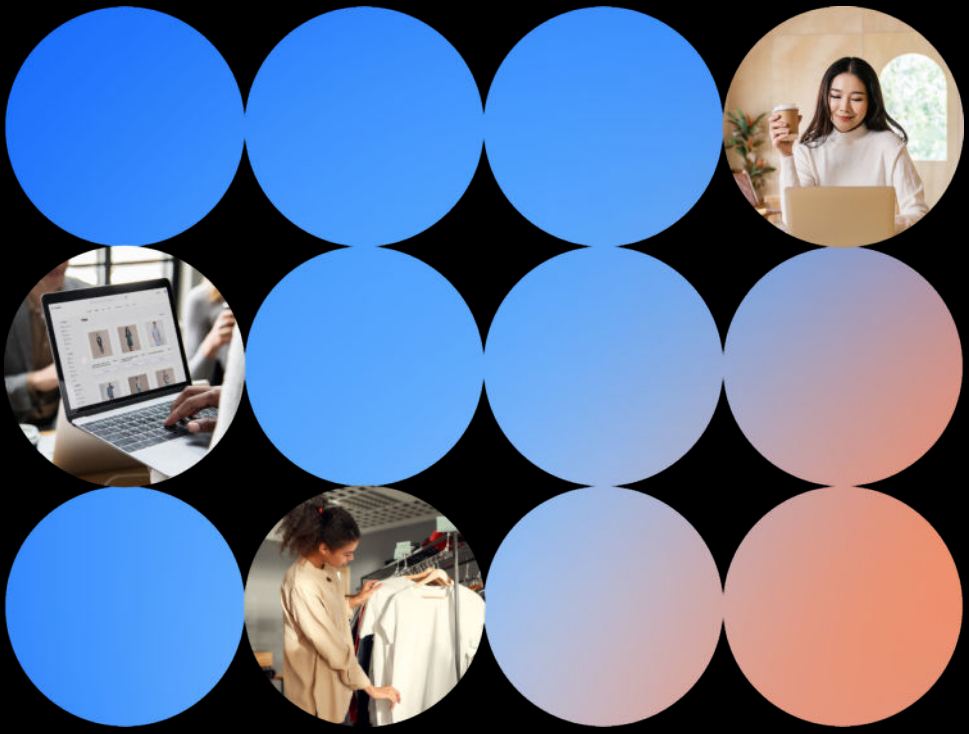
Constructor

The Ultimate Guide to A/B Testing for Search & Discovery

By: Constructor Data Science & Integrations Team



Why we wrote this

Optimizing search and discovery experience is still one of the most underleveraged opportunities in ecommerce.

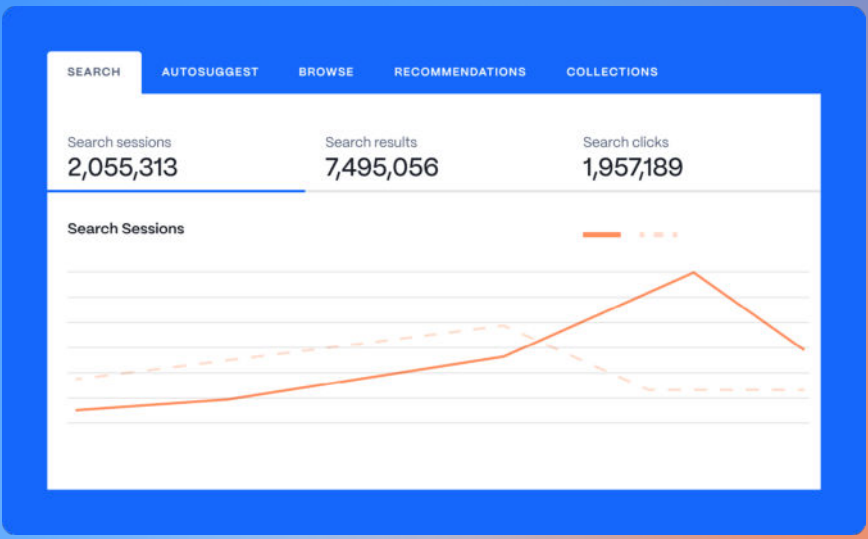
Even though AI and machine learning go a long way to self-optimize product ranking by relevance and attractiveness based on real-time user behavior, there are still a number of use cases for A/B testing merchandising and UX decisions across Search, Autocomplete, Browse, and Recommendations experiences that can lead to improved revenue and profitability.

At Constructor, our dedicated team of A/B testing experts help guide experimentation programs for our platform customers year-round. We've seen it all, and we believe every merchant should incorporate split testing into their digital operations. But testing is both art and science (okay, and math).

We wrote this Ultimate Guide to A/B Testing for Search & Product Discovery to demystify testing if you're new, reinforce technical and tactical best practices, and provide test case ideas to inspire your continuous A/B testing efforts.

What you'll learn:

- Why A/B testing search & discovery matters
- What you should be A/B testing as an online merchant or product manager
- Tips for running valid A/B tests and interpreting results
- Practical test ideas across Search, Autocomplete, Browse, and Recommendations



No matter how much you invest into marketing, web design and UX, browsers won't convert to buyers unless they find the products they want. And the larger and more diverse your catalog, the more critical search and discovery experiences are to digital performance.

Enterprise search platforms like Constructor provide a wide range of settings and rules that can help you optimize the browse-to-buy journey. The merchandising strategies you can apply are virtually endless.

The question is, which strategies work best for your business and under what conditions?

For merchants who take a data-driven approach, A/B testing allows you to run controlled experiments with real users to validate your assumptions and quantify the impacts of your decisions.

Table of contents

Why A/B Test Search and Discovery Experiences?

The short answer is to keep sales and revenue continually moving up and to the right.

But A/B testing also provides a quantitative method for understanding how customers respond to different strategies to ultimately serve better experiences for every context and touchpoint.

Make data-driven decisions

Web analytics only provide data on the decisions you've already employed — not what you could (or should) have done instead.

On the other hand, merchandising decisions are often made based on subjective ideals that are hard to measure, like maintaining brand value.

While these decisions have their place, it's also important to know what effects merchandising has in the short term — because we only really know if our strategies are profitable if we can measure them against alternative strategies with live customers.

Most everyone has an opinion on UX and merchandising "best practice," but without real data, it's just opinion. A/B testing provides a scientific way to measure real impacts on business metrics (and a diplomatic means to settle internal debates around strategy).

Adopt and adapt to innovation

AI is rapidly transforming search and discovery in exciting ways. For example, Constructor is continually investing in new AI-enabled capabilities to support novel experiences and 1:1 personalization.

It can be helpful to measure the business impact of these new capabilities against your human-driven, carefully curated experiences. For example, you may want to test dynamically generated facets against your standard facet taxonomy.

A/B testing provides hard numbers to help you make decisions on what's right for your digital product.

Drive incremental revenue over time

One-off A/B tests rarely move the needle for large enterprises. The best results come with consistency.

Companies do best with an approach to A/B testing that's strategic, systematic, and supported by the entire organization.

Because Product, UX, and CRM teams often have their own A/B testing programs, it's important that no function tests in a silo without awareness and communication across teams, as experiments in one area can compromise the validity of tests by other teams.

Common KPIs for search and discovery A/B testing

A/B testing can be used to address any KPI, but the most common primary metrics for search and discovery are:

Conversion rate (CVR) – percentage of users that completed a purchase among users who encountered the test experience

Average order value (AOV) – total revenue divided by number of orders among users who encountered the test experience

Revenue per visitor (RPV) – total revenue from users who encountered the test experience divided by the number of these users

Revenue per session (RPS) – total revenue divided by the number of user sessions tracked

Your A/B testing tool will determine a "winner" based on a single primary metric you choose. It will also provide reporting on secondary metrics you include in your experiment.

Because it's common for online shoppers to make more than one returning visit before checking out, the Constructor testing team recommends tracking per-user metrics rather than per session to avoid attribution-related distortions, and to calculate the entire value of each user.

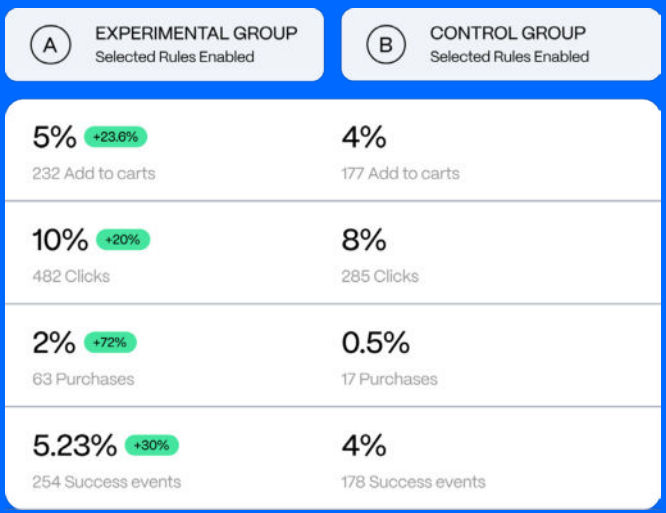
With search and discovery experiments, you also want to track micro-conversion events, such as:

- **Click through rate (CTR) –** percentage of users with at least one click from the search/discovery experience among all users served the experience
- **Add-to-cart rate (ATC) –** percentage of users that added a clicked product to cart from the search/discovery experience among all users served the experience
- **Exit rate –** percentage of users that left the site after being served an experience

Tracking micro-conversion events is particularly useful with search and discovery experiments. End-of-funnel conversion events happen less frequently than micro-conversions and are influenced by many other factors within the shopping journey. Revenue metrics are similarly "noisy" due to this large variance. More frequently occurring events, such as clicks and add-to-carts, can be used as proxies for CVR and RPV and as early indicators of successful test performance.

Testing more sensitive metrics also helps clarify the direct influence of search and discovery strategies and overcome the challenges with smaller sample sizes when experimenting within a specific category or search case.

You're not limited to the above metrics. You can track essentially any custom key event that's relevant to your objective, provided you send the event to both your search and web analytics applications.



What Should You A/B Test?

For most merchants, the main purpose for A/B testing is increasing business performance.

While there are an infinite number of hypotheses you can test, experiments need time to run and a reasonable amount of traffic (sample size) to reach statistical significance. There is an opportunity cost to running low-impact tests at the expense of higher-impact tests.

Although you technically can run multiple experiments at once, you risk them cross-pollinating each other and dirtying your data by introducing competing variables.

Just because you can doesn't mean you should. It's important to prioritize your tests based on business value.

Start with a business problem

Testing is a key part of a continuous optimization program, and it's wise to tackle your largest experience gaps first.

You may start from a website perspective and identify search terms and pages with high traffic and:

- Low click through rates relative to other pages
- High bounce and exit rates
- Low conversion rate and revenue per visitor relative to other pages

You may also start from a merchandising productivity perspective and look at underperforming brands, categories, and products and look for ways to improve their visibility across your site.

Not every A/B test needs to center around existing business problems, and not every business problem needs an A/B test to fix. It's wise, however, to ensure that underperforming terms and pages are regularly reviewed and considered for optimization.

Strong use cases for A/B testing

You're making a significant change to searchandising strategy

If your proposed strategies will be applied to a large number of experiences (categories, searches, landing pages, etc.) or significantly change the way products will be ranked, test them.

These are just a few of the scenarios that may warrant an A/B test:

Responding to consumer trends
A brand you carry recently received a wave of social media hype, and you want to test if everyday shoppers have been influenced by this trend by boosting the brand across search and browse.

Hitting revenue targets
Sales have been soft this quarter, and your current strategy boosts higher-margin house brands. You want to measure if boosting more recognized brands can beef up top line revenue.

Improving inventory turns
You're overstocked on expensive-to-store SKUs, and they're poor sellers. You want to both discount and boost them to give them a better chance, but you're concerned about the impact on revenue.

Bringing in fresh talent
A new Head of Merchandising has joined and believes promoting New products over Best Sellers is best for the brand. They want to change the default sort across all Category pages except Sale.

Your proposed change impacts UI/UX

Serving content, displaying badges, changing product tiles, and adding or reordering facets all change the user interface. So, it's important to quantify the impact of these changes to your UI.

For example:

- Adding banners or in-grid content cards in various grid slots or of various sizes
- Employing variant slicing (showing variants as individual listings)
- Adding or changing attribute badges
- Adding a recommendation pod or changing its location
- Changing autocomplete UI or behavior

You want to try customer segmentation strategies

You may have observed through data or simply hypothesize that different customer segments prefer different experiences.

For example, frequently returning visitors may want to see newest products first, or visitors coming directly from social ads may want to see trendier brands and products over staples. Or, different geographies with different seasonal weather may benefit from seeing products with specific attributes first.

Testing lets you answer these questions definitively.

You want to test manual vs machine

If you've ever wondered if pure 1:1 personalization or rules-led strategies drive more revenue, A/B testing can tell you.

But you may also be curious about the impact of specific personalization settings or features within your search application.

For example, Constructor offers personalization caching, which determines whether products should be reranked within a single session for a given user immediately after they perform actions that indicate specific intent or preferences.

Our optimization group at Constructor have A/B tested personalization caching across 12 customers spanning different languages, geos, and industries, and we've discovered no significant differences between states. This leads us to believe it's safe for merchants to simply employ their preference, but you can always A/B test in your own environment to make sure.

You're launching a new campaign

A campaign is a means of applying a set of searchandising rules across multiple search queries, browse categories, or facets together.

For example, to optimize for Holiday shopping, your team may be adding new 'Shop by' gift guide collections around recipients, price, and themes.

You want to test whether adding content cards to product grids linking to gift guides and gift finders within your core categories lifts or hurts revenue by guiding shoppers to relevant sections or distracting them from their current task.

An A/B test can help you capture data early in your campaign so you can roll out the winner to all users confidently. However, keep in mind that you need a large enough sample size to support a valid test in reasonable time.

You've identified cases where conversion may cannibalize revenue

Think about your ongoing promotion strategies. If you regularly run a rotation of flash events or week-long promotions on specific products versus entire categories or sitewide discounts (e.g., grocery and mass merchant), your engine may still credit these products as best sellers for a period after the sale ends (typically 14 days).

In such a case, you might test boosting currently 'on sale' products against your default ranking method. You may discover that sorting by best selling doesn't hurt clicks or conversions and results in more full-price merchandise sold with a positive impact on AOV and RPV, even if the 'sale boost' cell produced a higher conversion rate.

Your searchandising rules have become stale

The more seasonal or dynamic your catalog, the more prone your searchandising strategies are to becoming outdated.

We recommend time boxing your rules with a start and end date for this reason. But even so, it's a good idea to revisit your strategies from time to time. A/B testing can give you definitive data on if and to what degree your previously applied rules have become stale.

You're expanding your business model

Foraying into new territories like adding B2B bulk buying, third-party marketplace sellers, re-commerce (used or returned items), or sponsored listings to your catalog impacts search and discovery.

You may wish to control these rollouts conservatively with 90/10 or 75/25 splits to ensure your searchandising rules and UI changes are working and not cannibalizing revenue until you perfect the recipe.

Running Your A/B Test

Once you've decided what to test, there are a few critical steps to take before launch:

- ◯ Define your hypothesis
- ◯ Define your primary and secondary metrics
- ◯ Choose your split
- ◯ Estimate your sample size and duration
- ◯ Determine your testing environment
- ◯ Choose your excursions
- ◯ Consider your stakeholders

Define your hypothesis

A hypothesis in search and discovery testing states your assumptions of how a proposed change will impact business metrics.

When defining your hypothesis, you can follow this simple template:

> We have observed [facts that indicate your problem] by [your data source or method]. We wish to [business goal] for the segment [visitor attributes]. This will lead to [anticipated outcome] measured by [primary and secondary KPIs].

Define your primary and secondary metrics

The success metrics you define are part of your hypothesis statement, and they're the very reason for running your tests. It's important to choose your primary KPI wisely, as your testing tool will often use it in its Probability to Be Best calculation and when crowning a definitive winner.

Of course, in practice you're interested in how your experiment impacts a handful of metrics. While your testing tool may only accept a few secondary metrics in your test set-up, connecting your test to your web analytics enables deeper analysis.

Choose your split

Most A/B tests follow a 50/50 split, but when new rules or elements are expected to have a significant or unknown impact on your experience, you can reduce your risk by rolling out your testing variation to only a subset of users. For example, you can run a 90/10 or 75/25 split (control vs variation) rather than a 50/50 split.

Keep in mind that uneven splits typically require a longer testing time frame to reach statistical significance, given that fewer conversion events are captured within the smaller segment. It's best to reserve this strategy for specific test cases.

There are a couple cases where you should absolutely throttle your traffic split to enable a series of "ramp-up" rounds that enable your search and discovery's machine learning to catch up to normalized user behavior. For instance, this is the case when running an AABB test between a new and existing search solution or when turning on new AI-driven features for the first time.

Estimate your sample size and duration

Before starting an A/B test, it helps to check whether you have enough traffic for the test to complete within reasonable time. Estimating your sample size can flag which experiments don't have enough traffic gas to reach the finish line.

Most experts suggest running your experiment through at least two full business cycles (i.e., two calendar weeks) and not longer than eight weeks. The reason is to avoid the influence of seasonal or campaign-based changes to your traffic mix or customer behavior.

When planning tests, it's important to consider whether upcoming events or changes to your site may introduce new variables into your experience which may impact your 'controlled' conditions when judging your maximum test duration.

For example, an experiment that may last six weeks should not be launched three weeks before Black Friday.

You don't need to be a math whiz to estimate sample size (although it helps). Many testing platforms provide sample size calculators, or you can find a free online tool.

The conversion rate for your control and your expected Minimum Detectable Effect (MDE) also play into the equation. Larger expected lifts will calculate smaller sample sizes and test durations. Your control conversion rate is known, while your MDE is estimated. So, your sample size and duration estimates will be just that — estimates.

It's also essential to calculate the duration for all metrics that will influence the decision taken from test results. This doesn't mean every metric you're planning to measure, but specifically the success and guardrail metrics.

For instance, if CVR is your primary success metric and RPV your guardrail and the expected effect on purchase conversion is 2%, and you want to ensure revenue per user doesn't drop by more than 3%, you must calculate the duration for both metrics based on the given MDE and baseline values. You should then choose the longest calculated duration to ensure that changes in both metrics can be detected within this timeframe, as both are part of the decision-making process.

And note that if your test contains more than two variations (A/B/x) or you choose an uneven traffic allocation, you'll need a larger sample size and duration than for A/B.

About statistical significance

With any A/B test you run, there's always a possibility your results are due to chance even when your testing tool declares a winner.

This is the nature of statistics and probabilities across a distribution. You may have heard about Type I and Type II errors where you may accept a false positive or negative result and end up implementing a 'win' that didn't really win or rejecting a 'loser' when there was no true negative result outside of your reported data.

Commercial A/B testing tools generally follow a standard practice of targeting a 95% confidence level (equivalently, a 5% significance threshold), and an 80% power. This means there is a 5% risk of detecting a false effect that doesn't really exist (Type I error) and a 20% risk of failing to detect a real effect that does exist (Type II error).

Unless you have a strong reason to use your own acceptable thresholds for confidence and power, sticking with your tool's defaults is safe.

Determine your testing environment

Depending on your testing case, you may be able to run your experiment through your search application's native testing interface, or you may need to use a third-party testing platform like Adobe Target or Optimizely for more complex scenarios.

For example, within the Constructor dashboard, you can test global search and browse rules and recommendation strategies, but an experiment involving both searchandsing and page template UI changes would be run through a third-party tool.

Choose your exclusions

It's also important to keep your test cells clean from unwanted or irrelevant users. For example, exclude internal user IP addresses and bot traffic.

You may also want to exclude specific geographies, especially when their catalogs are very different (same goes for B2B accounts where a single bulk order can radically skew RPV and AOV values).

Consider excluding traffic sources such as email or SMS marketing from which visitors may exhibit their own set of behaviors or be motivated by personalized offers. It's also a good idea to exclude outlier customers, such as the top 1–5% of purchasers by revenue.

Consider your stakeholders

Be mindful that colleagues outside your group and future team members may review your tests and archives without being privy to the reasons you are running an experiment.

Follow intuitive naming conventions for tests and variants, and provide easy-to-understand documentation that summarizes not just your testing hypothesis, but also the specific strategies and elements involved in your variant design and notes on your testing rationale.

AABB test

If you're migrating from one search and discovery vendor to another, it's best practice to run an AABB test with both applications head-to-head to catch any problems with setup.

An AABB test essentially runs AA tests (i.e., no changes) between both platforms, and results are compared to each other (AA cells for incumbent and BB cells for your new platform).

If everything has been set up correctly, you should not observe any statistically significant differences between AA or BB cells. Otherwise, you should investigate where there may be a problem with traffic splitting or any other technical aspect of your setup.

The AABB test requires both applications to be configured in a similar way using the same production catalog and sending the same events, with identical searchandising rules and UI components across search and browse.

Constructor advises performing the same AABB test in 3 separate and discrete rounds to give time for your new tool to machine-learn:

- **Ramp-up 1:** Send 45% of traffic to each vendor's A variant and 5% to each B
- **Ramp-up 2:** Allocate 37.5% to each vendor's A variant and 12.5% to each B
- **Even split:** Send 25% to each test cell evenly

Between each ramp-up round, users should be "re-shuffled" between test cells to reduce user memory effects and maintain random distribution.

Interpreting Test Results

Although your testing tool will typically tell you if your test produced a winner and its probability for success, it's important to take in the full picture of your experiment data as it relates to your business goals, look for sources of potential bias, and apply data analysis best practices.

Don't jump to conclusions

When you first launch a test, it's normal to see a large difference between conversion events due to the effect of the smaller sample of users early in the experiment. This spread can get you very excited and may even tempt you to declare a winner prematurely.

As the experiment powers on, you'll often see a cooling off effect and a closer race. It's important to let your experiment run long enough for your tool to capture enough sessions and events to reach a reliable result.

Many experts say "don't peek!" while an experiment is running, but it's not the peeking that's the problem. It's drawing premature conclusions. Stopping a test too soon and implementing the leading variant is as reliable as not having run the test at all.

Likewise, resist the temptation to make adjustments mid-test, such as changing the percentage split towards a favored result. If you need to tweak, stop the test and restart fresh.

End tests that are dragging on

If your test fails to produce a statistically significant difference within the predefined testing duration, end the test and declare the result as inconclusive to make room for other tests.

It's reasonably safe to go forward with the variant you prefer as the odds are there was no true difference between the two experiences, or traffic was low enough that using 'the wrong one' won't make or break your business performance.

Or, you may choose to restart the test under conditions that would draw a larger sample, such as more search terms or browse pages. But weigh that decision against the opportunity cost of not testing something else in your queue.

Check for KPI cannibalization

Your A/B testing tool does a lot of the math for you to determine a winner based on your primary metric, but you want to pay attention to the details — especially if you observe a lift in one metric and a dip in another.

For example, your variation may win on CVR and lose on RPV. Make sure your data interpretation takes into consideration the full implications of accepting a winner and factor to what degree secondary metrics may be impacted.

Beware of bias

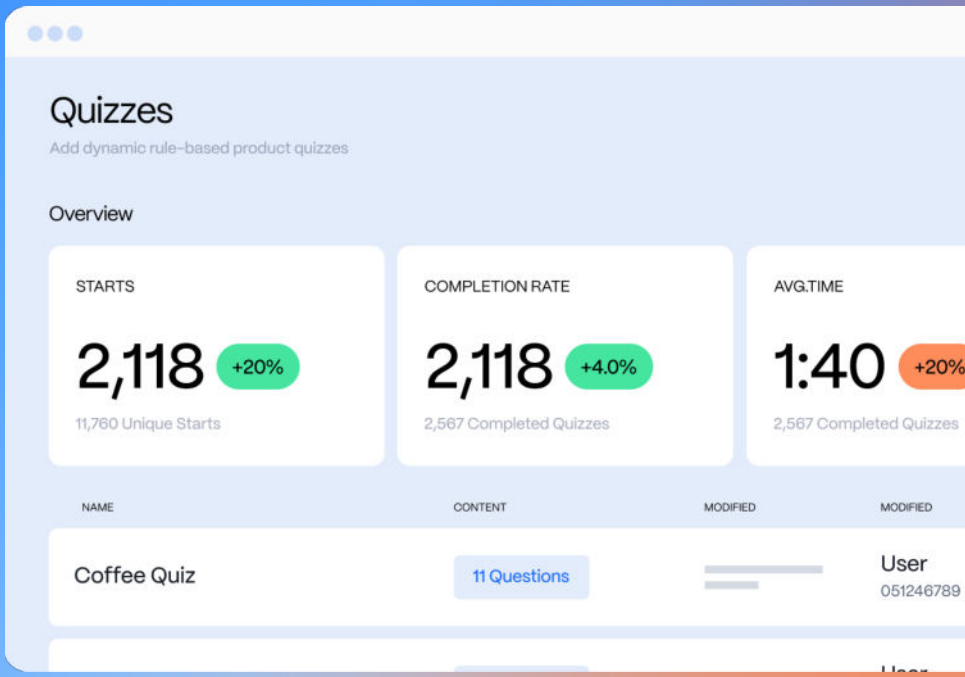
There are several sources of bias that can sneak into your A/B tests if you're not careful. Make sure to consider these factors both ahead of launching your test and during analysis:

- **Selection bias.** The groups exposed to your variations are not comparable. For example, in an AABB test between 2 search vendors, you've included all geographies in vendor A's cells and only US traffic in vendor B's cells
- **Sample size bias.** If your sample size is too small, your results may be skewed by random variations, leading to false positives or negatives
- **Novelty bias.** Users may respond differently to new designs and experiences simply because they're new, not because they're better. For instance, it's often observed that conversion drops after a UI redesign until shoppers 'learn' how it works
- **Device bias.** If one variant is not optimized for a specific browser or device context, it can underperform and skew results
- **Technical bias.** Server-side or client-side implementation errors may cause some users to experience glitches in one variation, causing premature abandonment that's not related to test variables themselves
- **Seasonal bias.** If external factors like sale events, holidays, unexpected traffic spikes, or significant changes to your catalog occur during your test, they may influence your data
- **Confirmation bias.** If you have a strong preference for a variant winning, you may be tempted to end a test early when it's ahead or use results to confirm pre-existing assumptions
- **Testing conflicts.** If other team members are running concurrent experiments that introduce new and competing variables, both tests may be compromised

Look for analytic anomalies

Your analytics can also uncover unusual changes to traffic or user behavior during your test period that your merchandising group wasn't aware of, such as unexpected downtime or a spike from a specific traffic source.

Use your judgment whether the anomaly is significant enough to call your test validity into question.



Search and Discovery A/B Testing Ideas

While there are infinite use cases for A/B testing search and discovery, these are a few practical and inspirational ideas to get you started.

Search

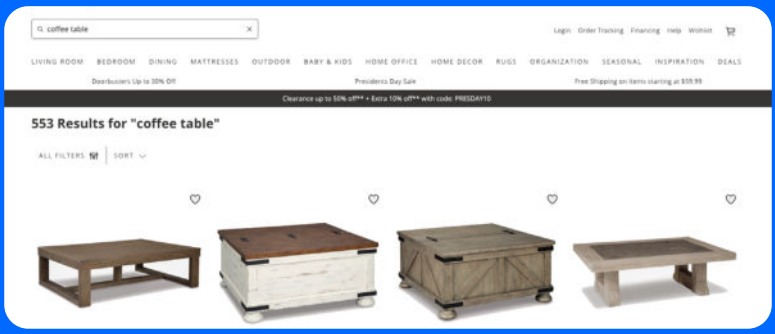
Category redirects

If you've invested in building beautiful browse pages, it makes sense to redirect exact match searches directly to them to leverage enhanced UI and curated merchandising.

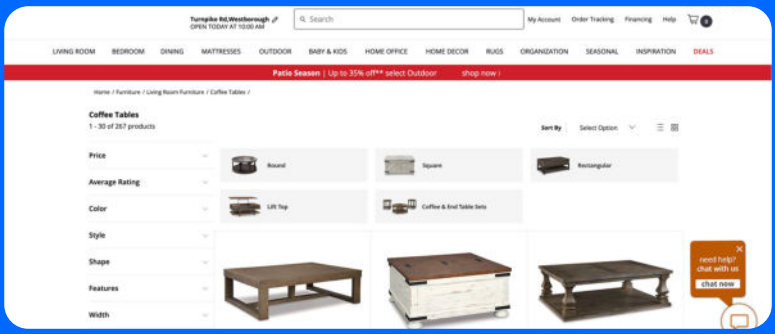
But sometimes you're unsure whether a category redirect or a search experience truly performs better while shoppers are in a search context.

Search results pages enable you to show suggested search refinements and show users a familiar UI that matches their expectations. Search and browse rankings are also influenced by separate rules and algorithm factors.

For example, this furniture retailer's search and browse pages display different grid and filter UI. Search returns 553 matches with four results per row and minimized facet menu, and browse includes 'Shop by Shape' filter links and displays three products per row with only 267 results.

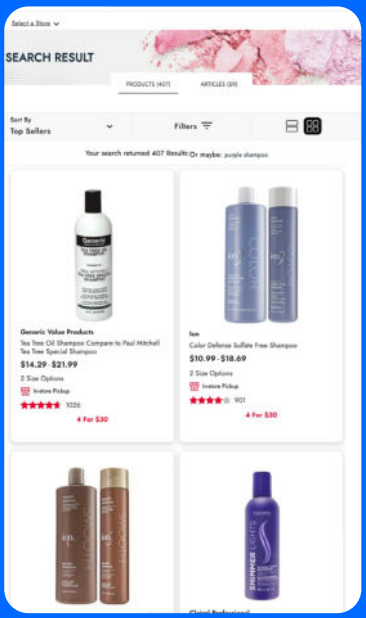


Search results for "coffee table" with 553 results

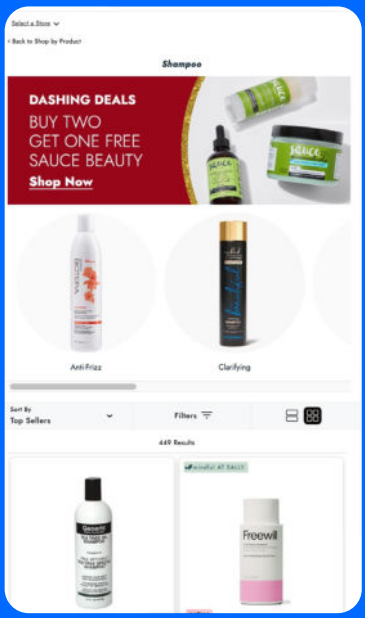


Coffee Table category with 267 results

Banners and other graphics on browse pages can also dramatically change mobile experience. Because redirects impact users across devices, it's important to test redirect performance against none, head-to-head.



Mobile search results for "shampoo"



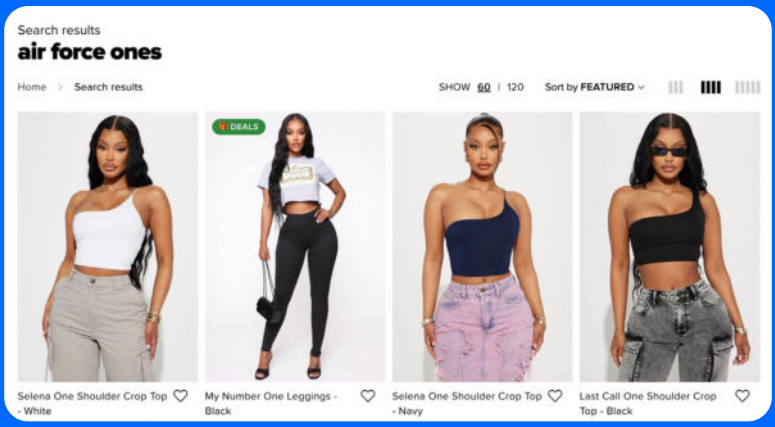
Mobile browse page for Shampoo

Before running your test, review which search redirects you have in place and their respective browse page experiences. You may choose to include all redirected searches in your test group, or only ones with similar UI elements on the browse page to keep like-with-like (provided your anticipated sample size will be large enough, by aggregate search volume).

Merchandising 'no results found'

The power of fuzzy matching means you never have to send visitors to an empty search results page. Today, it's common practice to just show results even if there are no exact matches.

Some of these experiences, however, will be better than others — and there's a risk that matches can be so poor for certain queries that users lose trust in your entire website.

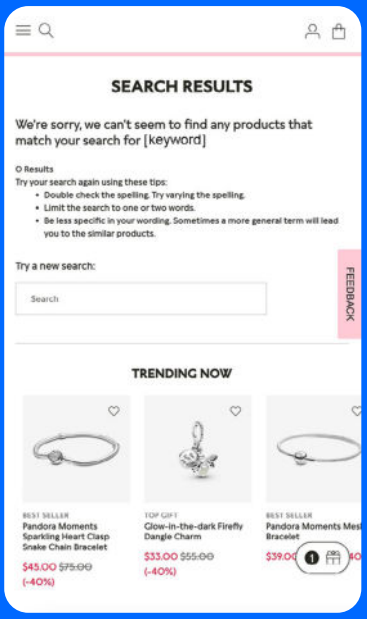


Sometimes search results have no relevance to a search query

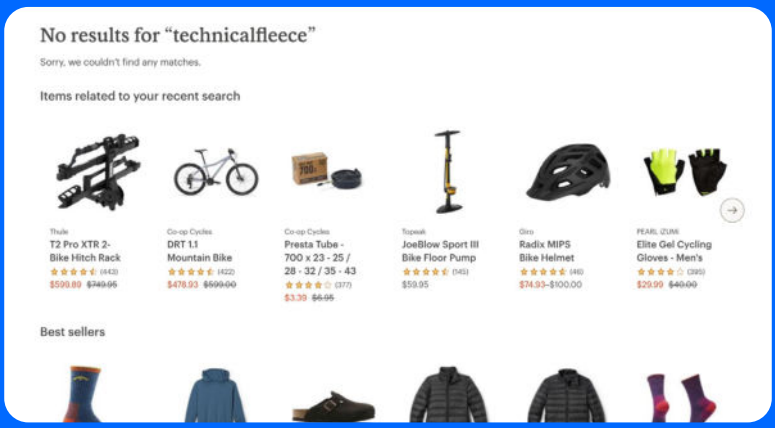
Consider testing this against the following alternatives.

Global 'no results' experience

If you've set your site to always show results, consider testing this strategy against serving a traditional 'no results' page with a prompt to refine their query and/or a few alternative recommendation pods.



A 'no results' strategy showing refinement tips, a visible search box, and a recommendations pod

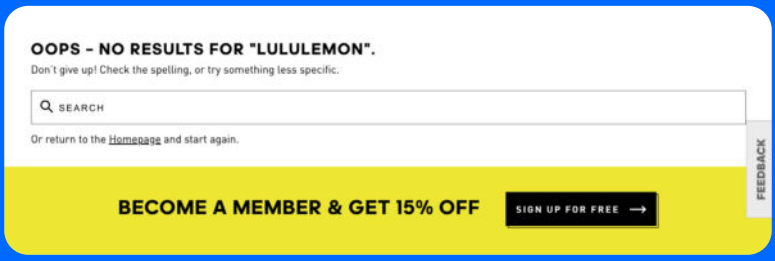


A 'no results' strategy showing related item and best seller recommendation pods

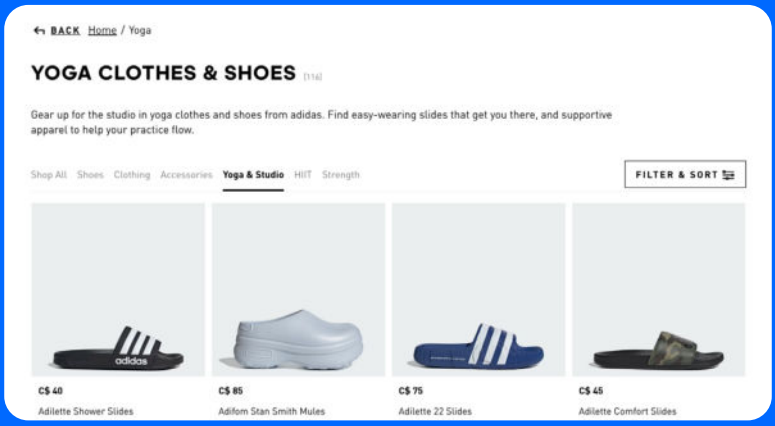
Brand not carried redirection

You may also wish to test strategies on individual or groups of queries (such as brands you don't carry) with a specific strategy, like showing a 'no results' page with explainer text and recommendation pods versus showing similar matches or redirecting to the nearest browse page.

For example, if a famous DTC brand receives many searches for other consumer brands, it may test its 'no results' page against redirects across a bucket of search terms like 'Lululemon,' 'HOKA,' 'On,' and 'Air Jordan.'

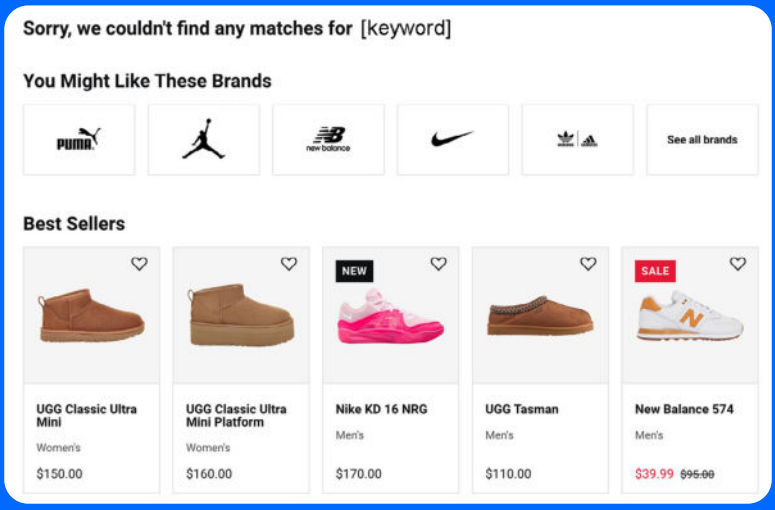


A 'no results' experience without similar-brand recommended products



Instead, search displaying near-match for 'Lululemon' search

Alternatively, another athletics retailer displays brand tiles and best sellers for brands and products they don't carry.



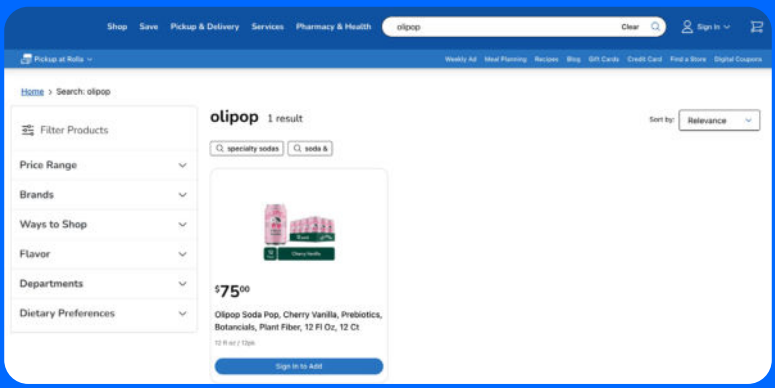
Athletic retailer's 'no results' page

Before attempting a 'no results' A/B test, ensure that you incur enough no results events per month to support a split test within a reasonable time frame.

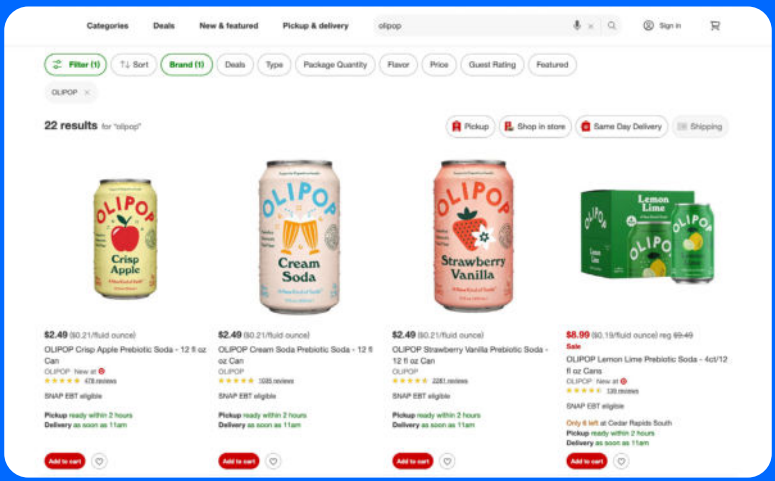
Variant slicing

Variant slicing splits a single product listing into individual product cards for each specified variant (for example, with apparel you may slice color, but not size).

You may have use cases where allowing shoppers to browse variants separately from the product list may be the better user experience, especially when they're associated with frequently applied facets like color or flavor.

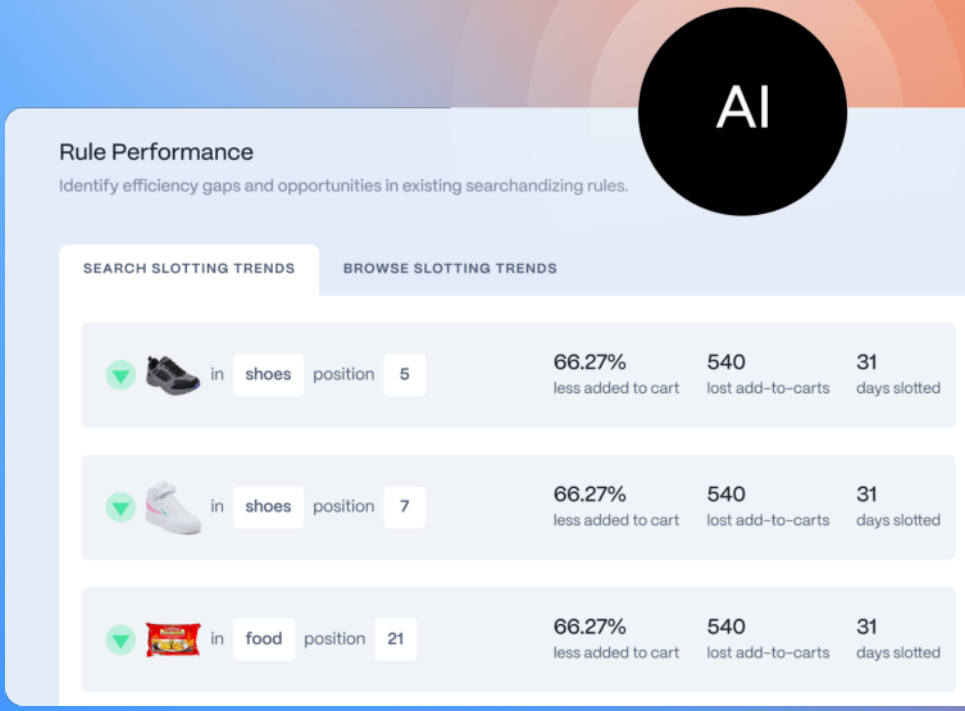


Search results for "olipop" without variant slicing



Search results for 'olipop' with variant slicing

If you spot search queries with high volume and low click-through and conversion, consider testing variant slicing against your default approach. Again, it's important that search volume is sufficient to support an A/B test for a single or group of related queries.

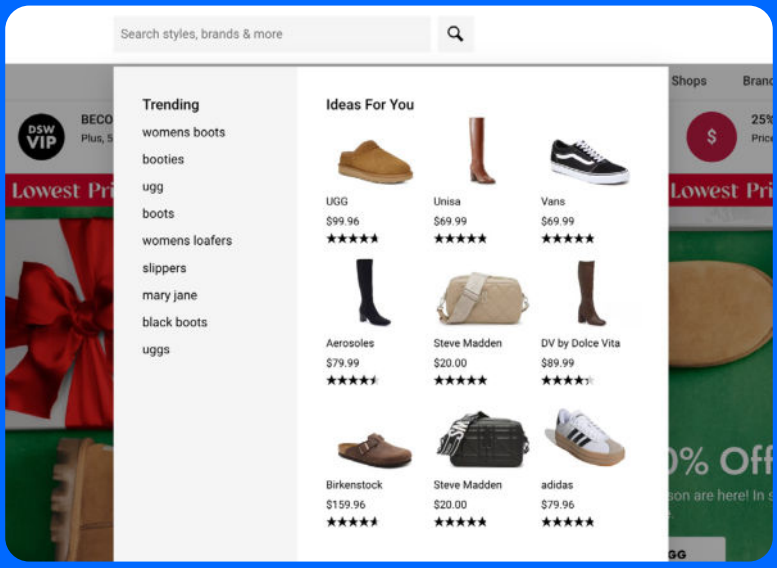


Autocomplete

Pre-merchandising

Many merchants choose to pre-populate their autocomplete widget with trending terms and products to inspire shoppers.

These suggestions disappear once the user starts typing a query and are replaced with search matches.



Autosuggest widget with pre-populated suggestions

While this is a popular practice, it's important to validate whether this tactic is the best pattern for your customers.

Visitors that touch the search box already have a query in mind and a task to complete. They are in 'search' versus 'browse' mode. Showing unrelated suggestions may be distracting enough to interrupt their task and lead to premature abandonment.

An A/B test that measures clicks, conversion, and revenue can validate whether showing suggestions or keeping search clean works better for your business and customers.

Should pre-merchandising win this experiment, consider testing different recipes within your widget UI.

Widget content and UI

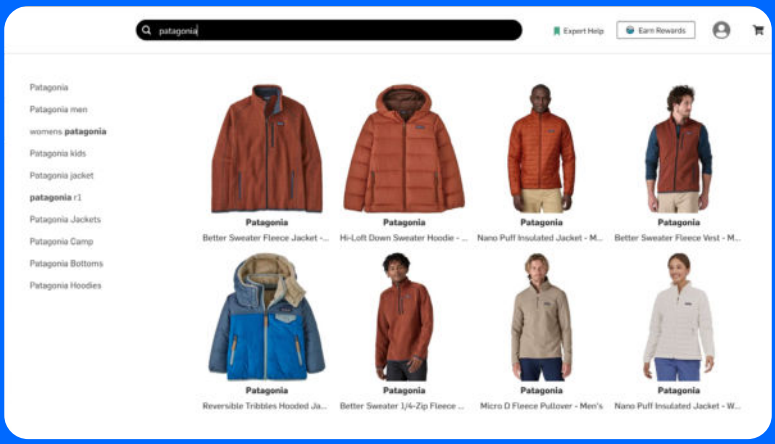
There are infinite ways you can style your autocomplete UI and populate it with suggestions.

Many widgets contain a mix of dimensions (suggested queries, categories, brand, and product hits). Some even include banners and other messages.

It's a great idea to treat your widget like any landing page and test strategies against each other.

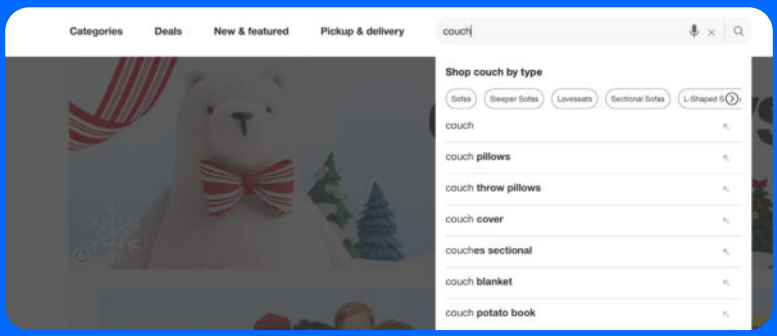
Show / hide product hits

For example, showing product hits provides instant gratification but may be perceived as your full selection. Consider testing a widget with and without product hits (and it's always a good idea to include a 'Show all' link with your product results).



Autocomplete widget with search terms and product hits

A widget without product hits steers shoppers to larger sets of results versus a single product and may enable a richer browse-to-click experience with positive impact on revenue.

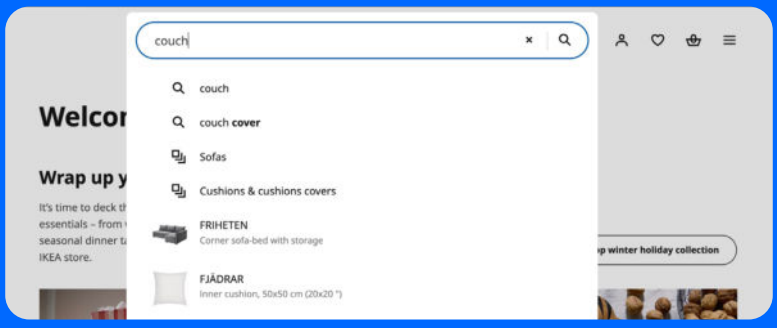


Autocomplete widget without product hits

Number of results per dimension

How many results you display per dimension may also impact user experience. Showing fewer search terms, category matches, or featured products may not provide enough inspiration. On the flip side, a simpler menu is easier to scan with less clutter — especially on mobile.

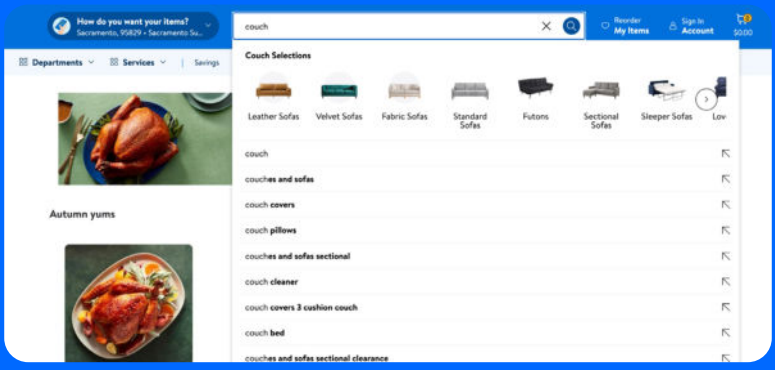
Consider testing different recipes of suggestion settings.



Autocomplete widget with a 2 x 2 x 4 result setting across dimensions

Category thumbnails

You can also test showing image thumbnails with category or brand suggestions to provide visual context.



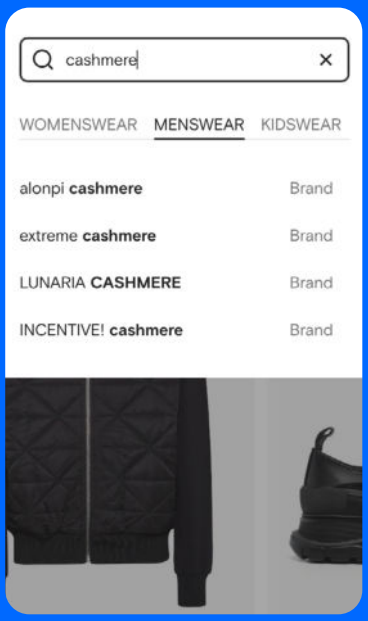
Autocomplete widget with category thumbnail images

Facet scopes

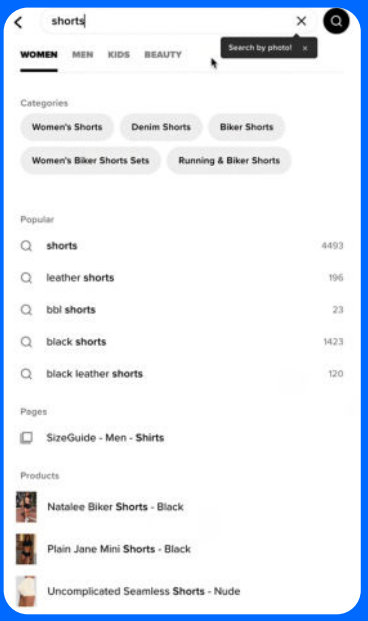
Facet scopes serve as a pre-filter to narrow results before the user submits their query or selects an autosuggestion.

Department scopes

If your catalog spans a few top-level departments such as Men, Women, Kids, and Beauty, giving searchers a way to select a department or facet before submitting their search will return a more focused and relevant results set.



Autocomplete widget with department scopes



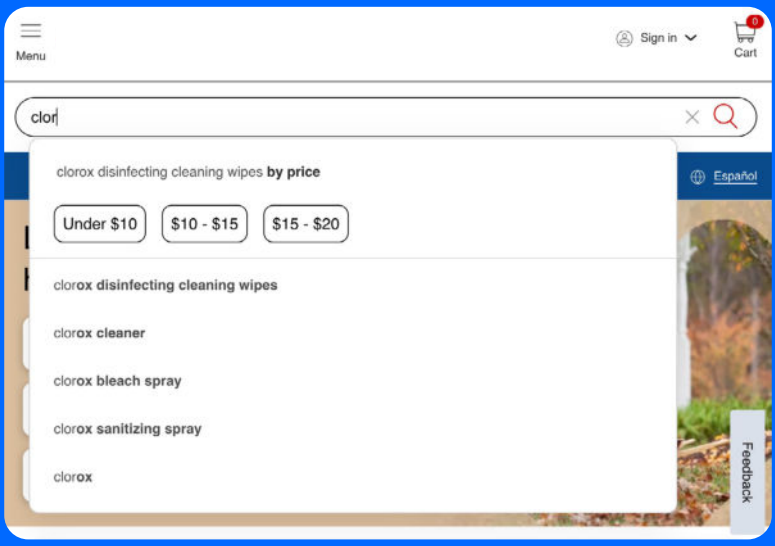
An autocomplete widget with department scoping and a large list of suggestions

It's important to note that as with all tabbed elements, they can get lost in the clutter. In addition to testing with-and-without scopes, you may want to test them in combination with simple versus complex widget population.

Sale and price scopes

You may also try sale and price scopes in your widget. For example, providing a 'Sale only' checkbox or including price filter chips.

Autocomplete widget with Sale only checkbox



Autocomplete widget with price filter chips

Browse & Collections

Searchandising rules

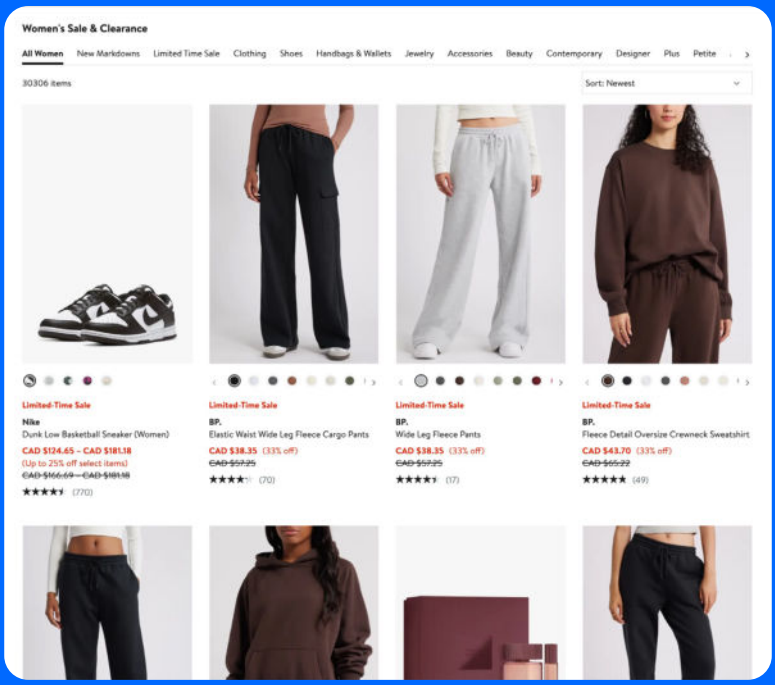
A/B testing use cases for searchandising rules are as broad as your imagination. Just keep in mind that not all rule conditions are significant enough to provide an A/B test lift.

To see the biggest revenue lifts, try testing more complex 'recipes' of search rules against your existing rules, rather than testing rule-by-rule.

Default sort

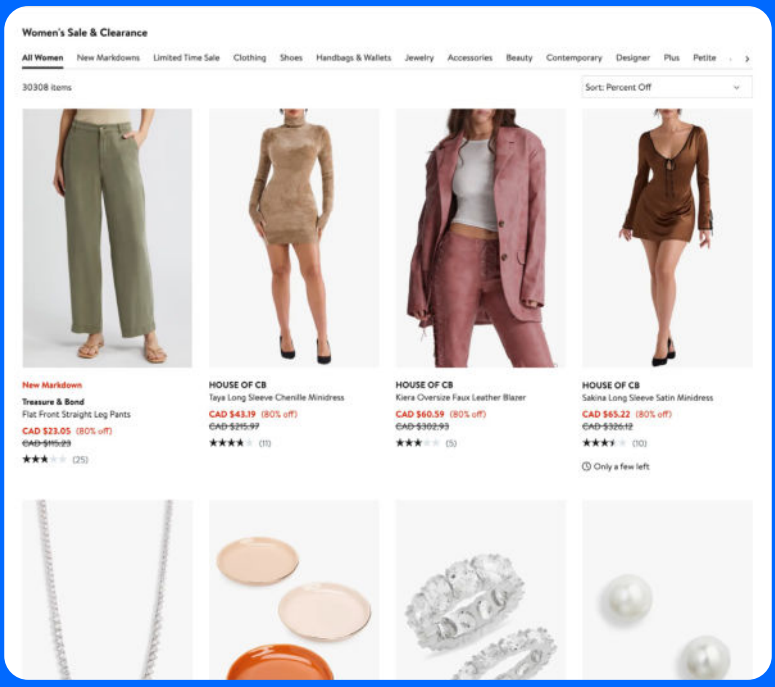
Look for opportunities where a new rule set may outperform your existing rules (for example, Sale and Clearance pages for a fashion catalog).

Is showing newer items first more appealing to your customer base than deeper discounts? Is your goal to turnover aging items more quickly or drive more daily revenue?



Sale collection results ranked by newest

Newer items feel more in-season, while deeper discounts feel like a score.



Sale collection results ranked by percent off

Remember, A/B tests thrive on sample sizes. Choose your testing strategies carefully. Global rules will get you significant results faster than testing a single page unless it consistently receives a lot of traffic per week.

Facets

Filters are powerful tools to help shoppers customize their results and can help them make faster buying decisions. But they're often difficult to use, especially on mobile.

The more granular your facet options, the more complicated this can be.

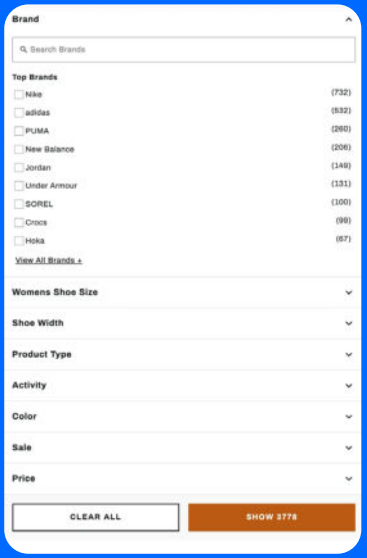
Facet rank

Optimizing facet rank can be achieved through your search and discovery tool when machine learning is enabled to re-rank based on user behavior.

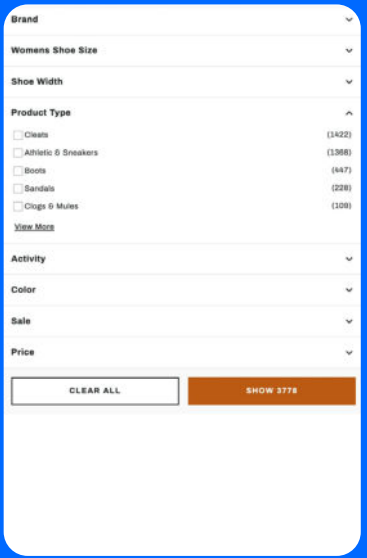
Facet visibility

Also consider which facet values to display open versus closed by default.

A/B testing lets you challenge your existing approach with alternatives. For example, on the menu shown below, only Brand is expanded by default when you open the filter. You may hypothesize that product type would do a better job guiding users to a focused result set than brand.

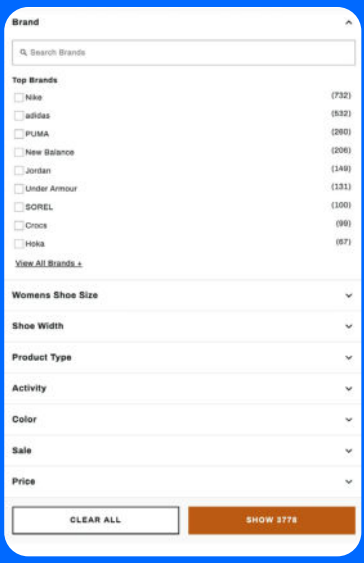


Filter menu with Top Brands facet open by default (Live)

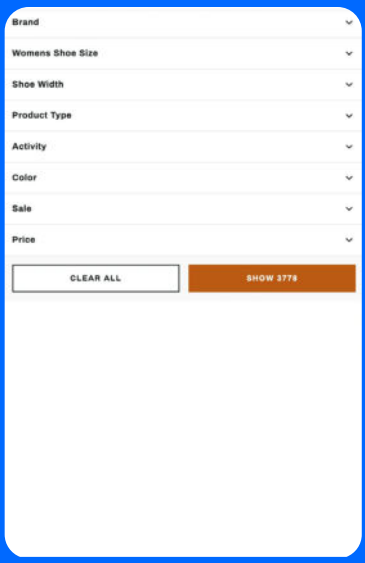


A proposed alternative with Product Type facet open by default

Or, you may hypothesize that a fully collapsed menu reduces visual clutter and streamlines the experience for better usability for most users.



Dick's Sporting Goods filter menu with Top Brands facet open by default (Live)



A proposed alternative with all facets collapsed

Consider running such A/B experiments across clusters of browse pages with similar attributes. For example, top-level department categories return more results than sub-categories and may benefit more from visible Product Type facets.

Besides testing facet visibility, you might also want to A/B test a fixed sort order against dynamic reranking if your search application offers this capability.

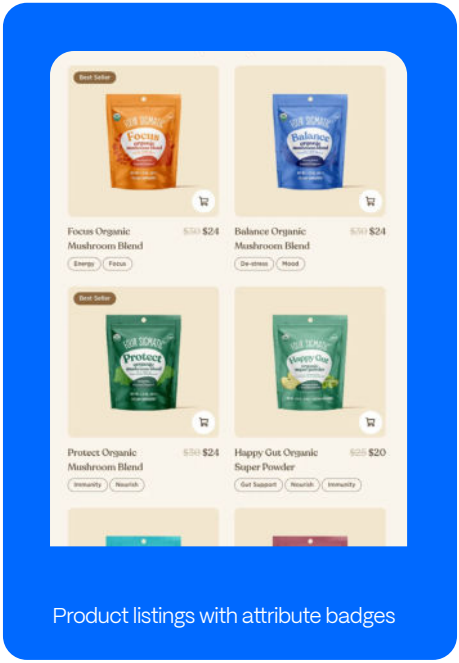
Facet slotting and merging

With Constructor, you can also test slotting (that is, pinning facets to specific spots) and merging facets to reduce complexity (e.g., for colors, size scales, and synonymous tags).

Attribute badges

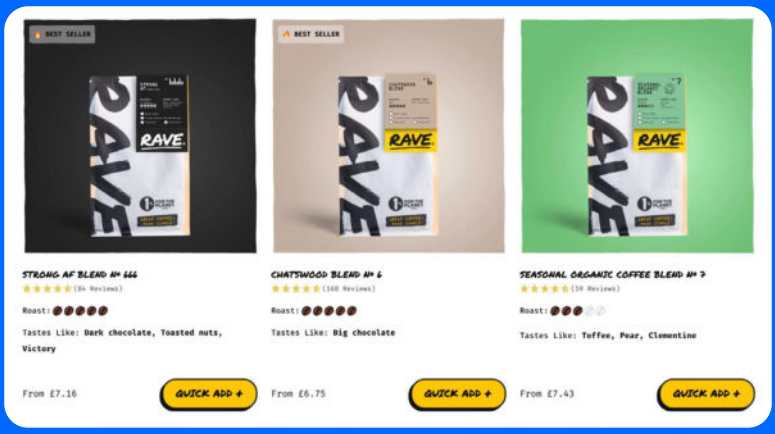
Highlighting useful product attributes within card details can help shoppers better understand their features and benefits without having to pogo-stick back and forth from product list to product pages.

Showing the right product attributes can also help 'pre-sell' shoppers pre-click to increase add to cart rates.



Product listings with attribute badges

This tactic may also underperform due to visual clutter and confusion. Consider A/B testing product listings without attribute detail and testing with different UI treatments.



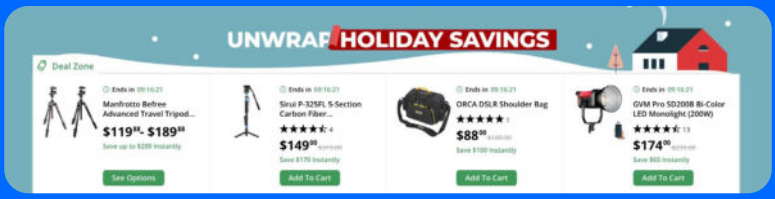
Product listings with attribute details displayed as 'Roast scale' and 'Tastes like'

Recommendations

Homepage recommendations

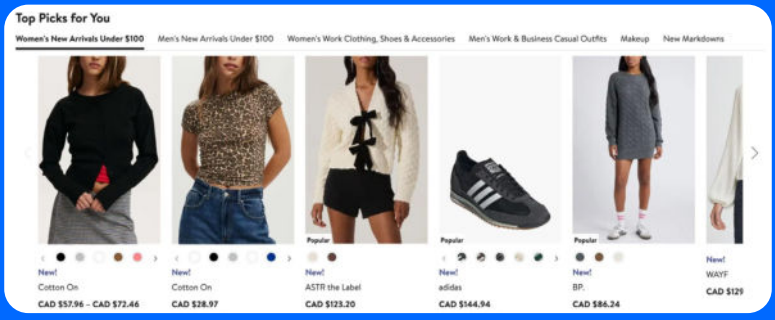
There are many creative ways to embed recommendation pods into homepages. Consider testing both merchandising themes and searchandising strategies within each section.

For example, this retailer's Deal Zone recommendation pod could be manually merchandised or dynamically populated based on rules or personalization. A/B testing can quantify which strategies drive more engagement and revenue.



Homepage 'Deal Zone' recommendation pod

Tabbed recommendation pods are trending on ecommerce homepages. For example, this department store's Top Picks for you lets you browse through several themes like new arrivals under $100, work apparel, makeup, and new markdowns. Consider testing default tabs (first visible) against each other.



Tabbed homepage recommendation pod

Product page recommendations

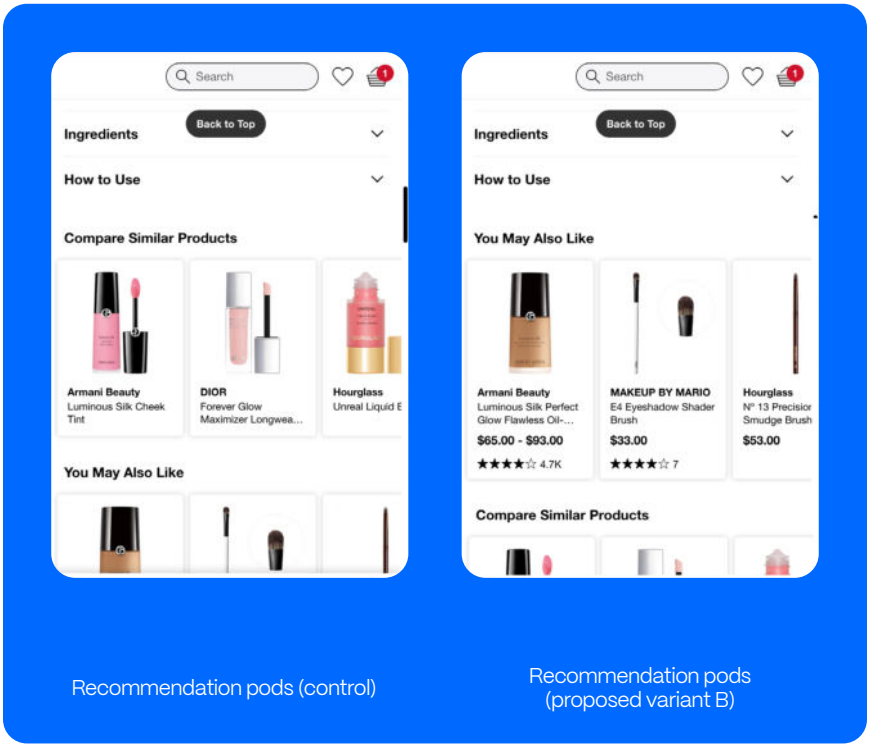
Because product detail pages are typically your most complex templates, it's important to test recommendation pod design and placement as well as your recommendation strategy.

Recommendation pods present a competing call to action to 'add to cart.' You're asking shoppers to make their decision a bit more complicated. But when used effectively, they can boost CVR, RPV, and AOV.

Pod recipes

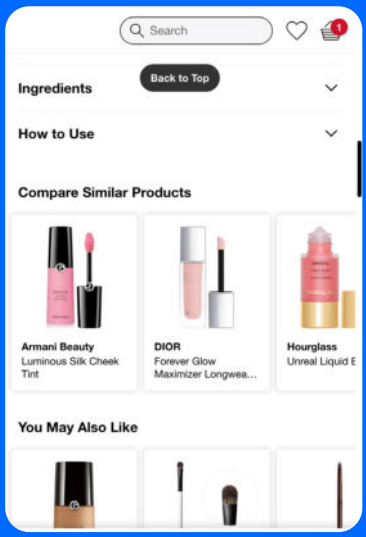
If you already present a combination of recommendation pod types on your PDPs (pod recipes), a great test to start with is their stack rank, or the order they're shown in.

For example, a beauty retailer uses 'You May Also Like,' 'Compare Similar Products,' and 'Use it With' pods. Because not all users will scroll to see all three (especially on mobile), a simple experiment testing pod order (A/B or A/B/C) is a great place to start before testing rules within the pods.

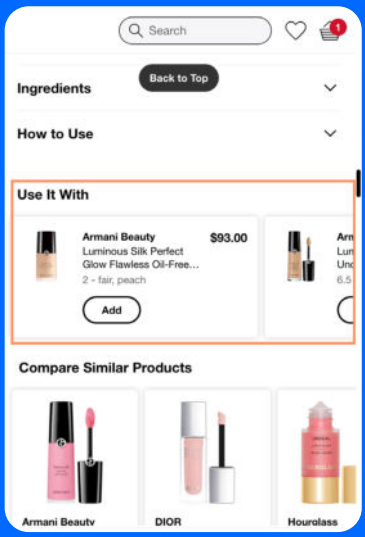


Recommendation pods (control)

Recommendation pods (proposed variant B)



Recommendation pods (control)

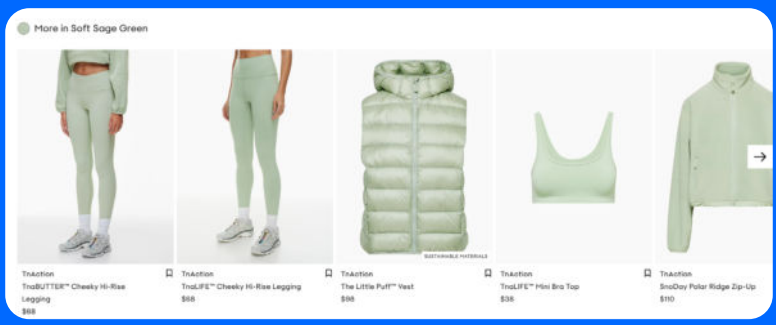


'Use It With' first (proposed variant C)

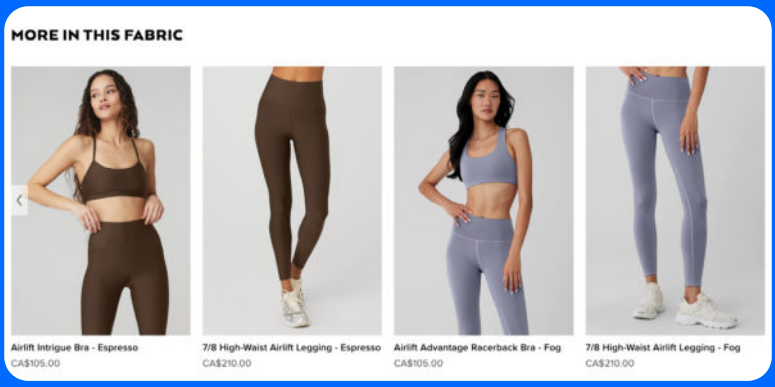
New thematic pods

Beyond the standard 'you might like' and 'more like this' varieties, consider testing attribute-level recommendations, like more in this color, pattern, fabric, scent, or finish.

A/B testing tells you how effective these pods are relative to your standard recommendations.

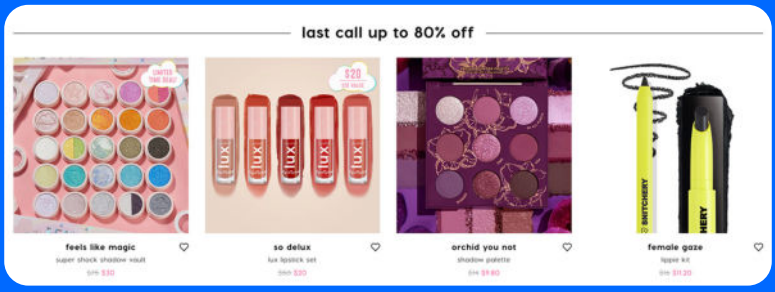


More in this [color]



More in this [fabric]

Or try a pod of deeply discounted recommendations (but remember, revenue and AOV may be lower compared to other pod strategies, even when click-to-conversion is high).

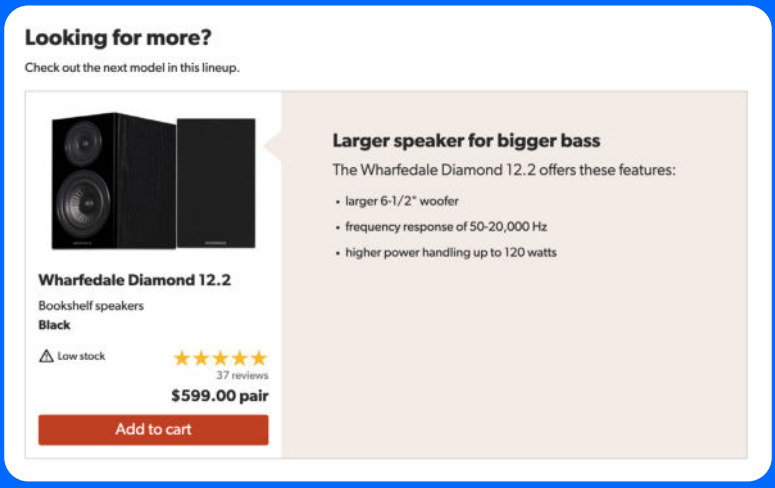


"Last call" recommendation pod

Alternative product

Within Constructor, you can also serve alternative product recommendations to steer shoppers toward a featured item, upsell, or otherwise attractive option, such as an item from a brand the customer has frequently bought from or added to favorites.

For example, this merchant suggests an upsell for the 'next model in this lineup' and includes a short summary of product specs inside the component.



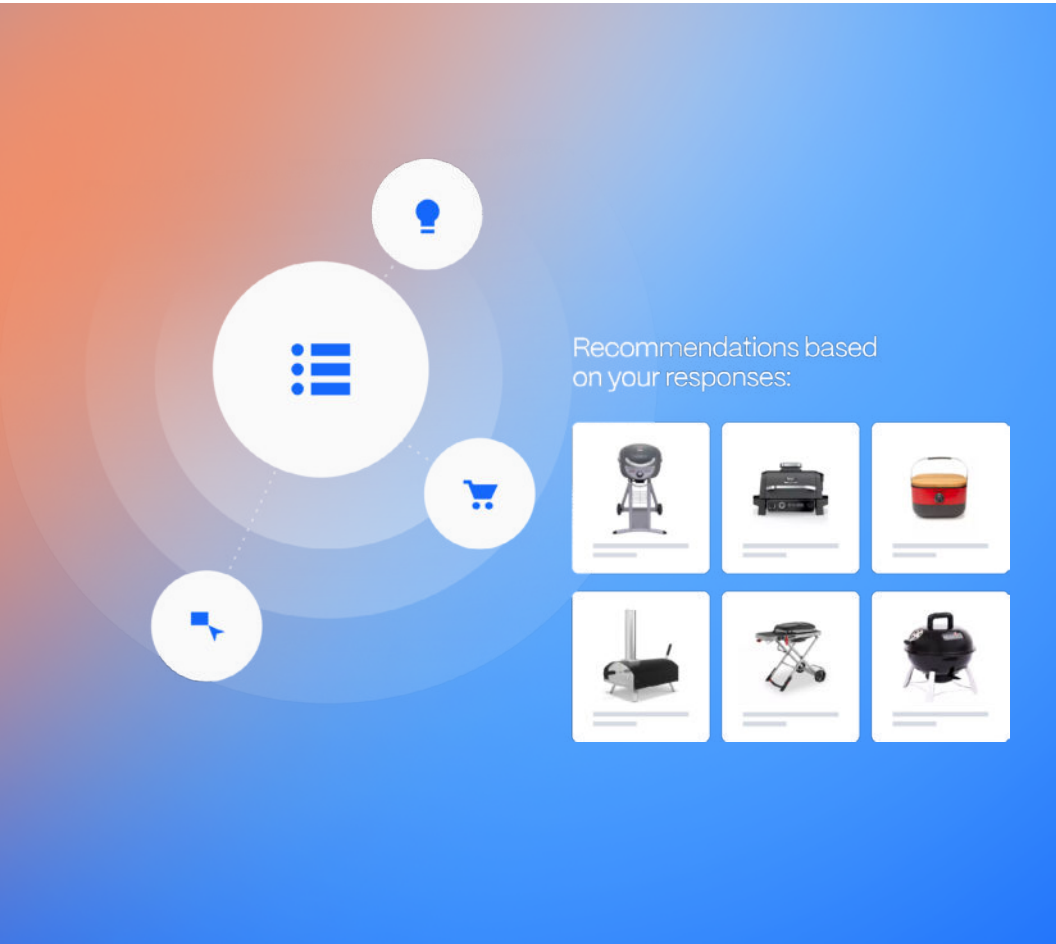
Alternative product upsell

At Constructor, in internal A/B tests, we found that complementary recommendation results lead to significantly more users clicking and eventually adding to cart compared to alternative items.

While more people clicked on alternative items, complementary recommendations drove conversions and larger basket sizes.

The alternative strategy (upselling) drove +8.1% in users clicking recommendations.

The complementary strategy drove +11.6% in users clicking on a recommendation followed by an add-to-cart and +13.6% in clicks followed by a purchase.



Cart cross-sells

As with all recommendation pods, your A/B testing use cases are virtually unlimited. But showing recommendations in the cart is a hot topic in itself. Many conversion optimization professionals argue that you should keep distractions out of the cart to increase proceed to checkout actions.

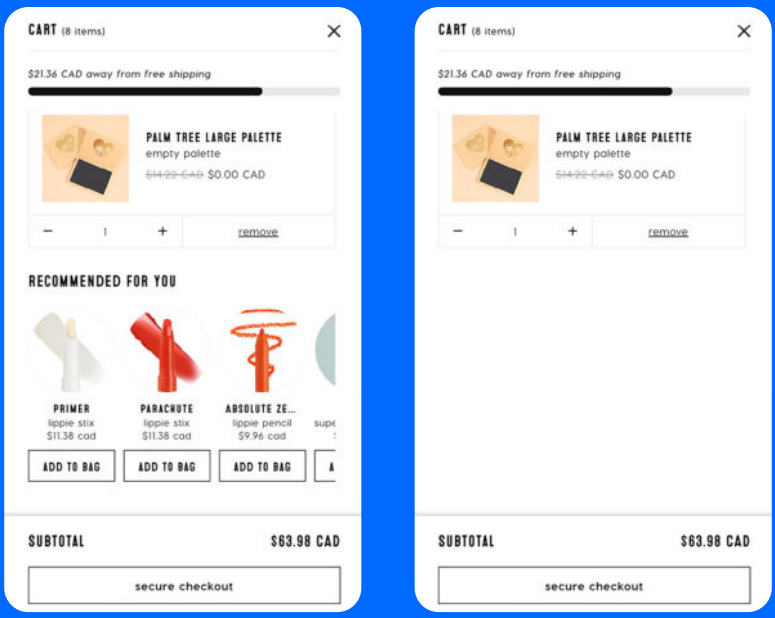
Others argue that cross-selling is a great tactic to increase AOV and revenue, especially on mobile where continuing your shopping journey through your navigation menu carries more friction.

It's not just a matter of showing cross-sells or not. It also depends on how your cart page (or mini cart drawer) is designed, your merchandising strategy, and how cross-sells display once a cart is full of items.

Global show-hide test

The simplest experiment (and the one you should start with) is a show/hide A/B test.

This enables you to gauge how receptive your shoppers are to cross-sells in general.



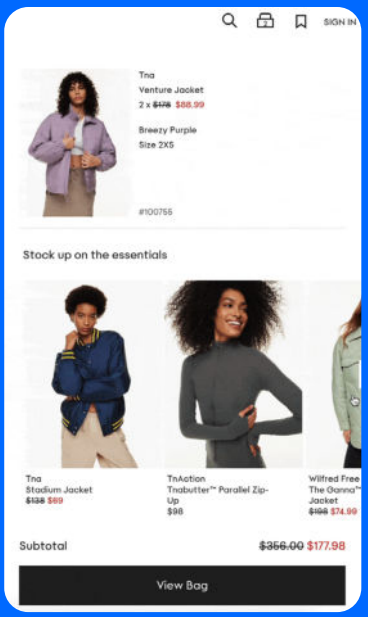
An example cart with and without cross-sells

A losing test doesn't mean you need to scrap the idea of cross-selling in your cart altogether. Consider re-testing with different merchandising strategies or UI patterns to see if you can close the gap.

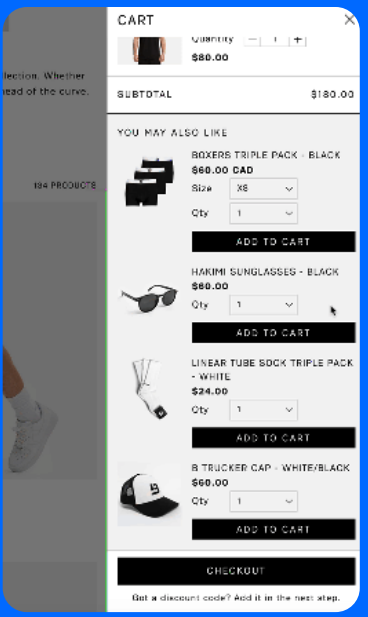
Once you have a winning recipe, you can more confidently proceed with keeping cart cross-sells live and experiment with more granular rules.

Recommendation orientation

Testing a horizontal scroller against a vertical pattern can help you determine what has the best impact on revenue.



A cart with horizontal cross-sells

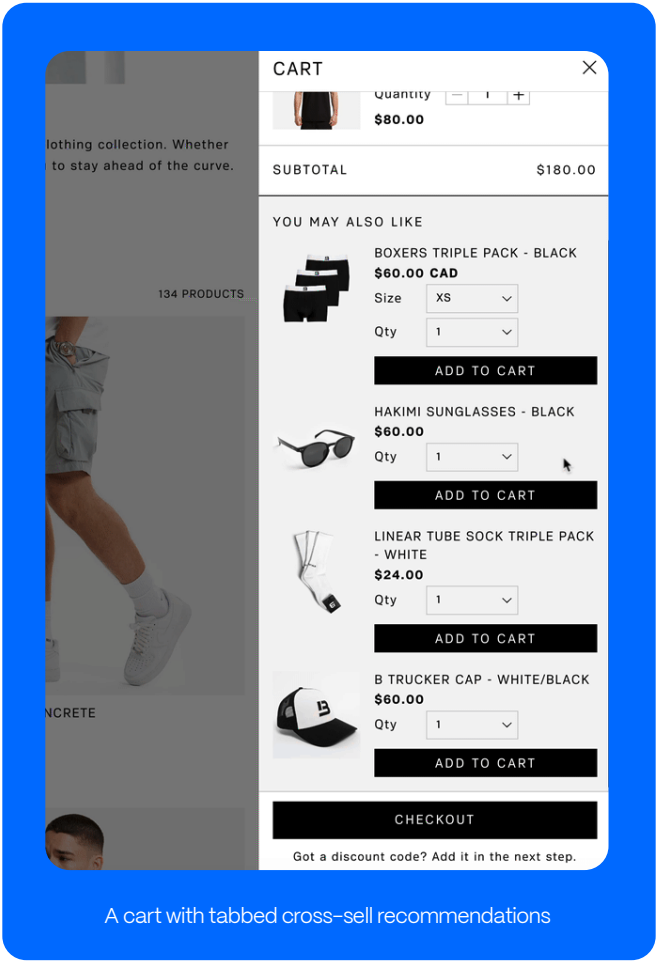


A cart with vertical cross-sells

Tabbed cross-sells

Merchandising the mini-cart is challenging with limited real estate on both mobile and desktop.

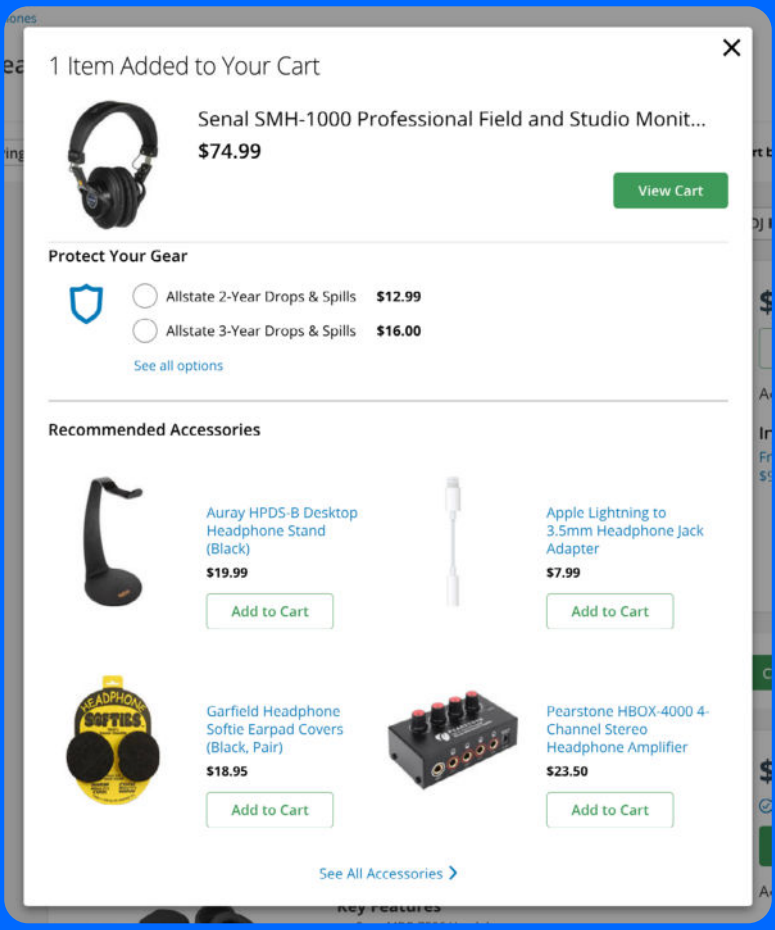
Tabbed recommendations enable you to pack more into the cart and give users some choice. You can also set up your experiment to switch the default visible tab from category A to B, for example.



A cart with tabbed cross-sell recommendations

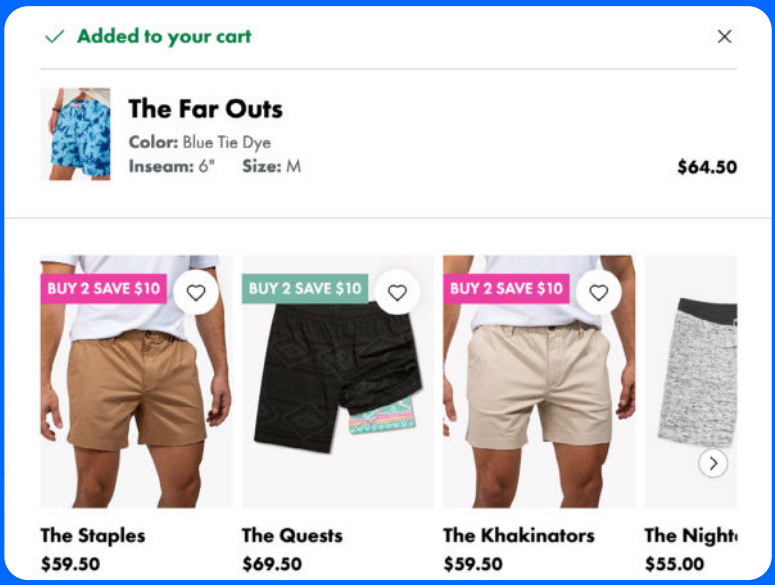
Add-to-cart interstitials

Modals give you more room to work with on both desktop and mobile. Consider testing them against traditional cart cross-sells.

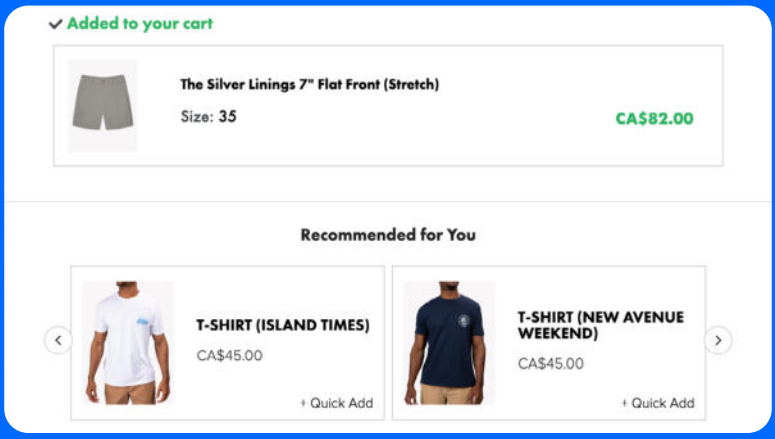


Add-to-cart confirmation modal populated with accessories related to the product added

You can also A/B test designs against each other. For example, this apparel merchant has redesigned their add-to-cart modal, which may involve A/B testing in the revision process.



Add-to-cart confirmation modal populated with complementary products



Add-to-cart confirmation modal with redesigned UI

Wrapping Up

Incorporating disciplined A/B testing into your search and discovery efforts is a powerful, often underutilized weapon in your revenue-optimizing arsenal.

Incorporating disciplined A/B testing into your search and discovery efforts is a powerful, often underutilized weapon in your revenue-optimizing arsenal. Experimentation across search, browse, autocomplete, and recommendations enables you to validate your hypotheses, de-risk your UI and merchandising decisions, learn more about your customers, and enhance overall user experience.

If you'd like to follow more insights on search and discovery A/B testing, check out the Constructor Experiments Blog and subscribe to receive notifications when we publish new test results.

And if you want to try out any of the concepts covered in this book, **we've created a companion set of templates and cheatsheets to make launching your A/B testing practice easier.**