Causal Meta-Mediation Analysis: Inferring Dose-Response Function From Summary Statistics of Many Randomized Experiments

Zenan Wang
UC Berkeley
Berkeley, California
zenan.wang@berkeley.edu

Xuan Yin
Etsy, Inc.
Brooklyn, New York
xyin@etsy.com

Tianbo Li
Etsy, Inc.
Brooklyn, New York
tli@etsy.com

Liangjie Hong
LinkedIn, Inc.
Sunnyvale, California
liahong@linkedin.com

ABSTRACT

It is common in the internet industry to use offline-developed algorithms to power online products that contribute to the success of a business. Offline-developed algorithms are guided by offline evaluation metrics, which are often different from online business key performance indicators (KPIs). To maximize business KPIs, it is important to pick a north star among all available offline evaluation metrics. By noting that online products can be measured by online evaluation metrics, the online counterparts of offline evaluation metrics, we decompose the problem into two parts. As the offline A/B test literature works out the first part: counterfactual estimators of offline evaluation metrics that move the same way as their online counterparts, we focus on the second part: causal effects of online evaluation metrics on business KPIs. The north star of offline evaluation metrics should be the one whose online counterpart causes the most significant lift in the business KPI. We model the online evaluation metric as a mediator and formalize its causality with the business KPI as dose-response function (DRF). Our novel approach, causal meta-mediation analysis, leverages summary statistics of many existing randomized experiments to identify, estimate, and test the mediator DRF. It is easy to implement and to scale up, and has many advantages over the literature of mediation analysis and meta-analysis. We demonstrate its effectiveness by simulation and implementation on real data.

KEYWORDS

causal inference; meta-analysis; mediation analysis; experiment; dose-response function; A/B test; evaluation metric; business KPI

1 INTRODUCTION

Nowadays it is common in the internet industry to develop algorithms that power online products using historical data. The one that improves evaluation metrics from historical data will be tested against the one that has been in production to assess the lift in key performance indicators (KPIs) of the business in online A/B tests. Here we refer to metrics calculated from historical data as *offline* metrics and metrics calculated in online A/B tests as *online* metrics. In many cases, offline evaluation metrics are different from online business KPIs. For instance, a ranking algorithm, which powers search pages in e-commerce platforms, typically optimizes for relevance by predicting purchase or click probabilities of items. It could be tested offline (offline A/B tests) for rank-aware evaluation metrics, for example, normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR) or mean average precision (MAP), which are calculated from the test set of historical purchase or click-through feedback of users. Most e-commerce platforms, however, deem sitewide gross merchandise value (GMV) as their business KPI and test for it online. There could be various reasons not to directly optimize for business KPIs offline or use business KPIs as offline evaluation metrics, such as technical difficulty, business reputation, or user loyalty. Nonetheless, the discrepancy between offline evaluation metrics and online business KPIs poses a challenge to product owners because it is not clear that, in order to maximize online business KPIs, which offline evaluation metric should be adopted to guide the offline development of algorithms.

The challenge essentially asks for the causal effects of increasing offline evaluation metrics on business KPIs, e.g., how business KPIs would change for a 10% increase in an offline evaluation metric. The offline evaluation metric in which a 10% increase could result in the most significant lift in business KPIs should be the north star to guide algorithm development. Algorithms developed offline power online products, and online products contribute to the success of the business (see Figure 1). By noting that online products

can be measured by online evaluation metrics, the online counterparts of offline evaluation metrics, we decompose the problem into two parts. The offline A/B test literature (see, e.g, Gilotte et al. [6]) works out the first part (the black arrow): counterfactual estimators of offline evaluation metrics to bridge the inconsistency between changes of offline and online evaluation metrics. We focus on the second part (the red arrow): the causality between online products (assessed by online evaluation metrics) and the business (assessed by online business KPIs). The offline evaluation metric whose online counterpart causes the most significant lift in online business KPIs should be the north star. Hence, the question for us becomes, how business KPIs would change for a 10% increase in an online evaluation metric.
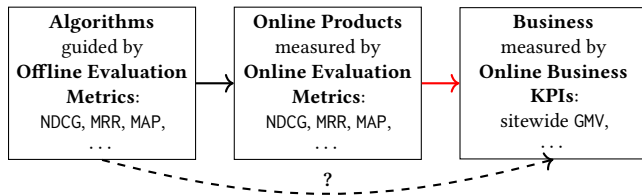
Figure 1: The Causal Path from Algorithms to Business

Randomized controlled trials, or online A/B tests, are popular to measure the causal effects of online product change on business KPIs. Unfortunately, they cannot answer our question directly. In online A/B tests, in order to compare the business KPIs caused by different values of an online evaluation metric, we need to fix the metric at its different values for treatment and control groups. Take the ranking algorithm as an example. If we could fix online NDCG of the search page at 0.22 and 0.2 for treatment and control groups respectively, then we would know how sitewide GMV would change for a 10% increase in online NDCG at 0.2. However, this experimental design is impossible, because most online evaluation metrics depend on users' feedback and thus cannot be directly controlled.

We address the question by developing a novel approach of causal inference. We model the causality between online evaluation metrics and business KPIs by *dose-response function* (DRF) in potential outcome framework [13, 14]. DRF originates from medicine and describes the magnitude of the response of an organism given different doses of a stimulus. Here we use it to depict the value of a business KPI given different values of an online evaluation metric. Different from doses of stimuli, values of online evaluation metrics cannot be directly manipulated. However, they could differ between treatment and control groups in experiments of treatments other than algorithms—user interface/user experience (UI/UX) design, marketing, etc. This could be due to the "fat hand" [19, 29] nature of online A/B tests that a single intervention can change many causal variables at once. A change of the tested feature, which is not algorithm, could induce users to change their engagement with algorithm-powered online products, so that values of online evaluation metrics would change. For instance, in an experiment of UI design, users might change their search behaviors because of the new UI design, so that values of online NDCG, which depends on search interaction, would change, even though ranking algorithm

does not change. The evidence suggests that online evaluation metrics could be *mediators* that (partially) transmit causal effects of treatments on business KPIs in experiments where treatments are not necessarily algorithm-related. Hence, we formalize the problem as the identification, estimation, and test of mediator DRF.

In mediation analysis literature, there are two popular identification techniques: *sequential ignorability* (SI) and *instrumental variable* (IV). SI assumes each potential mediator is independent of all potential outcomes conditional on the assigned treatment, whereas IV permits dependence between unknown factors and mediators but forbids the existence of direct effects of the treatment. Rather than making these stringent assumptions, we leverage trial characteristics to explain *average direct effect* (ADE) in each experiment so that we can tease it out from *average treatment effect* (ATE) to identify the causal mediation. The utilization of trial characteristics means we have to use data from many trials because we need variations in trial characteristics. Hence, we develop our framework as a meta-analysis and propose an algorithm that only uses summarized results from many existing experiments and gain the advantage of easy implementation to scale.

Most meta-analyses rely on summarized results from different studies with different raw data sources. Therefore, it is almost impossible to learn more beyond the distribution of ATEs. Fortunately, the internet industry produces plentiful randomized trials with consistently defined metrics, and thus presents an opportunity for performing a more complicated meta-analysis. Literature is lacking in this area while we create the framework of *causal meta-mediation analysis* (CMMA) to fill in the gap.

Another prominent strength of our approach in real application is, for a new product that has been shipped online but has few A/B tests, it is plausible to explore the causality between its online metrics and business KPIs from many A/B tests of other products. The values of online metrics of the new product can differ between treatment and control groups in experiments of other products ("fat hand" [19, 29]), which makes it possible to solve for mediator DRF of the new product without its own A/B tests.

Note that, our approach can be applied to any evaluation metric that is defined at experimental-unit level, like metrics discussed in offline A/B test literature. The experimental unit means the unit for randomization in online A/B tests. For example, in search page experiments, the experimental unit is typically the user. Also, the evaluation metric can be any combination of existing experimental-unit-level metrics.

To summarize, our contributions in this paper include:

(1) This is the first study that offers a framework to choose the north star among all available offline evaluation metrics for algorithm development to maximize business KPIs when offline evaluation metrics and business KPIs are different. We decompose the problem into two parts. Since the offline A/B test literature works out the first part: counterfactual estimators of offline evaluation metrics to bridge the inconsistency between changes of offline and online metrics, we work out the second part: inferring causal effects of online evaluation metrics on business KPIs. The offline evaluation metric whose online counterpart causes the most significant lift in business KPIs should be the north star. We show the implementation of our framework on data from Etsy.com.

(2) Our novel approach CMMA combines mediation analysis and meta-analysis to identify, estimate, and test mediator DRF. It relaxes standard SI assumption and overcomes the limitation of IV, both of which are popular in causal mediation literature. It extends meta-analysis to solve causal mediation while the meta-analysis literature only learns the distribution of ATE. We demonstrate its effectiveness by simulation and show its performance is superior to other methods.

(3) Our novel approach CMMA uses only trial-level summary statistics (i.e., meta-data) of many existing trials, which makes it easy to implement and to scale up. It can be applied to all experimental-unit-level evaluation metrics or any combination of them. Because it solves for causality problem of a product by leveraging trials of all products, it could be particularly useful in real applications for a new product that has been shipped online but has few A/B tests.

2 LITERATURE REVIEW

We draw on two strands of literature: mediation analysis and meta-analysis. We briefly discuss them in turn.

2.1 Mediation Analysis

Our framework expands on causal mediation analysis. Mediation analysis is actively conducted in various disciplines, such as psychology [15, 24], political science [7, 12], economics [9], and computer science [16]. The recent application in the internet industry reveals the performance of recommendation system could be cannibalized by search in e-commerce website [29]. Mediation analysis originates from the seminal paper of Baron and Kenny [3], where they proposed a parametric estimator based on the linear structural equation model (LSEM). LSEM, by far, is still widely used by applied researchers because of its simplicity. Since then, Robins and Greenland [21] and Pearl [16] and other causal inference researchers have formalized the definition of causal mediation and pinpointed assumptions for its identification [17, 20, 22] in various complicated scenarios. The progress features extensive usage of structural equation models and causal diagrams (e.g., NPSEM-IE of Pearl [16] and FRCISTG of Robins [20]).

As researchers extend the potential outcome framework of Rubin [23] to causal mediation, alternative identification, and more general estimation strategies have been developed. Imai et al. [12] achieved the non-parametric identification and estimation of mediation effects of a single mediator under the assumption of SI. After analyzing other well-known models such as LSEM [3] and FRCISTG [20], they concluded that assumptions of most models can be either boiled down to or replaced by SI. However, SI is stringent, which ignites many in-depth discussions around it (see, e.g., the discussion between Pearl [17, 18] and Imai et al. [11]).

Another popular identification strategy of causal mediation is IV, which is a signature technique in economics [1, 2]. Sobel [26] used treatment as IV to identify mediation effects without SI. However, as Imai et al. [12] pointed out, IV assumptions may be undesirable because they require all causal effects of the treatment pass through the mediator (i.e., complete mediation [3]). Small [25] proposed a new method to construct IV that allows direct effects of the

treatment (i.e., partial mediation [3]) but assumes that ADE of the treatment is the same for different segments of the population.

2.2 Meta-Analysis

Our method only uses summary statistics of many past experiments. Analyzing summarized results from many experiments is termed as meta-analysis and is common in analytical practice [5, 27]. In the literature, meta-analysis is used for mitigating the problem of external validity in a single experiment and learning knowledge that was hard to recover when analyzing data in isolation, such as heterogeneous treatment effects (see, e.g., Browne and Jones [4], Higgins and Thompson [10]). Besides, a significant advantage of meta-analysis is easy to scale, because it only takes summarized results from many different experiments.

Peysakhovich and Eckles [19] took one step toward the direction of performing mediation analysis using data from many experiments. They used treatment assignments as IVs to identify causal mediation, which is similar to Sobel [26], but lacks the justification why more than one experiment is needed and failed to address limitations of IV that we discussed above. Our framework shows that having access to many experiments enables identifying causal mediation without SI and overcoming the limitation of IV, both of which are hard to achieve with only one experiment.

3 CONCEPTUAL FRAMEWORK

We follow the literature of potential outcomes [12, 23, 25] to set up our framework.
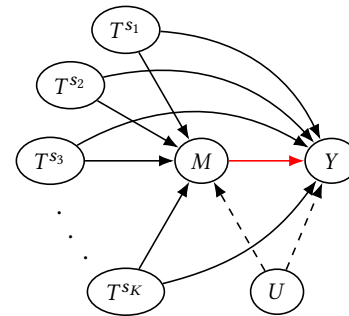
Figure 2: Directed Acyclic Graph of Conceptual Framework

As illustrated in Figure 2, we suppose there are many experiments with different treatments, and there is a mediator M that can be affected by any treatment T^s and will in turn influence outcome Y. Each treatment may also affect Y directly. But we are particularly interested in recovering their shared casual channel, the link between M and Y, marked in red in the figure. In the ranking algorithm example, M is online evaluation metric of search page, Y is online business KPI. A challenge to identify the red link emerges if a confounder U exists. U could be (user engagement of) any other web-page/module or user preference in the ranking algorithm example. In the literature, there are two approaches to solve this challenge, SI and IV. SI requires that U is observed and measured, and there are no other unmeasured/unobserved confounders after controlling for U. The standard IV approach allows unmeasured/unobserved U, but assumes no direct links between

T^ss and Y. An IV method proposed by Small [25] relaxes requirements of standard approach, and assumes all T^ss share a single direct link to Y. Our method allows unmeasured/unobserved U and direct links from each T^s to Y.

Suppose there are K randomized trials in total. For each trial, there exists one treatment group and one control group[1]. To simplify the discussion, here we assume experimental units are first randomly assigned to different trials, and then randomly assigned into the treatment or control group of that trial. However, our approach CMMA allows the same unit to participate in multiple trials. We will go back to this point in Section 5.1.

We consider the following model for potential outcomes.

Definition 1.

$$M_i^{(t,s)} = \boldsymbol{\tau}_i^\top \boldsymbol{s} \times t + \boldsymbol{\phi}^\top \boldsymbol{s} + M_i^* \tag{1}$$

$$Y_i^{(t,s,m)} = \mu_i(m) + \boldsymbol{\gamma}_i^\top \boldsymbol{s} \times t + \boldsymbol{\theta}^\top \boldsymbol{s} + Y_i^*, \tag{2}$$

where the $(\boldsymbol{\tau}_i, \boldsymbol{\gamma}_i, M_i^*, Y_i^*)$ are i.i.d random vectors, for all $t = 0$ or 1, $\boldsymbol{s} \in \mathcal{S}$ and $m \in \mathcal{M}$.
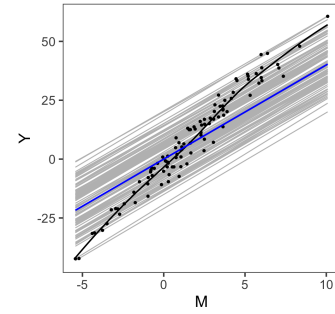
We use a one-hot vector $\boldsymbol{s} = (s_1, \cdots, s_K)$ to encode the trial assignment where $(0, \cdots, s_k = 1, \cdots, 0)$ indicates the assignment to trial k, and a binary variable $t \in \{0, 1\}$ to encode treatment assignment, with $t = 1$ if assigned to the treatment group. For experimental unit i with trial assignment $\boldsymbol{s}$ and treatment assignment t, the random variable $M_i^{(t,s)}$ represents the potential mediator, and the random variable $Y_i^{(t,s,m)}$ represents the potential outcome that would be observed if i were to receive or exhibit level m of the mediator through some hypothetical mechanism.

We are interested in mediator DRF, which represents the value of $Y_i^{(t,s,m)}$ given different values (m) of $M_i^{(t,s)}$ for the same t and $\boldsymbol{s}$. This effectively is the red arrow in the Figure 2 for individuals. Let $\mu_i(m)$ be the mediator DRF for individual i. Our goal is to estimate average mediator DRF: $\mathbb{E}[\mu_i(m)]$, with which we can compute the percentage change of $\mathbb{E}[\mu_i(m)]$ for a 10% increase in m *ceteris paribus*. The expectation here is taken over the population of i and so are other expectations in this paper if not specified. Here we consider polynomial mediator DRF: $\mu_i(m) = \sum_{p=1}^{P} \beta_{i,p} m^p$, which can capture the nonlinearity of the causality. Estimating $\mathbb{E}[\mu_i(m)]$ means to estimate $\beta_p = \mathbb{E}[\beta_{i,p}]$ for $p = 1, 2, \cdots, P$. Vectors $\boldsymbol{\tau}_i$ and $\boldsymbol{\gamma}_i$ are in $\mathbb{R}^K$. Each element of the vector, $\tau_{i,k}$ or $\gamma_{i,k}$, represents direct effect of the treatment on mediator or outcome in trial k, and is assumed not to depend on m.

Vectors $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are also in $\mathbb{R}^K$, representing trial fixed effects. Random variables M_i^* and Y_i^* represent idiosyncratic individual characteristics in the values of potential mediator and potential outcome. Assume $\mathbb{E}[M_i^*] = \mathbb{E}[Y_i^*] = 0$ [2]. Let $\mathcal{M}$ be the support of the distribution of mediator, and $\mathcal{S}$ be the set containing all possible trial assignments.

We only observe realized data $\{M_i, Y_i, T_i, S_i\}$. The observed mediator $M_i := M_i^{(T_i, S_i)}$, and observed outcome $Y_i := Y_i^{(T_i, S_i, M_i)}$.

The specification has two implications. First, it implies that being in a particular trial will not affect mediator DRF. If mediator DRF

---

[1]Experiments with multiple treatment arms can be considered as multiple trials with one treatment group and one control group.
[2]This assumption can always be satisfied by reparameterizing M_i^*, Y_i^* and $\boldsymbol{\phi}$, $\boldsymbol{\theta}$.

Figure 3: Mediator DRF and Observed Data in Control Group

is not trial independent, then having many trials only adds noises rather than provides additional information, and thus defeats the purpose of conducting a meta-analysis.

Remark (Trial-Irrelevant Mediator DRF).

$$Y_i^{(t,s,m')} - Y_i^{(t,s,m)} = \mu_i(m') - \mu_i(m), \tag{3}$$

for all $t = 0$ or 1, $m', m \in \mathcal{M}$ and $\boldsymbol{s} \in \mathcal{S}$, where $\mu_i(m') - \mu_i(m)$ does not depend on s.

Second, the specification implies that there are no interaction effects between treatment and mediator on the outcome. It means the individual direct effect in each trial is irrelevant of the value of mediator. It is common in the literature of causal mediation (see, e.g., NPSEM-IE of Pearl [16] and FRCISTG of Robins [20]).

Remark (No-Interaction Pearl [16], Robins [20]).

$$Y_i^{(1,s,m)} - Y_i^{(0,s,m)} = \boldsymbol{\gamma}_i^\top \boldsymbol{s}, \tag{4}$$

for all $m \in \mathcal{M}$ and $\boldsymbol{s} \in \mathcal{S}$, where $\boldsymbol{\gamma}_i$ does not depend on m.

When mediator M_i is not independent of Y_i (which is the case here), it is not trivial to recover $\mathbb{E}[\mu_i(m)]$ through observed data. Figure 3 illustrates the challenge using simulated data. The grey lines in Figure 3 are simulated individual mediator DRFs following Definition 1, which represents the true causality between M_is and Y_is. The blue line is the average mediator DRF. After randomly assign individuals to a trial, we can compute their observed mediator values M_i and outcome values Y_i when in the control group, which are depicted by black scattered points. The black line shows the result from fitting the observed points by a widely-used non-parametric machine learning algorithm, locally estimated scatterplot smoothing (LOESS). Although the black line fits the data almost perfectly, it significantly deviates from the true underlying causality: the average mediator DRF.

4 IDENTIFICATION

4.1 Random Assignment

We first formally define the assumption that trial assignment and treatment assignment are random.

Assumption 1 (Random Assignment).

$$\left\{ Y_i^{(t',s,m)}, M_i^{(t,s)} \right\} \perp\!\!\!\perp S_i \tag{5}$$

$$\left\{ Y_i^{(t',s,m)}, M_i^{(t,s)} \right\} \perp\!\!\!\perp T_i | S_i = s \tag{6}$$

for $t, t' = 0$ or 1, all $m \in \mathcal{M}$ and $s \in \mathcal{S}$ and it is also assumed that $0 < \mathbb{P}(T_i = t | S_i = s) < 1$ and $0 < \mathbb{P}(M_i(t) = m | T_i = t, S_i = s) < 1$ for $t = 0$ or 1, all $m \in \mathcal{M}$ and $s \in \mathcal{S}$.

Equation 6 can be guaranteed by random assignment of treatments in online A/B tests. Equation 5 means that individual's potential outcomes and potential mediators are independent of the trial assignment. In practice, users are randomly selected into A/B tests, thus this assumption is trivially satisfied [3].

4.2 Relaxed Sequential Ignorability

Assumption 2 (Relaxed Sequential Ignorability).

$$Y_i^{(t,s,m')} - Y_i^{(t,s,m)} \perp\!\!\!\perp M_i | T_i = t, S_i = s \tag{7}$$

for all $m, m' \in \mathcal{M}$, all $s \in \mathcal{S}$, $t = 0, 1$.

With Definition 1 and assumption that $\mu_i(m)$ is a polynomial function, this assumption is equivalent to

$$\beta_{i,1}, \cdots, \beta_{i,P} \perp\!\!\!\perp M_i | T_i = t, S_i = s$$

for all $s \in \mathcal{S}$, $t = 0, 1$. This assumption says that the effect of changing m on the outcome of i is independent of the idiosyncratic individual unobservable that affects M_i. A similar assumption is proposed by Small [25, (IV-A3)]. It means that the underlying causality between online product and the business is invariant even we observe some users produce higher online metrics than others for unknown reasons (i.e., the idiosyncratic unobservable). It is weaker than the SI. Whenever SI is satisfied, this assumption is naturally satisfied.[4] But the inverse is not true. It is possible to break SI and still fulfill this assumption. For example, when the mediator DRF is the same for all individuals and potential outcomes only depends on unobserved confounders, SI could be violated, while this assumption is still satisfied.

4.3 Trial-level Conditional Independence

In order to allow the presence of many direct treatment effects, we put some structure on the direct treatment effects. Let H_k represents a vector of trial characteristics for trial k.

Assumption 3 (Trial-Level Conditional Independence). We assume $\gamma_{i,k} | H_k$'s are independently and identically distributed with $\mathbb{E}[\gamma_{i,k} | H_k] = H_k^\top \pi$ and $\gamma_{i,k} \perp\!\!\!\perp \tau_{i,k} | H_k$ for all i.

This assumption allows correlation between individual direct treatment effects on outcome and on mediator in a trial, while assuming such correlation disappears once conditioned on the characteristics of the trial. In the example of ranking algorithm, we may believe that experiments that test new algorithms generally have high impacts on both online NDCG and GMV whereas experiments that test new UI designs generally have only modest impacts on both metrics. But, within the same type of experiments, how much

---

[3]This is true regardless of whether the same unit can participate in multiple trials.
[4]If $Y_i^{(t,s,m)} \perp\!\!\!\perp M_i | T_i = t, S_i = s$ for all $m \in \mathcal{M}$ then naturally $Y_i^{(t,s,m')} - Y_i^{(t,s,m)}$ is also independent of M_i given $T_i = t, S_i = s$.

a treatment affects online NDCG does not correlate with its effect on GMV. This assumption is unverifiable. However, it is weaker than standard IV assumption in the literature. If there were no direct effects (IV assumption), then this assumption is trivially satisfied.

We can stack vectors of trial characteristics of all the K trials into a matrix $H = \begin{bmatrix} H_1 & H_2 & \ldots & H_K \end{bmatrix}^\top$. Assumption 3 implies

$$\mathbb{E}[\gamma_i | H] = \begin{bmatrix} \mathbb{E}[\gamma_{i,1} | H_1] & \mathbb{E}[\gamma_{i,2} | H_2] & \ldots & \mathbb{E}[\gamma_{i,K} | H_K] \end{bmatrix}^\top = H\pi \tag{8}$$

for all i.

4.4 Relevance Condition

Assumption 4 (Relevance Condition).

$$Var(\tau_k | H_k) > 0; \tag{9}$$

This assumption means, for the same type of experiments, direct treatment effect on M (τ_k) varies between experiments. It implies that treatment in each trial is still helpful for predicting M after conditioning on trial-level covariates. This assumption is similar to the standard rank condition of IV identification (see Wooldridge [28, Chapter 5]). Because τ_k can be calculated easily by summary statistics, this assumption can be empirically verified. In practice, since we can decide which trials to be included into the analysis, we can make sure this assumption always holds.

4.5 Identification of Mediator DRF

To simplify the proof, let's assume $\mu_i(m) = \beta_i m$. The same proof works for the more general polynomial DRF. Based on Definition 1, random variables of observed mediator M_i and observed outcome Y_i can be written as

$$M_i = (S_i T_i)^\top \tau + S_i^\top \phi + \eta_i, \tag{10}$$

$$Y_i = M_i \beta + S_i^\top \theta + (S_i T_i)^\top \gamma + \epsilon_i, \tag{11}$$

where

$$\eta_i = (S_i T_i)^\top (\tau_i - \tau) + M_i^*,$$
$$\epsilon_i = (S_i T_i)^\top (\gamma_i - \gamma) + M_i(\beta_i - \beta) + Y_i^*,$$
$$\beta = \mathbb{E}[\beta_i], \tau = \mathbb{E}[\tau_i], \text{and } \gamma = \mathbb{E}[\gamma_i].$$

By plugging Equation 10 into Equation 11 for M_i, we obtain

$$Y_i = (S_i T_i)^\top \tau \beta + S_i^\top (\theta + \phi \beta) + (S_i T_i)^\top H\pi + \epsilon_i' \tag{12}$$

where $\epsilon_i' = (S_i T_i)^\top (\gamma_i - H\pi) + \eta_i \beta + M_i(\beta_i - \beta) + Y_i^*$.

With all the specifications and assumptions, we are ready to present the important result about the identification.

Theorem 1. *Consider the model specified in Definition 1, Under Assumption 1 - 4, the average mediator effect $\beta = \mathbb{E}[\beta_i]$ can be identified. A consistent estimator of β can be derived through a procedure of two-stage least squares (2SLS):*

(1) Following Equation 10, regress M_i on $S_i T_i$ and S_i via least squares to obtain a consistent estimator $\widehat{\tau}$;

(2) Plug the consistent estimator $\widehat{\tau}$ into Equation 12, and regress Y_i on $(S_i T_i)^\top \widehat{\tau}$, S_i, and $(S_i T_i)^\top H$ via least squares to obtain the consistent estimator $\widehat{\beta}$.

The proof is in Appendix A. Since τ can be identified from 10, we can use $\widehat{\tau}$ in place of τ. The proof shows that under Assumption 1 - 4, the covariances between ϵ'_i and all the covariates, $S_i T_i \widehat{\tau}$, S_i and $S_i T_i^{\top} H$ in Equation 12 are zeros. Therefore, structural parameters β can be identified (see, e.g., Wooldridge [28, Chapter 4] for more details of coefficient identification in linear regression).

The estimator $\widehat{\beta}$ from 2SLS of Theorem 1 is equivalent to an IV-2SLS estimator (see, e.g., Wooldridge [28, Chapter 5] for more details of IV-2SLS). To see it, let's rewrite Equation 11 as

$$Y_i = M_i \beta + S_i^{\top}(\theta + \phi\beta) + (S_i T_i)^{\top} H\pi + \epsilon''_i \qquad (13)$$

where $\epsilon''_i = (S_i T_i)^{\top}(\gamma_i - H\pi) + M_i(\beta_i - \beta) + Y_i^*$. We could get the same estimator of β as in Theorem 1 through applying IV-2SLS on Equation 13 and using $S_i T_i$ as instruments for M_i.

5 ESTIMATION AND HYPOTHESIS TESTING

5.1 Estimation

Theorem 1 implies that average mediator effect can be estimated by running two regressions with pooled data. Such estimation could be very costly when the sample size in each trial is huge so that pooling data from all trials becomes infeasible. We propose a simpler two-stage procedure: CMMA.

Algorithm 1 CMMA

Input Y_i, T_i, and M_i, S_i, H_k
Output $\widehat{\beta}$
 1: **for** trial $k = 1$ to K **do**
 2:     Estimate ATE of trial k treatment on Y using data from trial k and denote it as ATE_k^Y.
 3:     Estimate ATE of trial k treatment on M using data from trial k and denote it as ATE_k^M.
 4: **end for**
 5: Regress ATE_k^Y on ATE_k^M and additional trial-level covariates H_k.
 6: Save the coefficient for ATE_k^M as $\widehat{\beta}$.
 7: **return** $\widehat{\beta}$

Step 1-4 of CMMA calculates ATEs for each trial. ATE on Y and M could be done in the standard procedure of online A/B tests. Step 5 and 6 uses only trial-level summary statistics and covariates, making it very easy to implement. This estimator has the same identification strategy as in Theorem 1 and is equivalent to a weight-adjusted 2SLS estimator. The proof is in Appendix B.

Note that, CMMA allows the same unit to participate in multiple trials. We can always use regression/ANOVA with treatment interaction terms to estimate ATEs of each trial for units in multiple trials, and then implement Step 5 and 6 of CMMA on estimated ATEs to get $\widehat{\beta}$.

The most challenging part of applying CMMA is finding valid trial-level characteristics H_k to satisfy Assumption 3. A good H_k should have explanatory power for treatment effects on outcome and mediator across trials. However, similar to finding a valid instrument, there is no systematic way to produce H_k. Practitioners have to rely on available data and domain knowledge to argue for the validity of H_k. In Section 6 Table 2, we simulate the consequences of violating Assumption 3. In Section 7.2 Figure 4, we discuss the choice of H_k in

our real data application. Future work is required on the sensitivity of CMMA to H.

5.2 Hypothesis Testing

In general, the reported standard errors from the second stage regression is slightly different from theoretical values without access to residuals in the first stage. But this becomes less of an issue as sample size increases. Since the sample size is usually enormous in online A/B tests, we recommend using the reported standard errors in the second stage regression for convenience.

Although we have assumed that $\mu_i(m) = \beta_i m$ for discussion to this point, the same proof is still valid for polynomial DRF, $\mu_i(m) = \sum_{p=1}^{P} \beta_{i,p} m^p$. Let $\overline{M_i} = [M_i, \cdots, M_i^p, \cdots, M_i^P]^{\top}$ and $\overline{\beta} = [\beta_1, \cdots, \beta_p, \cdots, \beta_P]^{\top}$, where $\beta_p = \mathbb{E}[\beta_{i,p}]$. Then we can use $\overline{M_i}^{\top}\overline{\beta}$ in place of $M_i\beta$ in the proof. For our algorithm, in addition to ATE_k^Y and ATE_k^M, we also need to estimate ATE on the higher order terms of M_i, $\text{ATE}_k^{(M^p)}$, $p = 2, 3, \cdots, P$. [5]

To decide the highest p-th order term to include in the model of the second stage, we can use the common model selection tool: Wald test. The standard way is to run a regression with higher-order terms and then perform a series of tests to check whether coefficients of those terms are zeros. See Greene [8, Chapter 5] for more technical discussions on Wald test.

6 SIMULATION

We conduct Monte Carlo simulations to study the finite sample performance of our estimator. The details of our simulation set-up are described in Appendix C. The R code is available in our GitHub repository: https://github.com/znwang25/cmma.

Table 1: Finite-Sample Performance Comparison

| Estimator | $N_{per} = 200$ | | $N_{per} = 500$ | | $N_{per} = 1000$ | |
	Bias	95% CI coverage	Bias	95% CI coverage	Bias	95% CI coverage
LIML	0.003	93%	-0.002	94%	0.000	94%
CMMA	0.058	67%	0.020	88%	0.011	88%
Full Sample 2SLS	0.058	66%	0.020	83%	0.011	86%
Sobel [26]	0.308	3%	0.279	3%	0.283	3%
LSEM [3]	0.937	0%	0.936	0%	0.937	0%

We use Limited Information Maximum Likelihood (LIML) estimator as a benchmark to evaluate our CMMA estimator. An LIML estimator that is specified according to the simulation setup should have the best performance theoretically. We include two common estimators in the literature of mediation analysis into the comparison: Sobel [26] and LSEM [3], which are derived under identification approaches discussed in Section 2. Sobel [26] assumes complete mediation, and LSEM [3] relies on SI assumption [12]. Both assumptions are false under our setting. We also implement the Full Sample 2SLS estimator prescribed in Theorem 1, which should produce similar results to CMMA.

---

[5]Note: We need $\text{ATE}_k^{(M^p)}$, not the p-th order of ATE_k^M: $(\text{ATE}_k^M)^p$. $\text{ATE}_k^{(M^p)} \neq (\text{ATE}_k^M)^p$.

Table 2: Assumption Violation

| Violation | $N_{per} = 200$ | | $N_{per} = 500$ | | $N_{per} = 1000$ | |
	Bias	95% CI coverage	Bias	95% CI coverage	Bias	95% CI coverage
None	0.058	67%	0.020	88%	0.011	88%
A2	0.065	65%	0.024	87%	0.009	93%
A3	0.483	0%	0.470	0%	0.460	0%
A4	0.949	0%	0.931	0%	0.942	0%

We set the sample size per trial (N_{per}), to 200, 500, and 1000 and perform 100 simulations for each setting. Table 1 reports average biases and 95% confidence interval coverage of the estimators. The performance of CMMA is quite good and largely comparable to LIML's result. When the sample size per trial is small, our estimator is slightly biased but the bias is much smaller than those of Sobel and LSEM estimators. As N_{per} increases, the bias of our estimator decreases to zero, whereas the biases of the Sobel and LSEM estimators remain roughly the same. As the sample size is usually enormous in A/B tests (on average, one trial has millions of observations in the real data we obtained from an internet company), the bias of CMMA will be negligible in practice. Table 1 also shows that, as N_{per} increases, the 95% confidence interval coverage of our estimator converges to the nominal coverage. This means that the OLS variance estimated in the trial-level regression is valid for hypothesis testing when N_{per} is sufficiently large. In addition, the point estimate of Full Sample 2SLS estimator is numerically equal to CMMA, which empirically validates the equivalence claim made in Section 5.1.

In Table 2, we examine the performance of CMMA when each assumption fails. The results show that, the failure of Assumption 2 does not seem to affect the estimator's unbiasedness, whereas with failures of Assumption 3 or 4, CMMA is no longer unbiased. Comparing results across rows, it seems to suggest that violating Assumption 4 has worse consequence in terms of bias. Fortunately, Assumption 4 is testable as discussed in Section 4.

We also test the performance of the Wald test with a higher order of $\mu(m)$. Table 3 shows that Wald tests can successfully select the correct model. For example, in the second row of Table 3, the true mediator DRF is a quadratic function of m: $4m + 2m^2$. Wald tests in all simulations reject the null hypothesis that coefficients for the second-degree term and the third-degree term are zeros, whereas 97% of simulations fail to reject the hypothesis that coefficient for the third-degree term is zero. The result suggests that the highest order of $\text{ATE}_k^{(M^p)}$ should be 2. The last column of Table 4 shows that, with a correctly specified model, we can accurately estimate all the underlying parameters.

7 APPLICATION

We apply the approach on three most popular rank-aware evaluation metrics: NDCG, MRR, and MAP, to show, for ranking algorithms that power search page of Etsy.com, which one could lead to the most significant lift of sitewide GMV. Since the offline A/B test literature [6] bridges the inconsistency between changes of offline and online evaluation metrics, we only focus on, how sitewide GMV

would change for 10% lifts in online NDCG, MRR, and MAP of search page respectively. All metrics in the application, unless otherwise noted, are *online* metrics. Please note the approach has not been deployed in Etsy. This work is not intended to apply to, nor is it a prediction of, actual live performance metrics or performance changes on Etsy or any other property.

7.1 User-Level Evaluation Metrics

We follow the offline A/B test literature [6] and define the three online rank-aware evaluation metrics at the user level. Although the three metrics are originally defined at the query level in the test collection evaluation of information retrieval (IR) literature, the search page in the industry is an online product for users and thus the computation could be adapted to the user level. More specifically, the three metrics are constructed as follows: 1) query-level metrics are computed using rank positions on search page and user conversion status as binary relevance, and non-conversion associated queries have zero values[6]; 2) user-level metrics is the average of query-level metrics across all queries the user issues (including non-conversion associated queries), and users who do not search or convert have zero values. Also, all the three metrics are defined at rank position 48, the lowest position of the first page of search results in Etsy.com.

7.2 Data

We have access to summarized results of 190 randomly selected experiments from the online A/B test platform of Etsy.com. All the experiments in the data have the user as an experimental unit. The data include descriptive information about each experiment such as the tested product change, the product team that initiated the experiment, and summary statistics of each experiment such as average (user-level) NDCG per user in treatment and control groups. Note that, the difference between the average metric per user in treatment and control groups is ATE on the metric (Step 1 - 4 in Algorithm 1).

We use the taxonomy of product teams as our trial-level covariates H because team taxonomy is quite informative of experiments as Figure 4 suggests. Figure 4 shows the density of ATE on NDCG for selected teams including UI/UX design, marketing, search ranking. First, it is evident that distributions of ATE on NDCG vary by team, which implies that Assumption 3 is likely to be satisfied. In particular, most experiments of the search ranking team post positive gains on NDCG, whereas experiments of the UI/UX team barely affects NDCG. Second, within each team there are significant variances of ATE on NDCG, which suggests that Assumption 4 holds. Due to page limitation, only results for NDCG are presented here, but the distributions of ATE on MAP and MRR exhibit the similar pattern.

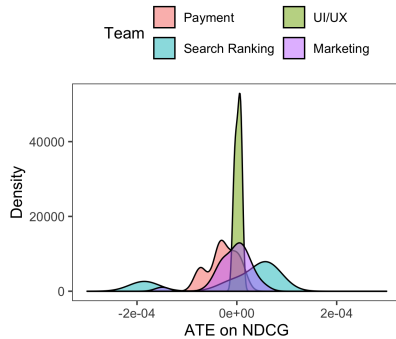### 7.3 Results

To decide the polynomial terms in the model, we perform Wald tests. The result from Wald tests is in Table 4. Since all the null hypothesis are rejected, the results suggest us including ATE on M, ATE on M^2 and ATE on M^3 in the model of each M (NDCG, MAP, MRR).

---

[6]If the user purchases the item that she has clicked on the search page, the relevance is 1; otherwise 0.

Table 3: Model Selection and Wald Tests

$\mu(m)$: $\beta_1 m + \beta_2 m^2 + \beta_3 m^3$	% of Wald Tests Rejecting H_0			CMMA Estimation	
	$H_0 : \beta_3 = 0$	$H_0 : \beta_2 = \beta_3 = 0$	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$	Highest Order of $\text{ATE}_k^{(MP)}$	Estimated β's of $\mu(m)$
$4m$	3%	5%	100%	1	4.016
$4m + 2m^2$	3%	100%	100%	2	4.019, 1.999
$4m + 0m^2 + 5m^3$	100%	100%	100%	3	4.022, -0.001, 5

Note: The sample size per trial is 1000 and the number of trials is 100. H_0 is rejected if p-value < 0.05.

Figure 4: Density of ATE on NDCG

Table 4: Wald Test Results

Null Hypothesis	P-Value		
	M: NDCG	M: MAP	M: MRR
$H_0 : \beta_3 = 0$	0.00119	0.02480	0.00346
$H_0 : \beta_2 = \beta_3 = 0$	0.00018	0.00003	0.00000
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$	0.00018	0.00003	0.00000

Figure 5 shows the estimation results. On the first row, blue lines depict estimated mediator DRFs. Scattered points represent summary statistics (the data for CMMA), ATE on ranking evaluation metrics and ATE on GMV, of all experiments. Black curves show results from fitting the data by LOESS. Note that, the range of three evaluation metrics could be much smaller than those in IR literature since they are defined at the user level (see Section 7.1). The estimated coefficients of mediator DRF are in Table D2 in the Appendix. The estimated mediator DRFs show that all three online metrics have positive causal effects on GMV. Note that, the causal relationships are different from the pattern of the data (scatter points and black curves). The differences between blue lines and black curves show the bias from fitting the data by machine learning methods without addressing omitted variables.

The second row of Figure 5 shows the *elasticity* of GMV: the percentage change of GMV for a 10% increase in each evaluation metric at its different values, which are derived from estimated mediator DRF. The downward slopes imply that, for all three evaluation metrics, as they increase, the benefit of continuous improving them on GMV decreases. For example, when average NDCG per user equals 0.0021, its 10% increase leads to a 9.88% increase in average GMV per

user. Yet, when it equals 0.006, its 10% increase only leads to a 9.65% increase in average GMV per user.

Now it is easy for product owners to pick the evaluation metric that could guide algorithm development to achieve the most significant lift in online GMV. Suppose the current average values (per user) of NDCG, MAP, and MRR from live data are 0.00210, 0.00156, and 0.00153 respectively. From estimated mediator DRFs, we can calculate their corresponding elasticities of GMV: 9.88%, 9.87%, and 9.90%, which are marked by red lines in Figure 5. Because online MRR has the highest elasticity of GMV, we should choose offline MRR, which is estimated based on offline A/B test literature [6] and thus has the same move as online MRR, to guide the development of ranking algorithms.

8 CONCLUSION

In the internet industry, the algorithms developed offline power online products and online products contribute to the success of a business. In many cases, offline evaluation metrics, which guide algorithm development, are different from online business KPIs. It is important for product owners to pick the offline evaluation metric guided by which the algorithm could maximize online business KPIs. By noticing that online products could be assessed by online counterparts of offline evaluation metrics, we decompose the problem into two parts. Since the offline A/B test literature works out the first part: counterfactual estimators of offline evaluation metrics that move the same way as their online counterparts, we focus on the second part: inferring causal effects of online evaluation metrics on business KPIs. The offline evaluation metric whose online counterpart causes the most significant lift in online business KPIs should be the north star. We model online evaluation metrics as mediators and formalize the problem as to identify, estimate, and test mediator DRF. Our novel approach CMMA combines mediation analysis and meta-analysis and has many advantages over the two strands of literature. In particular, it takes as inputs only summary statistics from multiple past A/B tests, and thus it is easy to implement in scale. We apply the approach on Etsy's real data to uncover the causality between three most popular rank-aware online evaluation metrics and GMV, and show how we successfully identify MRR as the offline evaluation metric for GMV maximization.

REFERENCES

[1] Joshua Angrist, Guido Imbens, and Donald Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *J. Amer. Statist. Assoc.* 91, 434 (6 1996), 444.
[2] Joshua Angrist and Alan Krueger. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15, 4 (11 2001), 69–85.
[3] Reuben Baron and David Kenny. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical
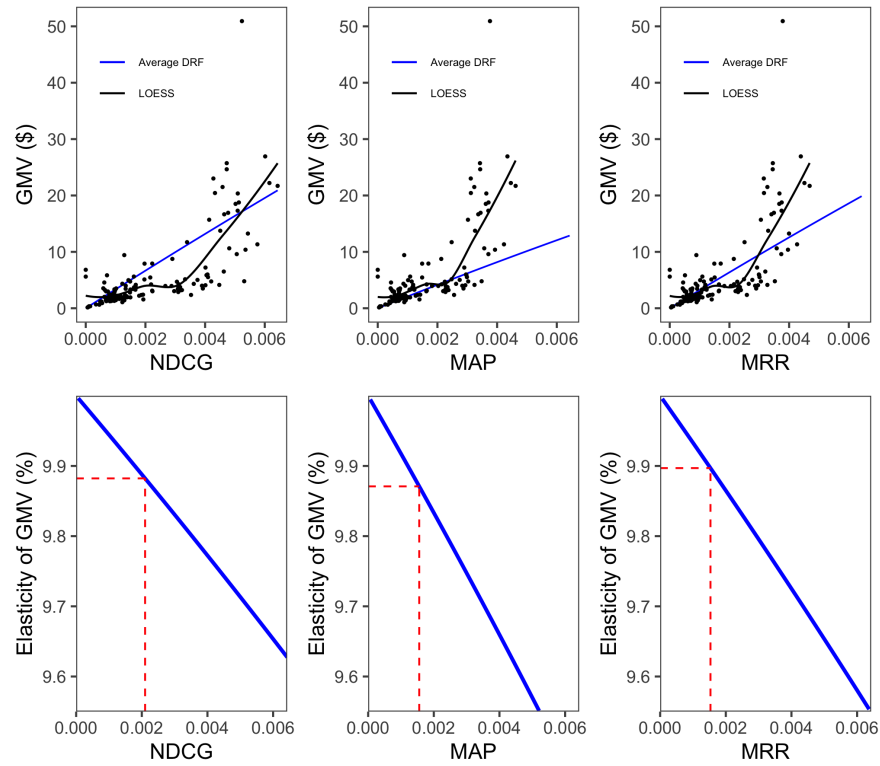
Figure 5: Estimated Mediator Dose Response Function

considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173–1182.

[4] Will Browne and Mike Jones. 2017. What works in e-commerce - a meta-analysis of 6700 online experiments. *Qubit Digital Ltd* (2017), 1–21.

[5] Harris Cooper, Larry Hedges, and Jeffrey Valentine. 2009. *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

[6] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18).* Association for Computing Machinery, New York, NY, USA, 198–206.

[7] Donald Green, Shang Ha, and John Bullock. 2010. Enough already about "Black Box" experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Science* 628, 1 (2010), 200–208.

[8] William Greene. 2011. *Econometric analysis* (7 ed.). Pearson Education Inc. 1232 pages.

[9] James Heckman and Rodrigo Pinto. 2015. Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs. *Econometric Reviews* 34 (2015), 6–31.

[10] Julian Higgins and Simon Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21, 11 (6 2002), 1539–1558.

[11] Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2014. Comment on Pearl: Practical implications of theoretical results for causal mediation analysis. *Psychological Methods* 19, 4 (2014), 482–487.

[12] Kosuke Imai, Luke Keele, and Teppei Yamamoto. 2010. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statist. Sci.* (2010).

[13] Guido Imbens. 2000. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika* 87, 3 (2000), 706–710.

[14] Guido Imbens and Keisuke Hirano. 2004. The Propensity Score with Continuous Treatments. (2004).

[15] David MacKinnon, Amanda Fairchild, and Matthew Fritz. 2006. Mediation Analysis. *Annual Review of Psychology* 58, 1 (12 2006), 593–614.

[16] Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers

Inc., 411–420.

[17] Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological Methods* 19, 4 (2014), 459–481.

[18] Judea Pearl. 2014. Reply to Commentary by Imai, Keele, Tingley, and Yamamoto Concerning Causal Mediation Analysis. *Psychological Methods* 19, 4 (2014), 488–492.

[19] Alexander Peysakhovich and Dean Eckles. 2018. Learning causal effects from many randomized experiments using regularized instrumental variables. In *The Web Conference 2018 (WWW 2018).* ACM, New York, NY.

[20] James Robins. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems* (1 2003), 70–82.

[21] James Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3, 2 (1992), 143–155.

[22] James Robins and Thomas Richardson. 2010. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: finding the determinants of disorders and their cures* (2010).

[23] Donald Rubin. 2003. Basic concepts of statistical inference for causal effects in experiments and observational studies. (2003).

[24] Derek Rucker, Kristopher Preacher, Zakary Tormala, and Richard Petty. 2011. Mediation Analysis in Social Psychology: Current Practices and New Recommendations. *Social and Personality Psychology Compass* 5, 6 (2011), 359–371.

[25] Dylan Small. 2012. Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *Journal of Statistical Research* 46, 2 (2012), 91–103.

[26] Michael Sobel. 2008. Identification of Causal Parameters in Randomized Studies With Mediating Variables. *Journal of Educational and Behavioral Statistics* 33, 2 (2008), 230–251.

[27] Tom Stanley and Hristos Doucouliagos. 2012. *Meta-regression analysis in economics and business.* Routledge.

[28] Jeffrey Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press, Cambridge, MA. 1096 pages.

[29] Xuan Yin and Liangjie Hong. 2019. The Identification and Estimation of Direct and Indirect Effects in A/B Tests Through Causal Mediation Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19).* ACM, New York, NY, USA, 2989–2999.

A PROOF OF THEOREM 1

PROOF. Let $X_i = \begin{bmatrix} (S_i T_i)^\top \tau & S_i^\top & (S_i T_i)^\top H \end{bmatrix}$, and
$B = \begin{bmatrix} \beta & \theta + \phi\beta & \pi \end{bmatrix}^\top$, so Equation 12 can be written as

$$Y_i = X_i B + \epsilon_i'$$

Because of assumption 4, $\mathbb{E}[X_i^\top X_i]$ is guaranteed to have full
rank, thus is invertible. If we are able to show that $\mathbb{E}[X_i^\top \epsilon_i'] = \mathbf{0}$,
then B can be estimated by linear projection $B^* = \mathbb{E}[X_i^\top X_i]^{-1}\mathbb{E}[X_i^\top Y_i]$.
While this is infeasible, we can first estimate τ with its estimate
$\widehat{\tau}$ and use $\widehat{X_i} = \begin{bmatrix} (S_i T_i)^\top \widehat{\tau} & S_i^\top & (S_i T_i)^\top H \end{bmatrix}$ in place of X_i. Let
$\widehat{X}$ be N-component data vector with ith element $\widehat{X_i}$. The resulted
estimator is the two-stage least squares(2SLS) estimator,

$$\widehat{B}_{2SLS} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top Y$$

to estimate B.

Note that, because of Assumption 1, 2 and 3, the following equations are true.

$\mathbb{E}[\eta_i] = \mathbb{E}[S_i T_i]^\top \mathbb{E}[(\tau_i - \tau)] + \mathbb{E}[M_i^*] = 0$

$\mathbb{E}[M_i(\beta_i - \beta)] = \mathbb{E}[\mathbb{E}[M_i(\beta_i - \beta)|S_i, T_i]] = \mathbb{E}[M_i \mathbb{E}[(\beta_i - \beta)|S_i, T_i]] = 0$

$\mathbb{E}[\gamma_i - H\pi] = \mathbb{E}[\mathbb{E}[\gamma_i|H] - H\pi] = \mathbf{0}$

With the first component in X_i,

$\mathbb{E}\left[(S_i T_i)^\top \tau \epsilon_i'\right]$

$= \mathbb{E}\left[(S_i T_i)^\top \tau \left[(S_i T_i)^\top(\gamma_i - H\pi) + \eta_i \beta + M_i(\beta_i - \beta) + Y_i^*\right]\right]$

$= \mathbb{E}\left[(S_i T_i)^\top \tau (S_i T_i)^\top(\gamma_i - H\pi)\right]$
$\quad + \mathbb{E}\left[(S_i T_i)^\top \tau\right] \mathbb{E}\left[\eta_i \beta + M_i(\beta_i - \beta) + Y_i^*\right]$

$= \mathbb{E}\left[(S_i T_i)^\top \tau (S_i T_i)^\top(\gamma_i - H\pi)\right]$

$= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[S_{i,k} T_i \tau_{i,k}(\gamma_{i,k} - H_k \pi)\right]\right]$

$= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{E}\left[S_{i,k} T_i \tau_{i,k}(\gamma_{i,k} - H_k \pi)|H_k\right]\right]\right]$

$= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[S_{i,k} T_i \tau_{i,k} \left(\mathbb{E}\left[\gamma_{i,k}|H_k\right] - H_k \pi\right)\right]\right]$

$= 0,$

where the forth equality expands the matrix multiplication and the
sixth equality follows from Assumption 3.

For the second component,

$\mathbb{E}\left[S_i \epsilon_i'\right]$

$= \mathbb{E}\left[S_i (S_i T_i)^\top(\gamma_i - H\pi)\right]$
$\quad + \mathbb{E}[S_i] \mathbb{E}\left[\eta_i \beta + M_i(\beta_i - \beta) + Y_i^*\right]$

$= \mathbb{E}\left[S_i (S_i T_i)^\top\right] \mathbb{E}\left[\gamma_i - H\pi\right]$

$= 0,$

where both the second and third equality follows from Assumption
1.

With respect to the third component,

$\mathbb{E}\left[(S_i T_i)^\top H \epsilon_i'\right]$

$= \mathbb{E}\left[(S_i T_i)^\top H (S_i T_i)^\top(\gamma_i - H\pi)\right]$
$\quad + \mathbb{E}\left[(S_i T_i)^\top H\right] \mathbb{E}\left[\eta_i \beta + M_i(\beta_i - \beta) + Y_i^*\right]$

$= \mathbb{E}\left[(S_i T_i)^\top H (S_i T_i)^\top \mathbb{E}\left[(\gamma_i - H\pi)|H\right]\right]$

$= 0,$

where both the second and third equality follow from Assumption
1.

Taken together, we have shown that $\mathbb{E}[X_i^\top \epsilon_i'] = \mathbf{0}$, therefore B
can be estimated by

$$\widehat{B}_{2SLS} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top Y.$$

□

B EQUIVALENCE OF CMMA AND WEIGHT-ADJUSTED 2SLS ESTIMATOR

Let $\mathbf{Z}$ be an N-component data vector with ith element $(S_i T_i)^\top$,
$\mathbf{S}$ be an N-component data vector with ith element $S_i^\top$, and $\mathbf{P_S} = \mathbf{S}(\mathbf{S}^\top \mathbf{S})^{-1}\mathbf{S}^\top$, $\mathbf{M_S} = \mathbf{I} - \mathbf{P_S}$. The following proposition shows the link
between CMMA and 2SLS estimator.

Proposition 1. *The* CMMA *method defined in Algorithm 1 is equivalent to a weight-adjusted 2SLS IV estimator with weight* $\mathbf{W}^\top \mathbf{W}$*, where*
$\mathbf{W} = (\mathbf{Z}^\top \mathbf{M_S Z})^{-1}\mathbf{Z}^\top$

PROOF. Let $\beta = \begin{pmatrix} \beta \\ \pi \end{pmatrix}, \overline{\mathbf{Z}} = \begin{bmatrix} \mathbf{Z}\widehat{\tau} & \mathbf{Z}H \end{bmatrix}$. Since we are not particu-
larly interested in coefficients in front of S_i, using Frisch–Waugh–Lovell
theorem, we can get 2SLS estimator for β,

$$\widehat{\beta}_{2SLS} = \left(\overline{\mathbf{Z}}^\top \mathbf{M_S M_S} \overline{\mathbf{Z}}\right)^{-1} \overline{\mathbf{Z}}^\top \mathbf{M_S M_S} Y.$$

We could use a weighting matrix $\mathbf{W}$ and still get a consistent esti-
mator for β,

$$\tilde{\beta} = (\overline{\mathbf{Z}}^\top \mathbf{M_S W}^\top \mathbf{W M_S} \overline{\mathbf{Z}})^{-1} \overline{\mathbf{Z}}^\top \mathbf{M_S W}^\top \mathbf{W M_S} Y$$

.

Use $\mathbf{W} = (\mathbf{Z}^\top \mathbf{M_S Z})^{-1}\mathbf{Z}^\top$

$$\overline{\mathbf{Z}}^\top \mathbf{M_S W}^\top = \begin{pmatrix} \widehat{\tau}^\top \mathbf{Z}^\top \mathbf{M_S Z}(\mathbf{Z}^\top \mathbf{M_S Z})^{-1} \\ H^\top \mathbf{Z}^\top \mathbf{M_S Z}(\mathbf{Z}^\top \mathbf{M_S Z})^{-1} \end{pmatrix} = \begin{pmatrix} \widehat{\tau}^\top \\ H^\top \end{pmatrix}$$

$$\widehat{\beta}_{CMMA} = (\begin{pmatrix} \widehat{\tau}^\top \\ H^\top \end{pmatrix} \begin{pmatrix} \widehat{\tau} & H \end{pmatrix})^{-1} \begin{pmatrix} \widehat{\tau}^\top \\ H^\top \end{pmatrix} \mathbf{W M_S} Y$$

$\mathbf{W M_S} Y = (\mathbf{Z}^\top \mathbf{M_S Z})^{-1}\mathbf{Z}^\top \mathbf{M_S} Y$ is equivalent to regressing Y_i on
$S_i T_i$ and S_i and take coefficients of $S_i T_i$. And it is equivalent to
regress on T_i for each trial.   □

C SIMULATION SETUP

We follow the specification described in Equation 10 and Equation
12 and let the M_i^* and Y_i^* be jointly normally distributed:

$$\begin{pmatrix} M_i^* \\ Y_i^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_{M^*}^2 & , & \rho\sigma_{M^*}\sigma_{Y^*} \\ \rho\sigma_{M^*}\sigma_{Y^*} & , & \sigma_{Y^*}^2 \end{bmatrix}\right)$$

with $(\rho, \sigma_{M^*}, \sigma_{Y^*}) = (0.95, 3, 3)$. We fix the number of trials to be 50,
and used independent uniform distributions to specify parameter
value for each element of θ, ϕ. To satisfy Assumption 3 and 4, τ
is set to be the sum of a H dependent term and a random vector
drawing from a uniform distribution. All the other parameter values
are listed in the Table C1. The innovations in the error terms such

as $\beta_i - \beta$, elements of $\tau_i - \tau$, and $\gamma_i - H\pi$ are all drawn from independent normal distributions $\mathcal{N}(0, 0.5)$.

Table C1: Parameter Values

Parameters	Value
β	4
π	$(0, 1.5, 3)^\top$
θ	$(\theta_1, \cdots, \theta_{50})^\top, \theta_k \sim \mathcal{U}(-2, 2) \; \forall k$
ϕ	$(\phi_1, \cdots, \phi_{50})^\top, \phi_k \sim \mathcal{U}(-2, 2) \; \forall k$
τ	$\mathbf{T}^\top \mathbf{H}^\top + (\tau_1, \cdots, \tau_{50})^\top,$ $\tau_k \sim \mathcal{U}(-3, 3) \; \forall k$ and $\mathbf{T} = (0.5, 1, 2.5)^\top$

The one-hot group assignment variable S_i and treatment indicator T_i are randomly generated. We assume the 50 trials can be grouped into 3 experiment types and use experiment types as trial level covariates. Trials are randomly assigned into three types and H is a 3×50 matrix representing such assignment. Under this setup, all the assumptions are satisfied, and thus we are ready to estimate β.

D TABLES

Table D2: Second Stage Regression Results

	Dependent variable: GMS		
	M: NDCG	M: MPP	M: MRR
	(1)	(2)	(3)
ATE^M	3,369.9*	2,113.3	3,227.6
	(1,753.3)	(2,600.2)	(2,368.2)
ATE^{M^2}	−18,593.8***	−17,254.4*	−21,411.1**
	(6,606.5)	(9,718.6)	(8,882.6)
ATE^{M^3}	16,733.0***	16,191.1**	19,229.7***
	(5,162.6)	(7,213.5)	(6,576.5)
Observations	190	169	190
R^2	0.240	0.271	0.282
Note:		*p<0.1; **p<0.05; ***p<0.01	