the estimation variance, which could compromise the accuracy of the results. Sample calibration involves several critical processes, including managing missing values, identifying and addressing outliers, and handling repeated users. These steps are essential to maintain the stability and reliability of behavioral data. After applying the weighting adjustments, it is necessary to

conduct balance tests again. If significant imbalances persist, a double robust estimation method can be employed. This approach combines weighting adjustments with the final result model, thereby enhancing the robustness and technical credibility of the measurement of user behavior response differences. Such comprehensive measures ensure the methodological rigor and reliability of the analysis.

4. Measurement and Validation of User Behavior Response Differences Supported by Propensity Score Method

4.1. Measurement of Inter-Group Differences

The fundamental concept behind measuring inter-group differences is to evaluate the behavioral response intensities of the experimental group and the control group after conducting propensity score matching (PSM) and propensity weighting (PW) tests. This approach is designed to mitigate biases that may arise due to disparities in sampling structures when relying solely on simple mean differences. By employing these methods, researchers can ensure a more accurate comparison of user behavior across groups, thereby enhancing the reliability of the findings.

Let D_i denote whether a user accepts the intervention of a specific platform. If D_i equals 1, it signifies that the user has accepted the precise push; conversely, if D_i equals 0, it indicates that the user has not accepted the precise push. The variable Y_i represents the user's actual response, encompassing actions such as clicks, purchases, time spent on the platform, and repeat purchase rates. The propensity score, denoted as $e(Xi)=P(Di=1 \mid Xi)$, is utilized to match users with similar characteristics. Within the matched sample, the average treatment effect of the treatment group can be calculated to quantify the impact of the intervention. This method provides a robust framework for analyzing the incremental behavioral responses induced by targeted interventions [8, 9].

$$ATT=\frac{1}{N_1}\sum_{i:D_i=1}\left(Y_i-\sum_{j:D_j=0} w_{ij} Y_j\right)$$

In this context, N_1 represents the sample size of the treatment group, while w_{ij} denotes the matching weight constructed between user i in the treatment group and user j in the control group, based on the distance of their propensity scores. This formula enables the measurement of the average incremental response attributable to intervention behaviors under comparable user characteristics. When employing the inverse probability weighting method, researchers can further construct a weighted average difference to assess the overall impact. This approach is particularly effective in scenarios where sample sharing is well-supported, and the research objective is to evaluate the aggregate average response difference across groups.

$$ATE=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)}-\frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

This methodology is highly applicable in cases where the objective is to analyze the effectiveness of marketing strategies or interventions. For instance, in the context of coupon distribution by an e-commerce platform, propensity score estimation can be performed using variables such as click counts over the past month, historical consumption amounts, membership levels, product type preferences, and device types. Users can then be categorized into treatment groups, comprising those who received coupons, and control groups, consisting of users with similar characteristics who did not receive coupons [10, 11]. If the conversion rate for the experimental group is observed to be 18.6%, while the weighted control group exhibits a conversion rate of 13.2%, the response gap of 5.4 percentage points demonstrates the positive impact of coupon marketing. This finding underscores the importance of controlling for differences in customer characteristics to accurately assess the effectiveness of targeted interventions.

4.2. Identification of Heterogeneity Effects

The heterogeneity effect analysis seeks to reveal the varying degrees of user responses based on differing attributes to recommendation results, aiming to mitigate confounding effects that may obscure the true impact on specific user groups. Let the user intervention status be D_i, the behavioral response result be Y_i, the covariate set be X_i,

and the propensity score be $e(X_i)=P(D_i=1|X_i)$. Once matching or weighted correction is completed, the overall treatment effect can be extended to the conditional average treatment effect. This approach allows for a more nuanced understanding of how interventions influence distinct user segments, ensuring that the analysis accounts for variability in user characteristics and behaviors.

$$CATE(x)=E[Y_i(1)-Y_i(0) \mid X_i=x]$$

In this framework, Y_i (1) represents the potential response of users after receiving the intervention, while Y_i (0) denotes the potential response when users do not receive the intervention. By stratifying users based on activity levels, consumption tiers, or preference intensities, a grouped treatment effect can be derived. This enables researchers to identify how different user groups respond to the same intervention. For example, high-frequency users may exhibit stronger behavioral changes compared to low-frequency users, highlighting the importance of tailoring strategies to specific user segments to maximize effectiveness.

$$\tau g=\frac{1}{N_g}\sum_{i\in g}\left(\frac{D_iY_i}{e(X_i)}-\frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

Within this context, g represents a specific user subgroup, and N_g indicates the sample size of this subgroup. This formula facilitates comparisons of response strengths across various user groups subjected to identical interventions, enabling the identification of high-sensitivity and low-sensitivity groups [12]. To improve the precision of identification, an interaction term model can be constructed following propensity score weighting. This model accounts for the influence of user stratification variables and provides insights into how intervention effects vary across different user attributes, thereby enhancing the robustness of the analysis.

$$Y_i=\alpha+\beta D_i+\gamma' Z_i+\delta(D_i\times Z_i)+\theta X_i+\varepsilon_i$$

Here, Z_i represents the user stratification variable, while δ depicts the range of variation in intervention effects across different user attributes. By incorporating these variables, researchers can better understand the dynamics of intervention impacts and refine strategies to target specific user groups more effectively. This approach ensures that the analysis captures the complexity of user behaviors and the interplay between intervention effects and user characteristics, thereby providing actionable insights for optimizing recommendation systems.

For instance, consider personalized recommendations on short-video platforms. By analyzing the number of active sessions within a seven-day period, users can be categorized into low-frequency, medium-frequency, and high-frequency groups. Combining this classification with data such as age, device type, historical viewing duration, preferred content types, and follower count allows for a detailed analysis. If weighted statistics reveal that the completion rate increases by 8.7% for high-active users, 4.1% for medium-active users, and only 1.3% for low-active users, it becomes evident that personalized recommendation strategies are more effective for loyal users. Furthermore, if the interaction variable $D_i\times Z_i$ is significantly positively correlated among high-active users, it suggests that this factor contributes to the growth of high-active users. This demonstrates that activity level not only influences baseline behavior rates but also modulates the intensity of intervention effects, underscoring the importance of tailoring strategies to user activity levels for optimal outcomes.

4.3. Balance Test

The core of the balance test is to evaluate whether the differences in basic user characteristics between the treatment group and the control group have been effectively minimized after applying propensity score matching or weighting. It is insufficient to rely solely on metrics such as click-through rate, purchase rate, or duration of stay after exposure. Instead, it is essential to ensure that key variables, such as historical activity levels, purchasing power, interest preferences, terminal type, and usage frequency, exhibit homogeneity between the two groups. In practical applications, various methods can be employed to assess this balance, including mean difference analysis, coefficient of difference ratio, overlap graphs, and density curve graphs. Among these, the mean

difference index is particularly suitable for analyzing large-scale platform data, as it provides a more robust measure compared to traditional statistical significance tests. If the mean difference of the core correlation variables is controlled within 0.1, it indicates that the sample has achieved a balanced allocation after matching or weighting. This ensures that subsequent comparisons are based on more equitable and reliable group characteristics.

Consider the example of member discount coupons distributed by an e-commerce platform. After issuing these coupons, it was observed that users who received them exhibited significantly higher recent purchase intervals, historical average consumption amounts, and browsing frequencies compared to those who did not receive the coupons. Simply comparing transaction conversion rates in such cases can lead to an overestimation of the impact of the information dissemination [13, 14]. By applying the propensity score matching method to these two groups, it becomes evident that they achieve a high degree of consistency in variables such as purchase frequency, total consumption amount, preferred categories, and participation in promotional activities. Furthermore, the standard deviation of core influencing factors decreases significantly, from 0.30 to below 0.08. At this stage, comparing metrics such as purchase conversion rates or repeat purchase rates provides a more accurate measurement of the response effect generated by the member coupon information during its dissemination. This refined approach ensures that the evaluation of marketing strategies is based on a scientifically sound and balanced framework, reducing biases and enhancing the reliability of the results.

4.4. Robustness Verification

The robustness test is conducted to evaluate whether the modeling method, data selection, matching method, and tail observations influence the measurement results of user behavior response differences. While the propensity score method can control observable confounding variables, variations in propensity score methods, matching rules, and data cleaning approaches may still introduce estimation bias [12]. To ensure robustness, tests are performed from multiple perspectives, including altering the propensity score model settings, adjusting the matching radius, employing different matching methods, excluding samples with extreme propensity score values, modifying the observation period of action responses, and conducting placebo tests. If the results remain consistent in direction and show no significant changes under these varying conditions, it can be concluded that the measurement results are relatively robust. This comprehensive approach ensures that the findings are reliable and not influenced by methodological or data-related biases.

Using the course promotion of an online education platform as an example, the baseline model demonstrates that course recommendations increase the conversion rate of trial sessions by 6.2%. When kernel matching is applied, the effect rises to 5.9%. With the use of inverse probability weighting, the effect further increases to 6.5%. Even after excluding data with excessively high or low values, the effect remains stable at approximately 6.0%. Extending the observation period from 7 days to 14 days reveals that the recommendation intervention continues to be effective, indicating that the observed effect is not merely the result of short-term random fluctuations. Additionally, a placebo test is conducted by using historical purchase records prior to the recommendation as a pseudo-explanatory variable [1]. If no significant difference is observed between the treatment group and the control group in this scenario, the credibility of the conclusion is further strengthened. These findings collectively highlight the robustness of the results, demonstrating that the observed effects are consistent and reliable across different methodological adjustments and observational conditions.

5. Conclusion

By employing the propensity score method to quantify differences in users' responses, researchers can effectively mitigate selection bias and reduce interference from confounding factors in non-random sampling scenarios. This approach enhances the

statistical power of group comparisons and facilitates the exploration of diverse aspects, including response types, variable construction, covariate selection, propensity score estimation, sample matching, weighting adjustments, and result validation. Such a comprehensive methodology enables a more precise identification of counterfactual differences across various user actions, such as clicks, purchases, retention, and interactions. Furthermore, the application of balance tests and stability tests ensures the robustness and reliability of the conclusions, thereby increasing the practical utility of the findings in areas like platform operations, user segmentation, and strategic optimization. Looking ahead, the integration of advanced techniques such as machine learning algorithms, causal inference frameworks, time series analysis, and multi-dimensional identification methods holds significant potential to propel this research towards achieving greater refinement, real-time adaptability, and enhanced interpretability. These advancements could pave the way for more dynamic and actionable insights in both academic and practical domains.

References

1. J. A. Rios and J. Deng, "Is effort moderated scoring robust to multidimensional rapid guessing?," *Educational and Psychological Measurement*, vol. 85, no. 1, pp. 134-155, 2025.
2. L. L. Li, J. Fu, C. Xu, M. Ni, W. Chai, L. B. Hao, ... and J. Y. Chen, "Gender Differences in Ankylosing Spondylitis Patients with Advanced Hip Involvement: Results from A Matched Retrospective Cohort Study," *Orthopaedic Surgery*, vol. 14, no. 2, pp. 405-410, 2022.
3. Y. Wang, Q. Tu, and Z. Tao, "Optimizing customer engagement in fintech marketing: A telecom-centric approach to precision targeting using mobile app data," in *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 826-832, Oct. 2024.
4. B. Doleschal, D. Niedersüß-Beke, P. Kirchweger, A. Petzer, J. Thaler, and H. Rumpold, "Survival outcome in early-onset metastatic colorectal cancer: a multicenter-matched pair analysis," *Oncology*, vol. 102, no. 2, pp. 107-113, 2024.
5. X. Li, S. Zhang, and X. Song, "The impact of Internet use and involvement on residents' attitudes to healthcare in China: A propensity score matching analysis," *Plos one*, vol. 19, no. 8, p. e0305664, 2024.
6. X. Cao and Y. Fan, "Exploring the influences of density on travel behavior using propensity score matching," *Environment and Planning B: Planning and Design*, vol. 39, no. 3, pp. 459-470, 2012.
7. F. J. Thoemmes and E. S. Kim, "A systematic review of propensity score methods in the social sciences," *Multivariate Behavioral Research*, vol. 46, no. 1, pp. 90-118, 2011.
8. S. Valenzuela, A. Arriagada, and A. Scherman, "Facebook, Twitter, and youth engagement: A quasi-experimental study of social media use and protest behavior using propensity score matching," *International Journal of Communication*, vol. 8, p. 25, 2014.
9. X. Cao, "Exploring causal effects of neighborhood type on walking behavior using stratification on the propensity score," *Environment and Planning A*, vol. 42, no. 2, pp. 487-504, 2010.
10. M. Li, "Using the propensity score method to estimate causal effects: A review and practical guide," *Organizational Research Methods*, vol. 16, no. 2, pp. 188-226, 2013.
11. R. M. Pruzek, "Introduction to the special issue on propensity score methods in behavioral research," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 389-398, 2011.
12. B. Keller and E. Tipton, "Propensity score analysis in R: A software review," *Journal of Educational and Behavioral Statistics*, vol. 41, no. 3, pp. 326-348, 2016.
13. Z. Qin, S. J. Chen, D. Metzler, Y. Noh, J. Qin, and X. Wang, "Attribute-based propensity for unbiased learning in recommender systems: Algorithm and case studies," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2359-2367, Aug. 2020.
14. Y. Wang, H. Zhang, T. Feng, and H. Wang, "Does internet use affect levels of depression among older adults in China? A propensity score matching approach," *BMC Public Health*, vol. 19, no. 1, p. 1474, 2019.