Theory Instead of Experiment (TIE): A Creator Valuation System at Tencent

Citation: Lei Huang and Juanjuan Zhang. 2025. Theory Instead of Experiment (TIE): A Creator Valuation System at Tencent. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25). Association for Computing Machinery, New York, NY, USA, 4522–4532.

Published Version: https://doi.org/10.1145/3711896.3737267

Publisher: ACM|Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2

Permanent Link: https://hdl.handle.net/1721.1/165194

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Theory Instead of Experiment (TIE): A Creator Valuation System at Tencent

Lei Huang*
leihuang@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Juanjuan Zhang
jjzhang@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

Experiments are informative but should be used judiciously as a costly resource. Well-constructed theory may serve as a substitute. We develop a "Theory Instead of Experiment" (TIE) framework and, in collaboration with Tencent, apply the framework to assess how much value (e.g., user clicks) each creator contributes to its WeChat Official Accounts Platform. This TIE application models content demand and supply upon the counterfactual departure of a creator. The demand model predicts user clicks based on estimated user preferences, while the supply model captures the platform's content distribution response. Together, they predict how each creator influences user engagement through the platform's content distribution strategy. We test the predictions of the TIE system with 168 experiments, each examining a different mix of creators and involving more than 9 million unique users. The TIE system and the experiments demonstrate a 97% correlation on the key performance metric (change in user clicks). Based on its low costs, high accuracy, granular output, and minimal latency, Tencent has deployed the TIE system as the default approach to creator valuation, assessing tens of millions of creators each day while avoiding a 2.5% user click loss associated with a typical experiment.

CCS Concepts

• **General and reference → Experimentation**; • **Computing methodologies → Machine learning**; **Modeling and simulation**; • **Applied computing → E-commerce infrastructure**.

Keywords

Economic Theory, Experimentation, Counterfactual Prediction, Creator Valuation, Content Platform, Automation

ACM Reference Format:
Lei Huang and Juanjuan Zhang. 2025. Theory Instead of Experiment (TIE): A Creator Valuation System at Tencent. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3711896.3737267

*Authorship is alphabetical. Both authors contributed equally to this research.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737267

1 Introduction

One of the most consequential deployments in human history, the Little Boy atomic bomb used during World War II, did not undergo a full-scale test [38]. The cost of such a test would have been prohibitive, considering the scarcity of the key material, Uranium-235, and the need to maintain strategic secrecy. However, confidence in the bomb's performance was high, based on precise mathematical calculations and a series of smaller tests involving components of the underlying model. Ultimately, the costs of a full-scale test were deemed to outweigh any potential benefit.

The lesson from Little Boy remains valuable today. Informative as they are, experiments should be viewed as a strategic resource, used only after carefully balancing their benefits and costs. This is particularly true for experiments conducted in natural environments, testing the impact of a product or invention on its actual users or target audiences. Although such experiments provide insights into real-world responses, they come with implementation costs and opportunity costs—the losses incurred from deviating from optimal strategies for the sake of experimentation. Rather than investing blindly in experiments, it is more prudent to first consider lower-cost alternatives that offer comparable informational value, such as well-established domain theories.

We develop one such solution, a "Theory Instead of Experiment" (TIE) framework, to help companies conserve the use of costly experiments. The TIE framework is built upon the following pillars: (1) a high-fidelity theory model of the phenomenon in question, (2) if applicable, tests of the assumptions underlying the theory model, and (3) tests of the theory model's predictions with a smaller set of experiments; if the predictions of the theory model are sufficiently accurate, they can serve as a substitute for further experiments.

In collaboration with Tencent, we apply the TIE framework to one of its high-priority business problems: creator valuation. Tencent aims to measure the value, in terms of user engagement (with clicks being a key performance metric), that each of its content creators contributes to its WeChat Official Accounts Platform. Prior to TIE, experiments were the default way to answer this question at Tencent. A typical experiment would temporarily block user access to a set of creators, track changes in user engagement, and infer the creators' value based on such changes. These experiments are resource-intensive to implement and tend to result in losses in user engagement. For example, to identify valuable creators, an experiment would need to block access to these creators, which in turn would cause a noticeable drop in user engagement.

The TIE system for creator valuation follows a different approach, as outlined below. We theorize the problem as predicting how user engagement would change in the new equilibrium if a creator were to deviate from contributing content on the platform. The answer

depends on how users, other creators, and the platform may respond in this counterfactual scenario. We assume (with supporting evidence) that a click-prediction model based on user preferences calibrated on factual user engagement data predicts user engagement in the counterfactual scenario. We also assume (with supporting evidence) that other creators do not change their posting behaviors in response to a single creator's departure, given the enormous number of creators on the platform. The question then boils down to how the platform would respond to the creator's departure. We invoke the platform's content distribution mechanism to specify the list of content the platform would provide to each user in the counterfactual scenario. These counterfactual content lists, combined with the click-prediction model, form theoretical predictions of user engagement in the absence of the focal creator. The difference in user engagement with versus without a creator defines the creator's value.

The focus on microfounded theory enhances the model's generalizability and predictive robustness. This is a well-established principle in structural modeling, which emphasizes the measurement of behavioral primitives that are invariant to policy changes [33, 35]. In the context of creator valuation, the policy change is the (hypothetical) departure of a creator. The assumption that users' content preference primitives remain policy-invariant is standard in structural modeling. The main difference is that we have a valuable opportunity to observe—rather than assume—the platform's counterfactual gameplay under the new policy [17]. This is an important feature, as it helps prevent any misspecification of the counterfactual gameplay from compromising the predictive accuracy of the theory model.

We test the predictions of the TIE system against 168 experiments, each involving more than 9 million unique users and blocking user access to different profiles of creators. The TIE predictions achieve a 97% correlation with the experimental results, while incurring negligible implementation cost and no opportunity cost in terms of lost user clicks, compared with the 2.5% user click loss associated with a typical experiment in this setting. In light of its low costs, high accuracy, individual-level granularity, and minimal latency, the TIE framework has been deployed at Tencent on a full scale as the default system for creator valuation, assessing the value of tens of millions of creators each day.

2 Related Research

Experiments provide direct information about counterfactual environments that are otherwise unobservable or observed in a way that could be confounded [1, 22]. For this reason, experimentation has been a fundamental research methodology in various fields, including physics [37], chemistry [29], biology [12], economics [30], and social science [39], and has gained widespread popularity in the technology industry [18, 27, 32, 50].

However, the literature has acknowledged that experiments can be costly. Experiment infrastructures can be resource-intensive to build [25, 48]. Moreover, experiments almost inevitably incur opportunity costs, because not all experiment groups can be exposed to the optimal policy [6, 47], a fact that introduces both efficiency and ethical concerns [5, 34, 52].

Correspondingly, various methods have been developed to mitigate the costs of experiments. These include experiment platform enhancement [26], sample size optimization [15, 44], strategic treatment allocation [2, 24, 54], adaptive experimental design [9, 31], market-structure-informed experimental design [53], extrapolation from known experiments [16], and post-experiment analysis improvement [23, 40, 43]. The TIE framework shares the goal of this literature to address the costs of experimentation, but takes a different approach. Rather than focusing on improving experimental design or analysis, it relies on theory-based modeling as a potential substitute for experimentation.

Another well-established alternative to experimentation is counterfactual evaluation, which analyzes offline (i.e., historical) data to estimate the outcomes of policies had they been deployed online (i.e., in real environments); see [42] for a comprehensive review. A range of counterfactual evaluation techniques has been developed, including the direct method, importance sampling, and doubly robust estimators [13, 20, 51]. Counterfactual evaluation has yielded successful industry applications, with notable use cases including Microsoft Bing Ads [7] and Uber Eats Ads [55].

By contrast, the TIE system for creator valuation is an online simulator. It invokes the deployed click-prediction and content distribution functions in live environments to simulate counterfactual changes in user engagement. Correspondingly, the TIE system offers two advantages. First, because it runs inside the operational system, TIE automatically tracks algorithm updates and adapts to evolving user preferences and platform policies, whereas offline methods are built on historical data and may often need recalibration to handle various shifts in the environment [21, 56]. Second, with direct access to live content distribution functions, TIE retrieves the exact counterfactual actions the platform would have taken rather than approximations simulated from historical data. In this sense, the TIE system moves one step closer to digital twin systems, which emphasize real-time data flow between the physical entity and its simulated counterpart [45].[1]

Lastly, the TIE system uses prediction models to simulate causal experimental results. Prior research has explored various predictive modeling approaches to improve causal inference, including causal forest [3, 4], double machine learning [10, 14], and meta-learners [28]. Models in this stream of literature generally rely on exogeneity assumptions to establish causal relationships due to the fundamental problem of causal inference—the inability to observe multiple potential outcomes for the same observation unit [19, 41]. In comparison, the TIE system directly simulates counterfactual responses for each unit (e.g., user) based on models of the counterfactual environment, thus eliminating the explicit need for exogeneity assumptions.

3 Application Context: Creator Valuation

Content creators are the driving force behind the rapidly growing content economy [46]. The success of content platforms depends on

[1] This online simulation approach has limitations relative to offline simulation methods. Because the TIE system operates on live data and live system logic, it cannot be used to simulate arbitrary past policies or retroactively test hypothetical strategies. Additionally, because the click prediction and content distribution logic must be used as is, predefined metrics are necessary, and bootstrapping for inference by replaying the system multiple times is not directly applicable.

the contributions of these creators, and Tencent's WeChat Official Accounts Platform is no exception. As one of the largest content platforms in the world, WeChat Official Accounts serves over a billion users and hosts tens of millions of creators (called "official accounts"), who publish hundreds of millions of content items annually.[2] However, as in many ecosystems, not all creators contribute equally. One of Tencent's current business priorities is to assess the value of each creator, meaning the creator's contribution to user engagement with the platform. Gaining such insights will allow Tencent to better understand and incentivize the creation of valuable content.

Creator valuation is no simple task. A naïve approach is to rely on descriptive statistics, such as the number of clicks and followers. However, this approach does not capture the complex patterns of substitution or complementarity among creators. Even a highly popular creator's value can be diminished if other creators produce similar content. Ultimately, creator value is a counterfactual concept; it is determined by the difference in the utility users derive from the platform with versus without the presence of the focal creator. Simple descriptive statistics may not offer such counterfactual predictions.

Therefore, to assess creator value, Tencent has relied on experiments. This experimental approach involves blocking user access to a set of creators temporarily and observing the resulting change in user engagement (e.g., user clicks). A larger drop in engagement suggests that the blocked creators are more valuable. While experimentation provides the ground truth of creator value, it often comes with several costs.

First, as reviewed in Section 2, experimental design is a complex science. For example, blocked creators must be carefully selected to make the experiment informative without affecting user experience too much [24]. The number of users involved must be cautiously chosen to ensure statistical significance while keeping the interruptions to the platform manageable [15]. Moreover, observations from each experiment take time to stabilize. These experiments must be carefully designed to ensure the collection of stable observations without obstructing content to a degree that degrades creator experience. Implementing these experiments also requires nontrivial effort. Numerous experiments are needed, each blocking a different combination of creators, to fully uncover patterns of creator substitution or complementarity [49]. This can be computationally burdensome because computational complexity increases with the number of experiments. The experiments also need to be monitored in case adjustments are needed [9].

Second, these experiments carry opportunity costs. A creator is only identified as valuable if access to the creator is varied and a significant change in user engagement is observed. Such variations often require deviations from the optimal content distribution mechanism. In other words, information is gained at the expense of real business performance. This expense can be as high as a 2.5% loss of user clicks for a typical experiment to assess creator value on the platform.

[2]Exact numbers are withheld for confidentiality. According to public sources, as of 2024, WeChat Official Accounts had more than 1.37 billion users, 20 million creators, and 444 million pieces of content generated annually; see https://news.qq.com/rain/a/20250119A03DEG00 for further details (accessed on June 5, 2025).

Additionally, two practical challenges arise concerning the value of information that can be obtained through experiments. First, given the massive number of WeChat creators, a sufficiently large group of creators must be blocked at once to produce a measurable impact. As a result, each experiment usually only evaluates the aggregate value of a group of creators rather than the value of individual creators. This limits the platform's ability to run personalized campaigns, such as monetary rewards or traffic incentives tailored to individual creators.

Second, as mentioned, experimental results take time to stabilize. Therefore, a creator's value can only be assessed experimentally after a delay. This not only increases system latency, limiting business applications that need real-time access to creator value, but may also yield misleading information if creator value fluctuates over time, as we demonstrate is the case.

To circumvent the costs of experiments while retaining their informational value or even enhancing their informational value, we develop the TIE system as a new solution to the problem of creator valuation. We present the construction of the TIE system in Section 4 and its test in Section 5.

4 A TIE System for Creator Valuation

4.1 Theory Model

The TIE system formulates creator valuation as quantifying the change in user engagement if a creator were to leave the platform (the counterfactual scenario). Grounded in micro-level behaviors, the TIE system centers on a theory model of how users, the platform, and other creators would behave in this counterfactual scenario.

On the content-demand side, we model user clicks based on the standard microeconomic theory of user utility maximization. In each usage session t, user i clicks on content $j \in L_{it}$ if and only if the associated consumption utility, $U_{ijt} = f(X_{ijt})$, exceeds a reservation utility to be estimated. The set L_{it} denotes the list of content the platform shows to user i during session t. The matrix X_{ijt} includes characteristics of user i, features of content j, effects of session t, and any potential interaction effects among the three. For example, X_{ijt} may include other available content, besides content j, shown to user i during session t, which may affect the user's consumption utility for content j through substitution or complementarity. The function f is empirically calibrated using data on user-content interactions, where more clicks on a type of content imply higher user utility from consuming this type of content [36].

On the content-supply side, we explicitly construct the content list users would be exposed to in the counterfactual environment where a creator were absent from the platform. Intuitively, to assess a creator's value, it is critical to know the alternative content options available to users in the absence of that creator. We construct this counterfactual content list by re-optimizing the platform's content distribution mechanism over the available content set. Once this counterfactual content list is known, we apply the calibrated f function to impute users' counterfactual utilities and clicks in the absence of the focal creator. The difference between the actual and counterfactual clicks reflects the creator's value, by definition.

To keep the problem tractable, we make the simplifying assumption that other creators do not adjust their posting behavior in

response to the departure of a single creator. Given the large number of creators on WeChat, such a departure may even go entirely unnoticed. We provide empirical evidence supporting this assumption in Appendix A.[3]

The counterfactual prediction outlined above is enabled by the micro-level modeling of behaviors. First, the same f function calibrated from observed user-content interactions can be applied in the counterfactual environment if the f function captures users' utility primitives (e.g., preferences for different content features), which are assumed to remain constant across various policy environments [33, 35]. Second, we construct the exact content list that would have been shown to each user in the counterfactual environment, so that the f function can be applied directly to the list to predict counterfactual utilities and clicks.

In contrast, an ad-hoc model without a microfoundation may not extrapolate well to the counterfactual environment. For example, consider an ad-hoc model that links historical user clicks to creator identity. Such a model may fail to predict the counterfactual clicks where a creator is absent—user clicks for the remaining creators may have changed depending on the substitution or complementarity among creators, and the list of available content may have changed after re-optimizing the platform's content distribution mechanism. In other words, user clicks and content lists may not be policy-invariant, whereas user preferences that lead to the clicks may be policy-invariant, and we have the added benefit of observing the platform's exact content list under the counterfactual policy. We will report test results of both the TIE system and the ad-hoc model in Section 5 for further comparison.

4.2 Engineering Architecture

To implement the theory model of content demand and supply as outlined above, the TIE system calls two predefined functions: the prediction function and the matching function. These two functions capture user preference f on the demand side and the platform's content distribution mechanism on the supply side, respectively.

For each user exposure to a piece of content, the prediction function takes as input the user, the content, and context (such as time and user exposure history) to predict the click-through rate (pCTR) for that exposure. The prediction function is trained on vast amounts of user-content interaction (e.g., clicks) data.

For each user session, the matching function generates a list of content to which the user will be exposed, taking the user, context, and the content pool available at the time of the session as input. The objective of the matching function is to maximize the pCTR in that session, while considering factors such as content consumption diversity and creator incentives based on the platform's content distribution policy of the time.

While the prediction and matching functions can be developed from scratch, one advantage of the TIE system is that these two functions are often readily available for a content platform. At Tencent, these two functions have served as essential components of its content recommendation system on the WeChat platform. Therefore, the TIE system calls these two functions in real time. Doing so not only substantially reduces the development cost of the TIE system but also, as discussed, enables access to user click predictions and the platform's counterfactual gameplay in the precise instance of a creator's departure, rendering TIE an online simulator. For confidentiality, we do not report the technical details of these two functions.

The TIE system consists of two modules: the factual module and the counterfactual module. Each module calls both the prediction function and the matching function. While some form of the factual module often exists for content platforms, we have uniquely created the counterfactual module for the TIE system to simulate outcomes in the counterfactual environment if a creator were to leave the platform.

In the factual module, for each user session, the matching function first takes the user, context, and the entire content pool as input and outputs the content list to which the user will be exposed. This list is evaluated by the prediction function, generating a pCTR for the session, which is the sum of pCTRs of all individual user-content pairs in that session. Both the content list and the pCTR of the user session are stored in the system database.

Once the factual content list is generated, the counterfactual module starts running. For each piece of content in the factual content list, the counterfactual module first identifies all pieces of content produced by the same creator and removes them from the original content pool to construct the corresponding counterfactual content pool. This counterfactual content pool, along with user and context information, is then fed into the matching function to generate the counterfactual content list for the user session—the content list that would have been shown to the user had that focal creator been absent from the platform. This counterfactual content list is then evaluated by the prediction function to estimate the corresponding pCTR for the user session. Both the counterfactual content list and its pCTR of the user session are stored in the system database.

The counterfactual module runs separately for every piece of content in the factual content list. The differences between the factual pCTRs and the counterfactual pCTRs represent the values of the creators in that session's content list. We do not need to run the counterfactual module for content that is not part of the factual content list. The corresponding creator values are zero by construction; blocking these creators would have had no impact on the content actually shown in the user session.

Summing a creator's values across all user sessions yields the TIE prediction of the creator's overall value on the platform. To minimize computational load, we implement the TIE system on a randomly selected representative sample of users. We select the sampling rate by gradually increasing it until creator value estimates stabilize, resulting in a final sampling rate of $r \approx 0.8\%$ drawn from the total user base. The resulting creator value derived from this sample is then scaled by a factor of $1/r$ to derive creator value for the entire platform.

Algorithm 1 outlines the TIE system pseudocode as described above. Figure 1 presents a visual illustration of the TIE system architecture. Again, a key feature of the TIE system is that both the factual and counterfactual modules call the same two predefined processes: the prediction function and the matching function. While the *outputs* of these two functions may vary between the factual and counterfactual environments, these two *functions*

[3]We similarly assume that other creators do not change their posting behaviors in response to users' or the platform's reactions to a single creator's absence.

themselves remain the same. Furthermore, if these two functions remain valid descriptions of user preferences and content distribution mechanisms in the counterfactual environment, they satisfy the policy-invariance property that enables accurate predictions of the counterfactual environment [33, 35]. We discuss policy invariance in the following section.

Algorithm 1: TIE System for Creator Valuation

Input:
T: Set of user sessions from a random user sample (sample rate r)
J_t: Content pool at session t
$P(i, \text{ctx}, j)$: Prediction function
$M(i, \text{ctx}, J)$: Matching function
Output:
V: Mapping from creator to creator value
Initialization:
$\forall$ creator c, $V(c) \leftarrow 0$
foreach *session* $t \in T$ **do**
  $i \leftarrow t$.user
  ctx $\leftarrow t$.context
  `// Factual module`
  $L_F \leftarrow M(i, \text{ctx}, J_t)$
  $pCTR_F \leftarrow 0$
  foreach *content item* $j \in L_F$ **do**
    $pCTR_F \leftarrow pCTR_F + P(i, \text{ctx}, j)$
  end
  store $(L_F, pCTR_F)$
  `// Counterfactual module`
  $C \leftarrow \{\, j.\text{creator} \mid j \in L_F \,\}$
  foreach *creator* $c \in C$ **do**
    $J_{CF} \leftarrow \{\, j' \in J_t \mid j'.\text{creator} \neq c \,\}$
    $L_{CF} \leftarrow M(i, \text{ctx}, J_{CF})$
    $pCTR_{CF} \leftarrow 0$
    foreach *item* $j' \in L_{CF}$ **do**
      $pCTR_{CF} \leftarrow pCTR_{CF} + P(i, \text{ctx}, j')$
    end
    store $(L_{CF}, pCTR_{CF})$
    $\Delta \leftarrow pCTR_F - pCTR_{CF}$
    $V(c) \leftarrow V(c) + \Delta$
  end
end
`// Rescale for sampling rate`
foreach *creator* c **do**
  $V(c) \leftarrow V(c)/r$
end
return V

4.3 Policy Invariance

Policy invariance is strictly satisfied for the matching function because the same matching function is invoked directly to produce the content list in the counterfactual module, describing the platform's precise content distribution strategy in the event of a creator's departure. As discussed, this is a unique advantage of the TIE system over typical structural models that rely on assumptions about supply-side strategies [17]. In terms of model variance [42], policy invariance of the matching function reduces the uncertainty of the TIE system to that of the prediction function.

For the prediction function to satisfy policy invariance, user preferences as specified in the prediction function must remain the same in the counterfactual environment. We explain why this assumption is likely to hold in the context of this paper.

First, the prediction function is an industry-level, high-fidelity model developed specifically to capture user preference primitives and predict user engagement in various environments. It is a transformer-based model with billions of parameters, trained on combinations of naturally occurring and experimental data to measure arguably-exogenous user preferences, and is continuously trained on incoming user-content interaction data to capture changes in user preferences. In subsequent experiments (Section 5), we find a 99.8% correlation between the total pCTR across sessions in each of 168 experiments and the observed clicks, suggesting high predictive accuracy of the prediction function across a range of domains.

Second, the counterfactual scenario of interest corresponds to the departure of one creator. It is unlikely that users will reformulate their preferences [8] following the departure of one among tens of millions of creators. They may not even be aware of a creator's departure because a typical browsing session only features a small subset of creators on the platform.

Third, we nevertheless stress-test the prediction function in a wide range of alternative contexts. At Tencent, content can be delivered through multiple channels besides the recommendation channel this paper focuses on (see Section 5). These include subscriptions to creator accounts, social sharing from friends, and cold-start promotions. In total, the platform supports over 15 distinct content delivery channels (exact number and details withheld for confidentiality). The prediction function achieves above 99% correlations with actual clicks in the majority of these channels, and falls below 90% in only one channel (with an 88% correlation). The average correlation across all channels is 98.2%. These results suggest that user preferences captured in the prediction function are likely to be valid across a diverse range of contexts. In fact, if user preferences remain valid in contexts distinct from the recommendation channel, they may arguably remain valid with respect to a small perturbation (i.e., the absence of one creator) within the recommendation channel.

4.4 Advantages Over Experiments

The TIE system for creator valuation offers several advantages over experiments. First, experiments require a significant human effort to design and implement. In comparison, the TIE system is fully automated, its only additional cost being the backend computational cost to execute the counterfactual module. The additional load on the computation cluster is under 1%.

Second, the TIE system carries no opportunity cost. Users are exposed to the same factual content lists generated by the existing

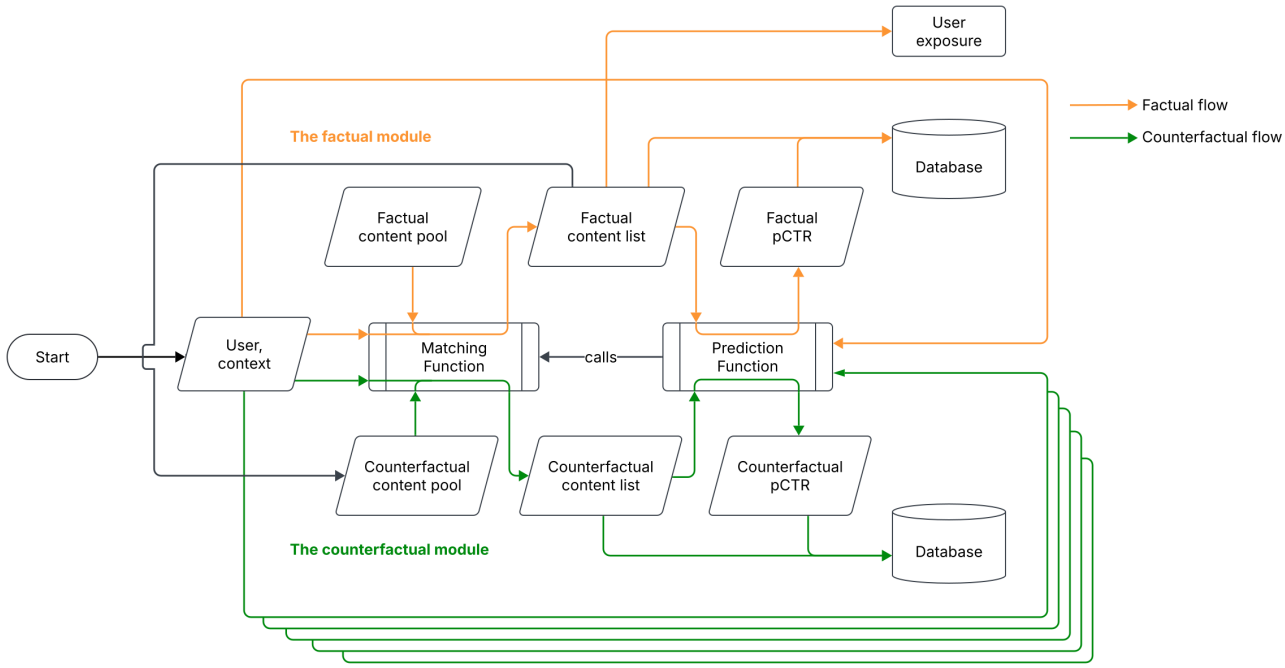

Figure 1: TIE System Architecture. For each user session, the factual module generates the factual content list to be shown to the user and calculates the corresponding predicted click-through rate (pCTR). A counterfactual module runs for each piece of content in the factual content list, generating the counterfactual content list and calculating the respective pCTR. Both the factual and counterfactual content lists, along with their pCTRs, are stored in the system database. The differences between the factual and counterfactual pCTRs are used to calculate creator value. The user is exposed only to the factual content list.

recommendation system, so that their experience remains uninterrupted. The counterfactual content lists are computed entirely in the backend, eliminating any interference with users.

Additionally, experiments on the platform can realistically only assess the aggregate value of a group of creators, making it difficult to differentiate the values of individual creators within the group. In contrast, the TIE system provides individual-level creator value assessments, enabling more personalized insights. Meanwhile, as explained earlier, experiments on this platform often require days or weeks to yield stable results, whereas the TIE system operates in real time, providing up-to-date assessment of creator value almost instantaneously.

It remains to test the TIE system's accuracy in assessing creator value. We present the test and its results next.

5 Testing the TIE System

We test the predictive accuracy of the TIE system against ground-truth data—data from experiments. In collaboration with Tencent, we conduct 168 such large-scale experiments on its WeChat Official Accounts Platform, following its current experimental methodology for creator valuation. (For confidentiality, we only present information about these experiments that are essential for assessing the performance of the TIE system.)

When users arrive on the platform, they first see new content from creators they subscribe to (subscription flow). After that, they see additional content that the recommender system chooses to show them (recommendation flow). While content can also be delivered through other channels such as social sharing or cold-start promotions, at the time of the test, over half of content exposures come from the recommendation flow. We focus on the recommendation flow, which is the primary application context for creator valuation at Tencent.

In each experiment, we randomly select a sample of active users to form the treatment group and block a specific set of creators from these users' recommendation flow. The blocked sets of creators are intentionally varied along meaningful dimensions across the 168 experiments, allowing the TIE system to be tested across a broad range of scenarios. To form the control group, we randomly select another set of active users of the same sample size, without blocking any creators. There is no overlap among the sets of users across experiments that overlap in time. The primary outcome of interest for an experiment is the change in user clicks—measured as user clicks in the treatment group minus user clicks in the control group—within the recommendation flow. Each experiment involves more than 9 million users in the treatment group and temporarily blocks user access to the involved creators for one day. Table 4 in Appendix B presents the summary statistics of the experiments.

The TIE system predicts creator value at the individual-creator level. We sum the values of all creators blocked in an experiment to generate the TIE-predicted click change for that experiment. This

summation should be viewed as an approximation because it does not fully account for potential substitution or complementarity among blocked creators. But the bias introduced is minimal if the number of blocked creators is small relative to the entire pool of creators (see Table 4), so that they are unlikely to show up in the same content list for many user sessions. To test the predictive accuracy of the TIE system, we compare the observed click change in each experiment with the corresponding TIE-predicted click change for that experiment. We scale both measures as percentages (or change rates) to preserve confidentiality.

Figure 2 plots the observed click change rate for each experiment against the corresponding click change rate predicted by TIE. The correlation between the TIE predictions and the observed experimental results is 97%. We run an ordinary least squares (OLS) regression with the observed click change rate as the dependent variable and the TIE prediction as the independent variable, as well as a constant term. The slope of the fitted line is 1.1236, close to 1, the value if the prediction is perfect.

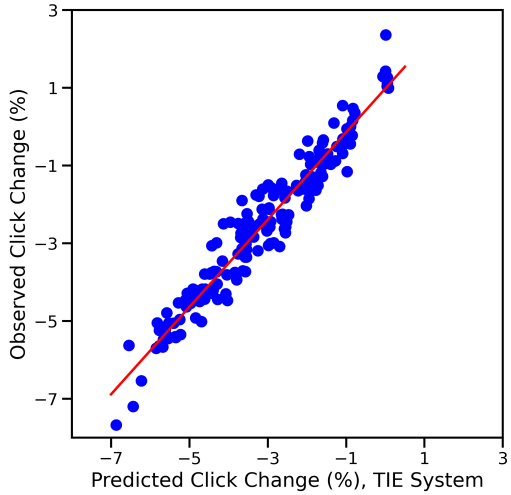

Figure 2: TIE-System Prediction vs. Experiment Outcome. The x-axis is an experiment's click change rate predicted by the TIE system, and the y-axis is the observed click change rate for that experiment. Each dot represents an experiment. The fitted line is based on OLS regression of y on x with a constant. The x and y axes are set on the same scale, so that perfect prediction corresponds to the diagonal line.

As discussed, each of the 168 experiments blocks a different mix of creators. This, by intention, leads to noticeable variation in observed click change rates across experiments. The TIE system accurately predicts the click change rate over the range of experiments, indicating strong robustness. Notably, in six experiments (represented by the six dots in the upper right of Figure 2), the blocked creators are selected from the bottom 10% of seven-day creator value predicted by TIE. Their associated positive click changes suggest that removing less valuable creators improves user engagement. The TIE system is able to predict this nuanced result.

For completeness, we test the predictive accuracy of the ad-hoc model described in Section 4, one that uses the total number of clicks for a creator's content as a measure of the creator's value. The click change rate predicted by this ad-hoc model is the sum of clicks for the blocked creators divided by the sum of clicks for all creators in the control group (using the sum of clicks in the treatment group yields almost identical results, which is to be expected given the randomized nature of the experiments).

Figure 3 plots the results. Relative to Figure 2, the ad-hoc model predicts higher creator values (i.e., larger click change rates upon blocking creators) than what the experiments indicate. The ad-hoc model's OLS-regression coefficient is 0.2340, which, compared with the coefficient of 1.1236 for TIE, means that the ad-hoc model overestimates creator values by almost four times. Further comparison indicates that the mean absolute error (MAE) of the ad-hoc model (0.1806) is 28 times the MAE of TIE (0.0065). As discussed, the ad-hoc model is expected to predict less accurately than TIE given its inability to account for substitution or complementarity among creators. The fact that the ad-hoc model overestimates creator values suggests that it likely fails to capture creator substitution.

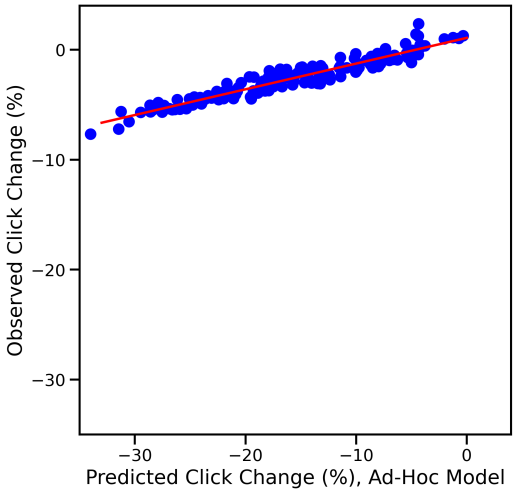

Figure 3: Ad-Hoc Model Prediction vs. Experiment Outcome. The x-axis is an experiment's click change rate predicted by the ad-hoc model, and the y-axis is the observed click change rate for that experiment. Each dot represents an experiment. The fitted line is based on OLS regression of y on x with a constant. The x and y axes are set on the same scale, so that perfect prediction corresponds to the diagonal line.

Table 1 summarizes the comparison between the experiments, the TIE system, and the ad-hoc model. Fully automated and running in the backend, the TIE system circumvents the implementation and opportunity costs of experiments while being reasonably accurate in assessing creator value and noticeably more accurate than the ad-hoc model. Moreover, the TIE system produces results that are both more granular and more immediate than those from experiments, enabling creator value assessment at the individual rather than group level, and within milliseconds rather than days or weeks.

To appreciate the importance of individual-level creator valuation, we examine the variability in creator value within the same group of creators simultaneously blocked in an experiment. As

Table 1: Comparison Between Experiments, the TIE System, and the Ad-Hoc Model

	Experiments	TIE System	Ad-Hoc Model
Implementation	Nontrivial human effort	Automated	Automated
Opportunity Cost	2.5% click decrease	0	0
OLS Coefficient	1	1.1236	0.2340
MAE	0	0.0065	0.1806
Prediction Granularity	Aggregate level	Individual level	Individual level
Latency (Time per Assessment)	Days or weeks	Milliseconds	Milliseconds

shown in Table 4 in the appendix, each of the 168 experiments blocks an average of 4,703 creators. For each creator, we compute the weekly average creator value over an eight-week period using the TIE system (we do so because creator value fluctuates over time, as shown below). Across the 168 experiments, the average coefficient of variation for creator value within the same group of creators reaches 11,538%. To put this number into perspective, a 30% coefficient of variation often indicates high variability [11]. This substantial within-group variability in creator value means that group-level, aggregate creator value produced by the experiments offers poor guidance for individual-level valuation.

Similarly, the TIE system's minimal latency not only enables applications that require real-time creator valuation but is also critical for ensuring accuracy when creator value fluctuates over time. To assess the extent of creator value fluctuation, we again analyze each creator's weekly average value in the TIE system over an eight-week period. Averaged across creators, we observe a 298% coefficient of variation over time. Furthermore, we examine the correlation between a creator's value in the first week and subsequent weeks. As shown in Table 2, across creators, the average creator value correlation between week one and week two is only 1.6%; it even becomes negative after five weeks. This result suggests high temporal volatility of creator value, so that one-time experiments may provide misleading insight. This result also suggests that, even if the ad-hoc model shows correlation with observed changes in clicks, it cannot be easily corrected with a simple scaling factor to duplicate the performance of the TIE system. Such a factor would itself vary substantially over time.

Table 2: Average Correlation Between Creator Value in Week 1 and Subsequent Weeks

Week Pair	Correlation	P-Value
Week 1 vs. Week 2	0.0160	0.000
Week 1 vs. Week 3	0.0081	0.000
Week 1 vs. Week 4	0.0165	0.000
Week 1 vs. Week 5	0.0338	0.000
Week 1 vs. Week 6	-0.0000	0.997
Week 1 vs. Week 7	-0.0000	0.997
Week 1 vs. Week 8	-0.0045	0.000

6 Deploying the TIE System

Based on its cost savings, predictive accuracy, granularity, and minimal latency compared with experiments, Tencent has deployed the TIE system on a full scale as a default approach to assess the value of all creators on its WeChat Official Accounts Platform. Since deployment, the TIE system has evaluated tens of millions of creators a day. Deploying TIE has saved the platform both implementation costs and the opportunity costs of experimentation, where a typical experiment to assess creator value would have otherwise caused a loss of 2.5% in user clicks. The TIE system has supported applications such as creator discovery and content incentivization.

In terms of system maintenance, the TIE system is a real-time online simulator, which directly calls the actively deployed and continuously updated prediction and matching functions. As a result, the TIE system requires no explicit recalibration—any algorithmic changes in the system are automatically reflected in the live functions and, in turn, the TIE system's outputs. Since its full-scale deployment in January 2025 (and as of the date of the current version of this paper in June 2025), the TIE system has not required any manual updates.

In terms of predictive stability, the baseline TIE system presented in this paper provides only a point estimate of creator value, but it can be extended to support statistical inference using common techniques such as model ensembles. For instance, one can independently train multiple versions of the prediction model (e.g., tens of neural networks with different random initializations or training splits, depending on computational resources). Running the full TIE pipeline with each model in the ensemble yields a distribution of creator value estimates. The average of these estimates serves as the point estimate, while the variance across models quantifies the uncertainty for applications that are sensitive to predictive stability. Again, a helpful feature of the TIE system is its access to the platform's exact counterfactual content distribution strategy. Therefore, the uncertainty in the TIE system arises solely from the prediction model, which remains low due to the prediction model's high accuracy.

One byproduct of the TIE system useful for certain business applications is its ability to explicitly capture substitution and complementarity effects between creators. When a creator is blocked, the counterfactual module logs which content takes its place, allowing the application to trace how user traffic is redistributed. By aggregating this data across sessions and mapping it to individual creators, one can construct a creator-to-creator substitution matrix of size $N \times N$ (where N is the number of creators). This matrix can serve as input to business applications that must account for externalities among creators, such as targeted incentive schemes designed to generate positive ripple effects across the creator ecosystem.

7 Conclusion

Experiments provide direct insights into the counterfactual world (e.g., what would happen if a creator were to leave the platform). However, they often come with significant implementation and opportunity costs. To address this gap, we develop the TIE framework as a substitute for experiments. If a theory can confidently describe the counterfactual environment, its predictions can serve as a reliable alternative for gaining counterfactual insights without the need for costly experiments.

In collaboration with Tencent, we apply the TIE framework to build an automated creator valuation system for its WeChat Official Accounts Platform. The TIE system is grounded in theories of utility maximization on the content-demand side and platform optimization on the content-supply side. The corresponding content demand model calibrated on factual data can plausibly—as we demonstrate—be used to accurately predict counterfactual demand. Moreover, knowing the platform's optimization model enables exact specification of the platform's counterfactual content supply. A creator's value is then derived, in real time, by comparing actual user clicks in the creator's presence with predicted clicks in response to the counterfactual content supply.

We test the TIE system with 168 experiments that block user access to various groups of creators in a controlled manner. The TIE system predicts experimental outcomes well, achieving a 97% correlation, while circumventing the costs of these experiments. Moreover, the TIE system delivers individual-level creator valuations with minimal latency, which greatly enhances the usability of the results given the substantial variation in creator value across individuals and over time. As a result, Tencent has deployed the TIE system as the default for creator valuation on its WeChat Official Accounts Platform, assessing the value of tens of millions of creators each day.

In the Tencent application, managerial judgment determines whether the TIE system is an adequate substitute for experiments (e.g., 97% correlation is considered good performance). One direction for future development is to automate this judgment process by quantifying the cost of experiments and their informational gain relative to known theories. Another direction is to automate theory improvement when its prediction falls short.

The TIE system is applicable beyond creator valuation. For example, it can be adapted to assess seller value on e-commerce platforms, where the opportunity costs of experiments include direct monetary consequences. More generally, the TIE system can be used to determine the marginal value of the members of an ecosystem, provided that two prerequisites are met. First, the TIE system relies on the ability to simulate exact counterfactual actions by calling the live decision function (e.g., the matching function). Such a function is in place for most companies with automated decisions and does not need to be sophisticated. The TIE system's minimum requirement is real-time access to this function. Second, the model of user behaviors (i.e., the prediction function) should be reasonably accurate—any prediction error will propagate through the simulation to affect performance. However, simpler prediction models can be used depending on the company's capabilities. The TIE system's minimum requirement is a prediction model that is likely to generalize out of sample. Methodologically, the TIE framework can also help identify and mitigate interference in network experiments, where theory is used to model how different nodes respond to various policy mixtures.

Acknowledgments

The authors thank the WeChat Official Accounts Platform Team at Tencent, especially Dehao Wu, Zebo Wu, Shichao Han, and Kangyi Lin, for their support. The authors received helpful comments from Jeremy Yang, the KDD '25 Reviewers, and the Area Chair. Lei Huang was a research intern at Tencent when the work was conducted. The views and conclusions expressed in this paper are those of the authors and do not represent the views of Tencent. All data presented in this paper have been anonymized and sanitized to protect user privacy, and no personally identifiable information is included or can be inferred from the data.

References

[1] Joshua D. Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, Princeton, NJ, USA.
[2] Eva Ascarza. 2018. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55, 1 (Feb. 2018), 80–98.
[3] Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *The Annals of Statistics* 47, 2 (Apr. 2019), 1148–1178.
[4] Susan Athey and Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. *Observational Studies* 5, 2 (2019), 37–51.
[5] Jackie Baek and Vivek F. Farias. 2024. Fair exploration via axiomatic bargaining. *Management Science* 70, 12 (Dec. 2024), 8922–8939.
[6] Sanjay Basu, Jeremy B. Sussman, and Rod A. Hayward. 2017. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment: A modeling study of randomized clinical trials. *Annals of Internal Medicine* 166, 5 (Jan. 2017), 354–360.
[7] Murat Ali Bayir, Mingsen Xu, Yaojia Zhu, and Yifan Shi. 2019. Genie: An open box counterfactual policy estimator for optimizing sponsored search marketplace. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining* (Melbourne, VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 465–473. https://doi.org/10.1145/3289600.3290969
[8] Xinyu Cao and Juanjuan Zhang. 2021. Preference learning and demand forecast. *Marketing Science* 40, 1 (Jan.–Feb. 2021), 62–79.
[9] Yi Cheng, Fusheng Su, and Donald A. Berry. 2003. Choosing sample size for a clinical trial using decision analysis. *Biometrika* 90, 4 (Dec. 2003), 923–936.
[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (Feb. 2018), C1–C68.
[11] Barbara M. Davit, Dale P Conner, Beth Fabian-Fritsch, Sam H Haidar, Xiaojian Jiang, Devvrat T. Patel, Paul RH. Seo, Keri Suh, Christina L. Thompson, and Lawrence X. Yu. 2008. Highly variable drugs: Observations from bioequivalence data submitted to the FDA for new generic drug applications. *The AAPS Journal* 10 (March 2008), 148–156.
[12] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210 (Aug. 2018), 2–21.
[13] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 1097–1104. https://dl.acm.org/doi/10.5555/3104482.3104620
[14] Max H. Farrell, Tengyuan Liang, and Sanjog Misra. 2025. Deep learning for individual heterogeneity: An automatic inference framework. arXiv:2010.14694 https://arxiv.org/abs/2010.14694
[15] Elea McDonnell Feit and Ron Berman. 2019. Test & roll: Profit-maximizing A/B tests. *Marketing Science* 38, 6 (Nov. 2019), 1038–1058.
[16] Brett R Gordon, Robert Moakler, and Florian Zettelmeyer. 2023. Predictive incrementality by experimentation (PIE) for ad measurement. arXiv:2304.06828 https://arxiv.org/abs/2304.06828
[17] Liang Guo. 2006. Removing the boundary between structural and reduced-form models. *Marketing Science* 25, 6 (Nov.–Dec. 2006), 629–632.
[18] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (June 2019), 20–35.

[19] Paul W. Holland. 1986. Statistics and causal inference. *J. Amer. Statist. Assoc.* 81, 396 (Dec. 1986), 945–960.

[20] Daniel G. Horvitz and Donovan J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 260 (Dec. 1952), 663–685.

[21] Chih-Wei Hsu, Martin Mladenov, Ofer Meshi, James Pine, Hubert Pham, Shane Li, Xujian Liang, Anton Polishko, Li Yang, Ben Scheetz, and Craig Boutilier. 2024. Minimizing live experiments in recommender systems: User simulation to evaluate preference elicitation policies. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington, DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2925–2929. https://doi.org/10.1145/3626772.3661358

[22] Guido W. Imbens. 2022. Causality in econometrics: Choice vs chance. *Econometrica* 90, 6 (Nov. 2022), 2541–2566.

[23] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2022. Always valid inference: Continuous monitoring of A/B tests. *Operations Research* 70, 3 (May 2022), 1806–1821.

[24] Toru Kitagawa and Aleksey Tetenov. 2018. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86, 2 (March 2018), 591–616.

[25] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. Association for Computing Machinery, New York, NY, USA, 1168–1176. https://doi.org/10.1145/2487575.2488217

[26] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery* 18 (Feb. 2009), 140–181.

[27] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, UK.

[28] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (March 2019), 4156–4165.

[29] Riccardo Leardi. 2009. Experimental design in chemistry: A tutorial. *Analytica Chimica Acta* 652, 1-2 (Oct. 2009), 161–172.

[30] Steven D. Levitt and John A. List. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* 53, 1 (Jan. 2009), 1–18.

[31] Gui Liberali, Eric Boersma, Hester Lingsma, Jasper Brugts, Diederik Dippel, Jan Tijssen, and John Hauser. 2025. Real-time adaptive randomization of clinical trials. *Journal of Clinical Epidemiology* 178 (Feb. 2025), 111612.

[32] Michael Luca and Max H. Bazerman. 2021. *The Power of Experiments: Decision Making in a Data-Driven World*. The MIT Press, Cambridge, MA, USA.

[33] Robert E. Lucas Jr. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1 (1976), 19–46.

[34] Alexander R. Luedtke and Mark J. van der Laan. 2016. Optimal individualized treatments in resource-limited settings. *The International Journal of Biostatistics* 12, 1 (May 2016), 283–303.

[35] Jacob Marschak. 1953. Economic measurement for policy and prediction. In *Studies in Econometric Method*, W. C. Hood and T. C. Koopmans (Eds.). Wiley, New York, NY, USA, 1–26.

[36] Daniel McFadden. 1986. The choice theory approach to market research. *Marketing science* 5, 4 (Fall 1986), 275–297.

[37] Adrian C. Melissinos and Jim Napolitano. 2003. *Experiments in Modern Physics*. Academic Press, San Diego, CA, USA.

[38] Julie Ann Miller. 2022. *A tale of two bomb designs: Why were both Little Boy and Fat Man created?* Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM, USA. doi:10.2172/1875784

[39] Richard J. Murnane. 2010. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, New York, NY, USA.

[40] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 235–244. https://doi.org/10.1145/2939672.2939688

[41] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (Dec. 1974), 688–701.

[42] Yuta Saito and Thorsten Joachims. 2022. Counterfactual evaluation and learning for interactive systems: Foundations, implementations, and recent advances. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 4824–4825. https://doi.org/10.1145/3534678.3542601

[43] Duncan Simester, Artem Timoshenko, and Spyros I. Zoumpoulis. 2020. Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science* 66, 8 (Aug. 2020), 3412–3424.

[44] Duncan Simester, Artem Timoshenko, and Spyros I Zoumpoulis. 2025. A sample size calculation for training and certifying targeting policies. *Management Science* (forthcoming 2025).

[45] Maulshree Singh, Evert Fuenmayor, Eoin P. Hinchy, Yuansong Qiao, Niall Murray, and Declan Devine. 2021. Digital twin: Origin to future. *Applied System Innovation* 4, 2 (June 2021), 36.

[46] Wentao Su and Weitao Duan. 2024. Improving ego-cluster for network effect measurement. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 5713–5722. https://doi.org/10.1145/3637528.3671557

[47] Hao Sun, Evan Munro, Georgy Kalashnov, Shuyang Du, and Stefan Wager. 2024. Treatment allocation under uncertain costs. arXiv:2103.11066 https://arxiv.org/abs/2103.11066

[48] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '10)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/1835804.1835810

[49] The Guardian Staff. 2021. *Google admits to running experiments which remove some media sites from its search results*. Retrieved May 24, 2025 from https://www.theguardian.com/technology/2021/jan/13/google-admits-to-running-experiments-which-remove-some-media-sites-from-its-search-results

[50] Stefan H. Thomke. 2020. *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Press, Brighton, MA, USA.

[51] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. 2022. A review of off-policy evaluation in reinforcement learning. arXiv:2212.06355 https://arxiv.org/abs/2212.06355

[52] U.S. Food and Drug Administration. 2018. *Adaptive designs for clinical trials of drugs and biologics*. Technical Report. U.S. Department of Health and Human Services, Center for Drug Evaluation and Research (CDER), Hillandale, MD, USA.

[53] Stefan Wager and Kuang Xu. 2021. Experimenting in equilibrium. *Management Science* 67, 11 (Nov. 2021), 6694–6715.

[54] Yizhe Xu, Tom H. Greene, Adam P. Bress, Brian C. Sauer, Brandon K. Bellows, Yue Zhang, William S. Weintraub, Andrew E. Moran, and Jincheng Shen. 2022. Estimating the optimal individualized treatment rule from a cost-effectiveness perspective. *Biometrics* 78, 1 (March 2022), 337–351.

[55] Yue Yin. 2023. *Accelerating advertising optimization: Unleashing the power of ads simulation*. Retrieved May 24, 2025 from https://www.uber.com/blog/unleashing-the-power-of-ads-simulation

[56] Shuxi Zeng, Murat Ali. Bayir, Joseph J. Pfeiffer III, Denis Charles, and Emre Kiciman. 2021. Causal transfer random forest: Combining logged data and randomized experiments for robust prediction. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 211–219. https://doi.org/10.1145/3437963.3441722

A Other Creators Do Not Respond to A Creator's Absence

For tractability, the TIE system makes a simplifying assumption that blocking a single creator does not affect the posting behaviors of other creators. With tens of millions of creators in the WeChat ecosystem, this assumption should generally hold.

Nevertheless, we supplement this argument with a conservative test. We make use of the same 168 experiments conducted to test the predictive accuracy of the TIE system (see Section 5). These experiments were conducted in early 2025. We test whether creators who were never blocked showed any change in their content posting behaviors during the experimental period. We use data from both 2024 and 2025, with the 2024 data serving as the control group, to control for any seasonality effect that may coincide with the experimental period. We focus on an approximately six-week window in each year surrounding the calendar dates of the experimental period; a longer window would have overlapped with another major holiday.

This design enables a difference-in-differences regression at the daily level. The dependent variable is a metric of posting behaviors of non-blocked creators on a day. We obtain data on two such

metrics: the number of non-blocked creators who actively posted content, and the number of posts they published. Both metrics are normalized to the [0, 1] range to preserve confidentiality. The independent variables are a year-2025 dummy, an experimental-period dummy (which equals 1 if the calendar date in either year falls within the experimental period in 2025), and their interaction term, along with day-of-the-week fixed effects.

We conduct a parallel-trend test by regressing either dependent variable on the year and day fixed effects and their interaction terms, excluding data from the experimental period. A joint test shows that the interaction terms are statistically insignificant, which confirms that trends in the outcome variables are comparable between year 2024 and year 2025 over days outside the experimental period. A pre-trend test yields the same conclusion.

Table 3 presents the results of the difference-in-differences regressions. For both metrics of posting behaviors of non-blocked creators, the interaction term between the year-2025 dummy and the experiment-period dummy is statistically insignificant. In other words, there is no significant evidence that non-blocked creators changed their posting behaviors during the experimental period.

The test above is conservative because it examines whether blocking a group of creators, as opposed to a single creator, affects the posting behaviors of other creators. The lack of a significant effect from blocking a group of creators offers further support to the argument that other creators are unlikely to change their posting behaviors in response to a single creator's absence.

B Summary Statistics of the 168 Experiments

Table 4 presents the summary statistics of the 168 experiments on the Tencent WeChat Official Accounts Platform. The variables are at the individual-experiment level.

Table 3: Non-blocked Creators' Responses to the Experiments

	# Active Non-blocked Creators	# Posts by Non-blocked Creators
Year2025 × ExperimentalPeriod	-0.050	-0.010
	(0.080)	(0.080)
Year2025	0.120**	-0.080
	(0.050)	(0.050)
ExperimentalPeriod	-0.200***	-0.280***
	(0.050)	(0.060)
Day-of-the-Week Fixed Effect	Included	Included
# Observations	86	86
R^2	0.512	0.549
Adjusted R^2	0.455	0.495
Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$		

Table 4: Summary Statistics of the 168 Experiments

Variable	Mean	SD	Min	Median	Max	N
# Users Involved in an Experiment	9,445,484.62	32,397.40	9,384,822	9,443,419	9,501,928	168
# Creators Blocked in an Experiment	4,703.07	2,611.57	67	4,290	15,280	168
% Change in User Clicks in an Experiment	-2.54	1.85	-7.67	-2.48	2.36	168