

From the research bench

What is propensity score matching?

Propensity score matching compares treated units with untreated ones that had a similar estimated chance of treatment. It balances only what was measured.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic)	18 October 2026	20 min	4,352
Commercial growth leader, lecturer and doctoral researcher			

Management summary
Propensity score matching estimates the effect of a treatment on the units that received it by pairing each with untreated units that had nearly the same estimated probability of being treated, given covariates measured beforehand. This concept guide defines the propensity score and explains why one score can stand in for many covariates (Rosenbaum and Rubin), sets out the two assumptions a causal reading needs, no unmeasured confounders and overlap, compares matching, stratification and weighting on the score, shows how balance is checked, and reports how matching and related methods fared against 15 randomized advertising experiments at Facebook (Gordon et al.). It gives a one-page record for a matched estimate. The sources are one statistics paper and one comparison study on one platform.

Contents	§ 1	What is propensity score matching?	3
	§ 2	Why can one score stand in for many covariates?	4
	§ 3	What must hold before a matched comparison counts as causal?	5
	§ 4	How do matching, stratification and weighting on the score differ?	6
	§ 5	How is balance checked after matching?	7
	§ 6	What happened when matching was held against Facebook's experiments?	8
	Figure 1	Share of the raw comparison's bias left after stratifying on the propensity score, 14 Facebook checkout studies	11
	§ 7	What should travel with a matched estimate?	12
	Figure 2	A matched-estimate record	13
	§ 8	Which parts of the Facebook result carry over to other data?	14

References

16

What is propensity score matching?

§ 1 of 8

It is a way to build a comparison group from data in which nobody assigned the treatment by chance. Rosenbaum and Rubin (1983)² define the score it rests on: “The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates.” Matching on the score pairs each treated unit with untreated units that had about the same estimated probability of being treated, and compares their outcomes. Matching in general, in their words, “is a method of sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group.” The covariates come from before the treatment: in their setup “all of the measurements in x are made prior to treatment assignment”, though x need not include every covariate that was used to assign the treatment.

² Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Gordon et al. (2019)¹, who test the method on advertising data, describe what the comparison estimates: “To estimate the ATT, matching methods find untreated individuals similar to the treated individuals and use the outcomes from the untreated individuals to impute the missing potential outcomes for the treated individuals.” The ATT is the average treatment effect on the treated: the effect on the units that received the treatment, not on everyone who might have. One appeal they name is that matching methods “do not assume a particular form for an outcome model”.

¹ Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2), 193–225.

In this guide, propensity score matching means estimating the effect of a treatment on the units that received it by pairing each with one or more untreated units of nearly the same estimated propensity score, computed from covariates measured before the treatment, and comparing their outcomes. The result is a causal estimate only under the two assumptions set out below, and only for the treated units that found comparable partners.

The incrementality guide uses the same Facebook study to show how far a raw comparison of exposed and unexposed users can overstate what advertising added. This page takes the method itself: what the

score does, what it needs, how its variants differ, and how well it did against randomized experiments.

Why can one score stand in for many covariates?

§ 2 of 8

Because, at any given value of the score, treated and untreated units have the same distribution of the covariates that went into it. Rosenbaum and Rubin (1983)² call a function of the covariates a balancing score when, at each of its values, treated and control units have the same covariate distribution: “Balancing scores, defined here, can be used to group treated and control units so that direct comparisons are more meaningful.” The full set of covariates “is the finest balancing score and the propensity score is the coarsest”: one number does the work of many. Hence, “if a subclass of units or a matched treatment-control pair is homogeneous in $e(x)$, then the treated and control units in that subclass or matched pair will have the same distribution of x ”, where $e(x)$ is the score and x the covariates. Gordon et al. (2019)¹ give the practical reason this matters: “Exact matching is only feasible using a small set of discrete observables.”

Balanced covariates are not yet a causal estimate. That step needs the assumption in the next section; if it holds, “units with the same value of the balancing score $b(x)$ but different treatments can act as controls for each other”, and “pair matching, subclassification and covariance adjustment on a balancing score can produce unbiased estimates of the average treatment effect”, where the average runs over the whole population; matching each treated unit, as Gordon et al.¹ do, averages over the treated instead. Outside a randomized trial, “the propensity score function is almost always unknown” and is estimated from the data, for example with a logit model.

Estimating it costs little in the authors’ simulation, which they take from earlier work by Rubin. With 50 treated units, bivariate normal covariates with a common covariance matrix and straight-line relations between covariates and outcome, parallel in both groups, matching on the estimated score, which under those conditions is the same as matching on a linear discriminant, removed between 67 and 92 per cent of the initial bias, depending on how far apart the groups started, when the pool of potential controls was twice the size of the treated group, and between 90 and 98 per cent when it was four times the size (Table 1, p. 51). The authors conclude that “the population and sample propensity scores are approximately equally effective in removing bias, so that no substantial loss is incurred by having to estimate the propensity score”.

What must hold before a matched comparison counts as causal?

§ 3 of 8

Two, beyond the usual condition that each unit receives one version of the treatment and that one unit's treatment does not change another's outcome, and the data can check only one of them. Gordon et al. (2019)¹ name both. The first is unconfoundedness: given the observed covariates, the potential outcomes are independent of treatment, which means that “no unobserved characteristics of individuals associated with the treatment and potential outcomes exist”. “This particular assumption is considered the most controversial and is untestable without an experiment.” The second is overlap: at every value of the covariates the probability of treatment lies strictly between 0 and 1, so that every kind of unit could have been either treated or untreated. Unlike the first, it can be checked: “Overlap can be assessed before and after adjustments are made to each group.”

Together they make up what Rosenbaum and Rubin (1983)² call strongly ignorable treatment assignment. In a randomized trial with properly collected data, the covariates are known to include everything that was used to assign treatment and may be related to the outcome; outside one, “This condition is usually not known to hold in a nonrandomized experiment.” A randomized trial also gives every unit a chance of each treatment. Where strong ignorability does hold, Gordon et al.¹ put the consequence plainly: “Given two individuals with the same propensity scores, exposure status is as good as random.”

Gordon et al.¹ dealt with overlap by dropping units with extreme estimated scores, where comparable partners are scarce. They removed users whose estimated score was below 0.05 or above 0.95, following Imbens, and estimated the score again on the rest, to “apply a simple trimming rule to improve overlap in the covariate distributions”. On this guide's reading, trimming changes whom the estimate describes: it is then the effect on the treated units that remained, not on all of them.

In advertising, the targeting itself puts overlap under pressure. If they observed the exact data and structure the ad-targeting systems use, Gordon et al.¹ write, “the propensity score distribution would collapse to discrete masses at 0 and 1”, because “a deterministic exposure system implies that common support in observables between treated and untreated observations cannot exist”. Their data still held chance variation in exposure, from Facebook's pacing of each campaign's delivery, the bids of advertisers whose interests are unrelated to the campaign's outcome, and the timing of users' visits, so the methods

need not fail; “However, as ad-targeting systems become more sophisticated, such failure is increasingly likely.”

How do matching, stratification and weighting on the score differ?

§ 4 of 8

In how they use the score to build the comparison, and in which average they return. Rosenbaum and Rubin (1983)² apply the score in “three standard techniques for adjustment in observational studies”: matched sampling, subclassification and covariance adjustment.

Matching pairs each treated unit with the untreated units closest on the score. Exact pairs are rare: “Unfortunately, exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used.” Gordon et al. (2019)¹ took, for each exposed user, the M unexposed users closest on the estimated log-odds of the score, and matched with replacement “because it can reduce bias, does not depend on the sort order of the data, and is less computationally burdensome”. Because the matched units stand in for the treated units’ missing outcomes, the estimate is the effect on the treated. Rosenbaum and Rubin² list reasons to prefer matching, the first being that “Matched treated and control pairs allow relatively unsophisticated researchers to appreciate immediately the equivalence of treatment and control groups”, and they note that matching on the score “can effectively balance rare binary variables for which it is not possible to match treated and control units adequately on an individual basis”.

Stratification, which Rosenbaum and Rubin² call subclassification, cuts the score into bands and compares treated and untreated units within each band. Grouping on many covariates at once soon leaves bands that hold only treated or only untreated units; “Subclassification on the propensity score is a natural way to avoid this problem.” In their medical example, they formed five bands of equal size at the quintiles of the score, “each containing 303 patients”, following Cochran’s observation that five subclasses remove “at least 90% of the bias for many continuous distributions”. The band estimates are then averaged. Weighted by each band’s share of the population, the average estimates the effect for everyone; Gordon et al.¹ weighted “corresponding to the fraction of treated users in the stratum relative to all treated users”, which gives the effect on the treated. In practice, bands are not perfectly even in the score, so “the directly adjusted estimate may contain some residual bias due to x ”.

Weighting uses the score to weight units rather than to pair or group them. The one weighting method in these sources is the inverse-probability-weighted regression adjustment that Gordon et al.¹ tested, which “uses the propensity score to weigh observations in the regression model”: “The estimated propensity scores are used to form weights to help control for correlation between treatment status and the covariates.” They place it among doubly robust procedures, which means that “the estimator is consistent even if one of the underlying models”, the propensity model or the outcome model, is misspecified. Gordon et al. use it, like their matching and stratification, for the effect on the treated: the weighted regression predicts what the exposed users would have done unexposed. Weighting without a regression is tested in neither source, and this guide does not compare it. Plain regression adjustment uses no score, and where treated and untreated users differ widely it “relies heavily on extrapolation, which is likely to produce biased estimates of the treatment effect”. Stratifying on the score and running a regression within each band combines the two, so that “the within-stratum regression is less prone to extrapolate”. Rosenbaum and Rubin² add a caution about regression on the covariates: it “cannot be relied upon to perform well unless the linear discriminant is highly correlated with the propensity score”.

Which variant works best, the Facebook data did not settle: “We do not find that one method consistently dominates”. In the study the authors walk through first, “whether a method relies on an outcome model seems unimportant”.

How is balance checked after matching?

§ 5 of 8

By comparing the treated and the matched untreated units on the score and on every covariate, before and after. Rosenbaum and Rubin (1983)² treat leftover imbalance as a warning: “evidence of residual bias in the propensity score is evidence of potential bias in estimated treatment effects”. In their study of coronary artery bypass surgery against drug therapy, they compared the two groups on each of 74 covariates with an F statistic before and after forming the five bands; the median fell from 10.9 to 0.2 (Table 2, p. 52), and they report that “within the constructed subclasses there is greater balance than would have been expected if treatments had been assigned at random within each subclass”. The check need not be elaborate: “The comparability of units within subclasses can be verified by the simplest methods, such as bar charts of means.”

Gordon et al. (2019)¹ checked two things. For the score, “we inspect the histograms of the estimated propensity scores by treatment group”.

For the covariates, “we examine standardized differences in covariates between the exposed and unexposed users, before and after matching”, dividing each difference in means by the pooled standard deviation. They prefer this to a t-statistic, which “might be large if the sample is large, even if the difference is substantively small”, and cite a suggested ceiling of 0.25 standard deviations from Stuart and Rubin. In their study 4, a score built from the first set of variables brought differences of up to about 0.4 standard deviations below 0.1, but “balancing on the FB variables does little to reduce the mean differences for the other variables, some of which have differences approaching 1 standard deviation.” On this evidence, a score can be relied on to balance only what went into it; variables left out of the model may stay far apart.

Balance was not what failed in the Facebook studies. “Balancing on the propensity score substantially reduces the mean differences in all the studies and achieves good balance on the underlying observables”, the authors report, and yet “Many of the observational models are unable to recover the RCT treatment effect, even though matching on the propensity score achieves good balance on the propensities themselves and on the underlying distribution of covariates.” Balance shows that the measured covariates are alike; “the unconfoundedness assumption requires not only balance on observables but also that no unobservables exist that might be correlated with treatment and potential outcomes”.

What happened when matching was held against Facebook’s experiments?

§ 6 of 8

The estimates were often far off, in both directions. Gordon et al. (2019)¹ used “15 large-scale advertising campaigns conducted on Facebook as RCTs in 2015”, collected between January and September 2015, all in the United States and restricted to users aged 18 and over. In each, users in the target audience were assigned by chance to a test group that could see the campaign or a control group that never saw it; control users were “shown the full distribution of ads they would have seen if the advertiser” had not run the campaign. Two of the authors chose the studies by criteria that made them suitable for comparing measurement methods, across retail, financial services, e-commerce, telecom and technology advertisers, with about 2 million to 140 million users each, and the authors state that “the sample of studies is not representative of all Facebook advertising (nor is it intended to be)”. Results are given as lift: the extra conversion among exposed users as a percentage of the conversion they would have had without the ads.

To mimic an advertiser without an experiment, the authors set the control group aside and compared exposed with unexposed users inside the test group. That comparison is biased because exposure is not random: “activity bias arises because a user must visit Facebook during the campaign to be exposed”, the delivery system “learns which types of users are most likely to meet the objective” and favours them, and other advertisers’ bids shift who sees which ad. They tested six methods: exact matching on age and gender, stratification on the score, propensity score matching, regression adjustment, inverse-probability-weighted regression adjustment, and stratification with regression. Except for exact matching, each ran on four nested sets of variables: common Facebook variables such as age, gender and friend counts; “nearly 40 variables drawn from the most recent Census and American Communities Surveys” by zip code; deciles of activity on Facebook in the week before the study; and a match score, “a composite metric of Facebook data that summarizes thousands of behavioral variables”. What they lacked was the advertiser’s own record of past outcomes: “we do not observe precampaign outcome histories that vendors might have received from an advertiser”, so the exercise is “most comparable to what an advertiser would have at their disposal for new customer acquisition”.

Three of the patterns they read off their summary table: “the observational methods we study mostly overestimate the RCT lift”, although “in some cases they can significantly underestimate RCT lift”; the point estimates “in 7 of the 14 studies with a checkout-conversion outcome are consistently off by more than a factor of three”; and the methods “do a better job of approximating RCT outcomes” for registrations and page views than for purchases. In study 4, the case the incrementality guide reports, propensity score matching with all four variable sets combined came to 95 per cent against the experiment’s 73 (Figure 8, p. 21 of the Articles in Advance version), a difference the authors could not test for matching. In study 9, with an experimental lift of 2.4 per cent, “The closest lift estimate is 1,306% (RA4), a massive overestimate”, which the authors “speculate” is partly due to an exposure rate of only 6.6 per cent. In study 15, “Except for exact matching on age and gender, all methods underestimate lift.”

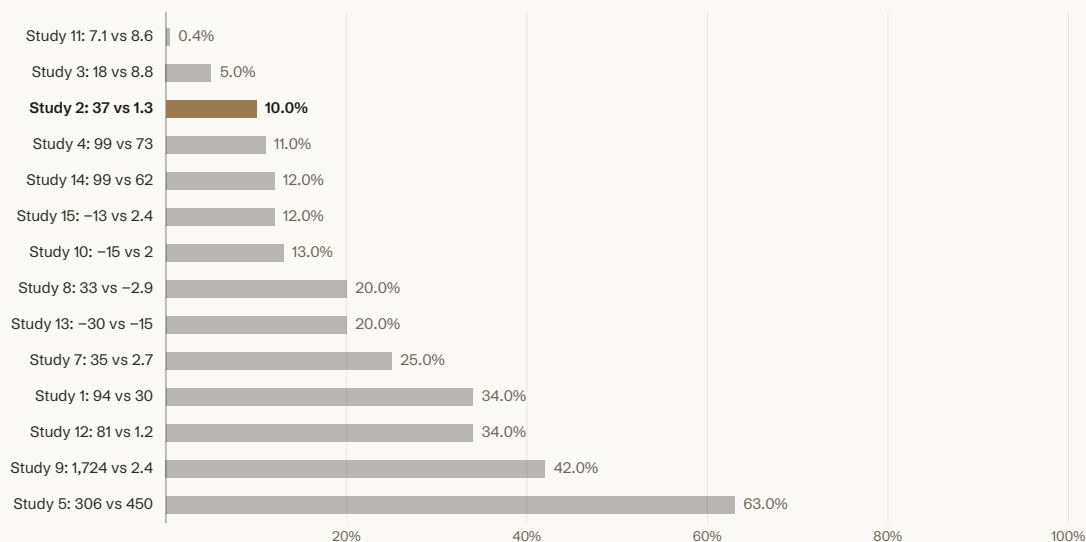
The benchmark has its own noise. “Of the 14 studies with a checkout conversion, six failed to produce statistically significant lifts at the 5% significance level”, and the authors warn that “One downside of using lift is that differences between methods can seem large when the treated group’s baseline conversion rate is small.” For the 14 checkout studies, they also report how much of the selection effect was left after

stratifying on the score: the distance between the stratified lift and the experimental lift, as a share of the distance between the raw exposed-versus-unexposed lift and the experimental lift. Figure 1 shows that share.

The authors give this ratio for stratification on the score, not for matching, and run their sensitivity analysis with stratification too, because it is quick enough to repeat over many simulated draws. Their study-by-study table puts the two side by side: with all four variable sets combined, the matching estimate lay within four percentage points of the stratified one in 10 of the 14 checkout studies, for example 95 against 99 per cent in study 4, and no more than 14 points away in any of them; the widest gap was study 9's, 1,710 against 1,724 per cent (Figure 10, p. 24 of the Articles in Advance version).

Figure 1 Share of the raw comparison's bias left after stratifying on the propensity score, 14 Facebook checkout studies

Stratifying on the propensity score left between 0.4 and 63 per cent of the raw comparison's bias. A small remainder could still exceed the effect: in study 2, 10 per cent left was 37 against 1.3.



Horizontal axis: share of the raw comparison's bias left after stratifying on the propensity score, from 0 to 100 per cent. Labels read: lift from stratifying on the score against lift in the experiment, in per cent. The 95 per cent intervals in the paper's Table 4 for the experimental lifts of studies 2, 7, 8, 9, 10 and 13 include zero. In studies 5, 10, 11, 13 and 15 the stratified lift lies on the other side of the experimental lift from the raw one, so the share there measures an overcorrection rather than a remainder. Table 6 prints the stratified lift of study 7 as 35; the paper's Figure 10 and Figure A.1 (pp. 24 and 29 of the Articles in Advance version) print it as -35 and -35.3, which would put that study's share at about 30 per cent, on the overcorrection side.

Source Gordon et al. (2019), Table 6, p. 27 (pages of the Articles in Advance version held, pp. 1 to 33): the bias reduction ratio for the 14 studies with a checkout outcome, in the order of the table. The ratio is the distance between the lift estimated by stratifying on the propensity score and the experimental lift, divided by the distance between the raw exposed-versus-unexposed lift and the experimental lift. Labels give the stratified and the experimental lift from the same table, in per cent. Ratios are shown to one decimal place because the table prints 0.4 for study 11; the others are whole numbers in the table. The authors' figures, not a re-estimation.

Stratifying on the score removed more than half of the raw comparison's bias in 13 of the 14 checkout studies, but what it left could still be many times the experimental lift. The authors' own example: although they "eliminated 90% of the bias in study 2 (brr = 10%), we need data that would allow us to reduce the remaining bias from 37%

to 1.3%". In study 9, a remaining 42 per cent was 1,724 against an experimental 2.4.

How strong would a missing variable have to be? The authors simulated an unobserved variable that removes the remaining bias and compared its explanatory power with that of the variables they had. In study 4, "a relatively weak unobservable is required to remove the bias entirely", about as strong as the census variables; in study 1, one as strong as their fourth variable set; in study 9, "the unobservable would need to have between 5 and 10 times as much explanatory power as the observables in our data". Their reading: "This suggests that eliminating bias from observational methods would be hard, even for industry insiders with access to additional data."

What should travel with a matched estimate?

§ 7 of 8

Six fields, in one record per estimate: the estimand, the covariates and when they were measured, overlap and what trimming removed, balance on every covariate before and after, the unmeasured causes that could drive both treatment and outcome, and the benchmark: an experiment the estimate has been held against or, failing one, a sensitivity analysis of how strong an unmeasured cause would have to be. The fields follow Rosenbaum and Rubin's² conditions and checks and the steps Gordon et al.¹ took and reported. The [uplift specification essay](#) gives a sheet for what any uplift claim is about; this record covers what a matched estimate needs on top of it.

Figure 2 A matched-estimate record

Write down the estimand, covariates, overlap, balance, unmeasured causes and benchmark before a matched estimate is used as an effect.

Estimand	Covariates	Overlap	Balance	Unmeasured	Benchmark
Effect on which treated units, over what window?	Measured before treatment? Any it could move?	How many treated units were trimmed?	Standardized differences, before and after?	What drives treatment and outcome unrecorded?	Experiment or sensitivity analysis, and result?

Without an experiment, report an adjusted association and how strong an unmeasured cause would have to be to remove it.

One row per estimate. List balance for every covariate, including those left out of the score model.

Source Author's decision worksheet. The estimand and covariate fields follow Rosenbaum and Rubin (1983), pp. 42 and 43, and Gordon et al. (2019), pp. 7 to 10; the overlap field follows Gordon et al. (2019), pp. 10, 12 and 16; the balance field follows Rosenbaum and Rubin (1983), pp. 48 and 52, and Gordon et al. (2019), pp. 16 to 19; the unmeasured and benchmark fields follow Gordon et al. (2019), pp. 22 to 27. Pages of Gordon et al. are those of the Articles in Advance version held. The fields are reader inputs; no company data is supplied.

An illustration, with no company data: a B2B software firm's marketing team wants to know what a customer webinar series did for expansion revenue, and only accounts that chose to attend took part. The record names the estimand as the effect on attending accounts over the two quarters after they attended. Its covariates are seats, plan, tenure, product usage and support tickets from the quarter before the invitation; usage after attendance is left out, because the webinars could have changed it. It reports how many attending accounts had no comparable non-attending account and were dropped, so that the estimate describes the rest. It lists standardized differences for every covariate, including those not used in the score model. Under unmeasured causes it names an account already evaluating an upgrade, which is more likely both to attend and to expand and which no field records. Until an experiment is run, the benchmark field names

a sensitivity analysis instead, and the estimate is reported as an adjusted association, with how strong an unmeasured cause would have to be to remove it. The team could withhold invitations by lot from a share of eligible accounts for two quarters, provided that accounts without an invitation cannot attend, record each group's expansion in the quarter before as a baseline, and compare the matched estimate with the experiment's. Gordon et al.¹ turned such a comparison into the effect on the exposed by dividing it by the share exposed, which assumes that assignment works "only through receipt of the treatment". Their users did not know their assignment; an invitation is itself a message to the account, so the team would have to argue that being invited without attending leaves expansion unchanged, or report the effect of the invitation, the intention-to-treat effect, on its own. With a few hundred accounts, both estimates will be far less precise than in studies of millions of users. These are the author's examples and have not been tested.

Which parts of the Facebook result carry over to other data?

§ 8 of 8

The authors draw the line themselves. "Of course, our results should not be interpreted to suggest that causal inference methods are unable to replicate experimental findings in all settings." They report that Eckles and Bakshy, measuring peer effects rather than advertising on Facebook data, found that high-dimensional models remove most of the bias of naive comparisons, and that the methods "may work to recover the causal effects of advertising that is not subject to the same degree of targeting as advertising on Facebook". In their conclusion, "the failure of observational methods is specific to the Facebook data we had at our disposal", "performing this analysis on a different advertising platform may yield different results", and "With our small number of studies, we could not identify campaign characteristics that are associated with strong biases." Against that, they count their data as rich, and Facebook's closed system "makes solving selection issues an easier problem than many other ad-effectiveness applications". Two of the authors list Facebook as their affiliation; the paper discloses that the other two were not paid by Facebook for the research and were contingent Facebook employees on a token salary, which they gave to charity.

The question is older than digital advertising. Gordon et al.¹ recount that LaLonde compared observational methods with randomized experiments in employment and training programmes and found that many of them did not reproduce the experimental results, and that a

later application of improved methods to the same data by Imbens did. Their reason to expect more trouble in advertising: “Because of targeting, the endogeneity problems associated with digital advertising are potentially more severe”. Rosenbaum and Rubin (1983)² supply the method’s theory, with an applied example from an observational study of coronary artery disease and no experimental benchmark.

On this guide’s reading, two points carry over to B2B marketing data with some confidence: balance on measured covariates cannot show that a comparison is unconfounded, and a matched estimate needs a benchmark before it carries a budget decision, a point the [attribution essay](#) makes for attribution output. How large the bias is when accounts, not users, are matched, when buying cycles are long, and when covariates include past purchases, which the Facebook data lacked, the sources do not say.

 End of essay

References

- 1 Gordon, B. R., Zettermeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2), 193-225. <https://doi.org/10.1287/mksc.2018.1135>
- 2 Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. <https://doi.org/10.1093/biomet/70.1.41>

How to cite this essay

Isoglu, S. (2026, October 18). What is propensity score matching? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-propensity-score-matching/>
Online version:
<https://isoglu.com/insights/journal/what-is-propensity-score-matching/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved