

CAN SYNTHETIC TESTERS SHOP LIKE REAL PEOPLE?

Human vs. Synthetic Testers on Amazon: A Dyson vs. Shark Case Study

Lihong Hicken

TheySaid Research

Case study · October 2026

ABSTRACT

Can Synthetic Testers reproduce how real people make a purchase decision, not just what they choose? We ran a matched-pair study in which 25 US shoppers and 25 Synthetic Testers completed the same live shopping task on Amazon.com: research a Dyson and a Shark cordless vacuum priced at \$300 or less, compare them, and choose one to buy. Both groups worked with the same AI moderator. Each Synthetic Tester was matched to one human participant and built from Motivation, Worldview and Knowledge inputs that TheySaid's AI Research Assistant derived from the recorded human sessions, rather than from invented personas. At the individual level, Synthetic Testers chose the same brand as their human counterpart in 84% of pairs (21 of 25), reproduced 76% of their counterpart's top purchase factors, and reproduced about 72% of their counterpart's shopping behavior. At the group level, each study produced 10 major findings; Synthetic Testers reproduced 50% of the human findings directly and 70% when partial matches are included. Both studies selected the same winning brand (Shark: 56% of humans, 64% of Synthetic Testers, a difference of two shoppers), reported almost identical purchase confidence, ranked price and value as the top purchase driver, and pointed to the same four business recommendations. The largest difference was research style: Synthetic Testers compared more products (often 2–4 per brand versus 1–2), read reviews more deeply and gathered evidence more consistently, while humans brought emotion, taste, shortcuts and lived experience into the decision. Synthetic Testers also surfaced an additional hypothesis that humans did not: shoppers want concrete product proof, such as battery runtime by power mode, suction figures and warranty length. We conclude that Synthetic Testers grounded in real human behavior reproduce the overall direction of a consumer research study, and that they deliver the most value when run alongside human research.

Keywords Synthetic Testers · AI testers · synthetic users · user testing · think-aloud research · purchase decisions · e-commerce · matched-pair study · research fidelity

1 Introduction

Research teams increasingly want to test ideas faster than they can recruit people. Recruiting the right human participants takes time and budget, and many product decisions get made with little or no research at all. Synthetic testers, AI participants that complete research tasks the way a real user would, offer a way to close that gap. But they are only useful if what they produce resembles what real people would have done.

Most evaluations of synthetic participants ask a narrow question: do they give the same answers? This study asks a broader one. We wanted to know whether Synthetic Testers could reproduce not only *which* product people choose, but *how* they research, what information matters to them, and how they arrive at a final decision.

To test this, we chose a realistic, high-consideration consumer purchase: a cordless vacuum from one of two well-known brands, bought on a live website with a fixed budget. We ran the task first with real people, built

Synthetic Testers from those sessions, ran the identical task again with the Synthetic Testers, and compared the two at two levels: one shopper at a time, and one study against the other.

This report makes four contributions:

1. A **matched-pair design** in which every Synthetic Tester corresponds to a specific human participant, allowing individual-level comparison.
2. A **behavior-grounded persona workflow**: Synthetic Testers were built from what human participants actually said and did on video, not from predefined segments.
3. A **two-level fidelity analysis** covering final choice, purchase factors, shopping behavior, major findings and business conclusions.
4. A practical account of **where Synthetic Testers match humans, where they differ, and what each contributes**.

2 Background

Earlier work on synthetic participants has shown that AI models can reproduce aggregate patterns in human responses, replicate classic behavioral experiments with directional agreement, and match individual people more closely when grounded in real data about those people.

This study extends that idea in three practical directions. First, the grounding source is *video of real behavior*, including what participants clicked, compared and said aloud, rather than survey answers or interview text alone. Second, the task takes place on a *live commercial website*, with real listings, prices, reviews and friction. Third, the evaluation goes beyond answer matching to compare research behavior, findings and the business decisions each study would support.

3 Study Design

3.1 Objective

The study tests how closely Synthetic Testers can simulate real human shopping behavior in a consumer purchase decision. Both groups completed the same Amazon shopping task with the same AI moderator. Their browsing behavior, reasoning, product comparisons and final purchase decisions were recorded and compared.

3.2 Participants

The human study used a broad US consumer audience of 25 participants. To qualify, participants had to be US residents aged 25 or older, use a vacuum at home, be willing to consider purchasing a cordless vacuum, and be involved in household purchase decisions. No household-income requirement was used. Household and product-background information was collected at the end of the session for analysis only and did not disqualify anyone. The synthetic study used 25 Synthetic Testers, each matched to one human participant (Section 4).

3.3 Shopping scenario and task

Participants imagined they were buying a new cordless vacuum for their home. They had a maximum budget of \$300, had narrowed their consideration to Dyson and Shark, and used Amazon.com to research available

products. They were required to evaluate one serious option from each brand before deciding. Table 1 summarizes the setup.

Table 1: Study setup.

Element	Specification
Human participants	25 US residents, age 25+, vacuum users involved in household purchases
Synthetic Testers	25, each matched to one human participant
Environment	Live Amazon.com, on the open web
Budget	\$300 maximum
Brands	Dyson or Shark, any cordless model within budget
Moderation	The same AI moderator and instructions for both groups
Outcome	Final choice left in cart with reasons; no order placed

The task had three parts, completed while thinking aloud:

1. **Find a Dyson option.** Search Amazon for Dyson cordless vacuums at \$300 or less, open the one you would seriously consider, review the product details and customer reviews that matter most to you, and add it to the cart.
2. **Find a Shark option.** Do the same for Shark (for example, the Shark PowerPro Flex Reveal Plus IZ382H or IZ388H, though any Shark cordless model within budget was allowed), and explain how it compares with the Dyson.
3. **Make the final decision.** Review both vacuums in the cart, compare them on the factors that matter most, decide which one you would actually buy, remove the other, and explain why the winner won.

3.4 Post-task questionnaire

After the task, all participants answered the same structured questions: brand chosen, specific model chosen, the top five factors in their decision (ranked from a list of 13, including price, suction and cleaning performance, battery life, pet-hair performance, weight, attachments, customer ratings, customer reviews, brand reputation, warranty, ease of maintenance and design), why they chose that vacuum, and how confident they were. Background questions covered current vacuum type and brand, familiarity with Dyson and Shark, pets, flooring, home type and prior brand preference. The full questionnaire appears in Appendix A.

3.5 Technical pilot

Before the main studies, we ran one human and one synthetic technical pilot to confirm two things: that a Synthetic Tester could complete the task on Amazon end to end, and that Amazon would not block it. The pilot Synthetic Tester completed the full task in 34 minutes without being blocked and without needing the alternative recording option. The pilot persona appears in Appendix B.

4 Building Synthetic Testers from Human Behavior

A central design choice in this study is that the Synthetic Testers were not written by hand. They were derived from the human sessions (Figure 1).

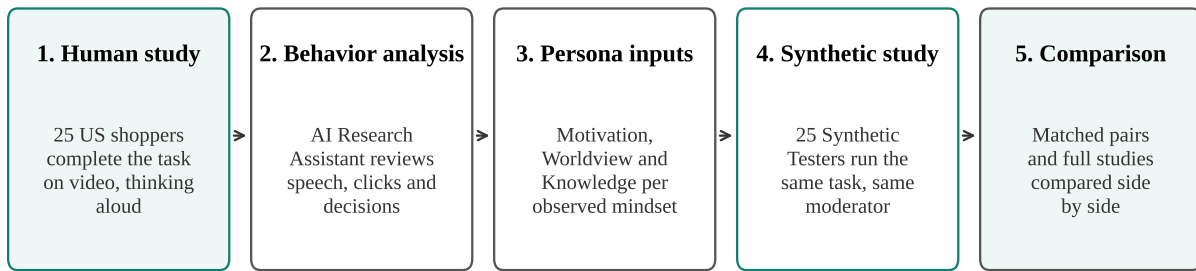


Figure 1: Study workflow. Human behavior is analyzed first; the resulting persona inputs drive Synthetic Testers that complete the identical task.

4.1 From recorded sessions to decision mindsets

After the 25 human sessions were complete, TheySaid’s AI Research Assistant reviewed the recordings, including what participants said, what they clicked, how they researched, which products they considered, and what ultimately drove their decisions. Its goal was to identify recurring decision mindsets across the sample. These mindsets were not predefined; they emerged from observed behavior and feedback.

4.2 Three building blocks

For each mindset, the AI Research Assistant generated the three inputs used to create a Synthetic Tester:

- **Motivation:** what the buyer is trying to accomplish, what matters most to them, and the mindset they bring to the decision.
- **Worldview:** the attitudes, preferences, biases, trust levels and beliefs that shape how they interpret products and brands.
- **Knowledge:** what they already know about the category, products and brands before and during the task.

The 25 Synthetic Testers were then distributed across these personas, matched to their human counterparts, and given the same task with the same AI moderator. The workflow is therefore:

Humans behave → AI analyzes behavior → AI identifies mindsets → AI writes Motivation, Worldview and Knowledge → Synthetic Testers run the same task

5 Analysis Framework

We compared the two groups at two levels, using the AI Research Assistant to produce side-by-side comparisons with supporting evidence from each session.

Individual fidelity asks whether a Synthetic Tester reproduces the specific person it was built from. For each of the 25 matched pairs we compared the final brand choice, the overlap in top purchase factors, and the similarity of shopping behavior (what was searched, opened, compared and read).

Group-level fidelity asks whether the synthetic study reaches the same conclusions as the human study. We compared the final brand split, purchase confidence, top purchase criteria, the major findings each study produced (counting findings reproduced in full and in part), differences in research behavior, findings unique to each group, and the business recommendations each study would support.

6 Results

6.1 Individual fidelity

Synthetic Testers matched their human counterpart's final brand choice in 84% of pairs (21 of 25). They reproduced 76% of their counterpart's top purchase factors and about 72% of their counterpart's shopping behavior (Table 2, Figure 2).

Table 2: Individual fidelity across 25 matched human–Synthetic Tester pairs.

Measure	Result	Basis
Same final brand choice	84%	21 of 25 pairs
Top purchase factors reproduced	76%	Overlap with counterpart's top factors
Shopping behavior reproduced	≈72%	Search, product views, comparisons, reviews

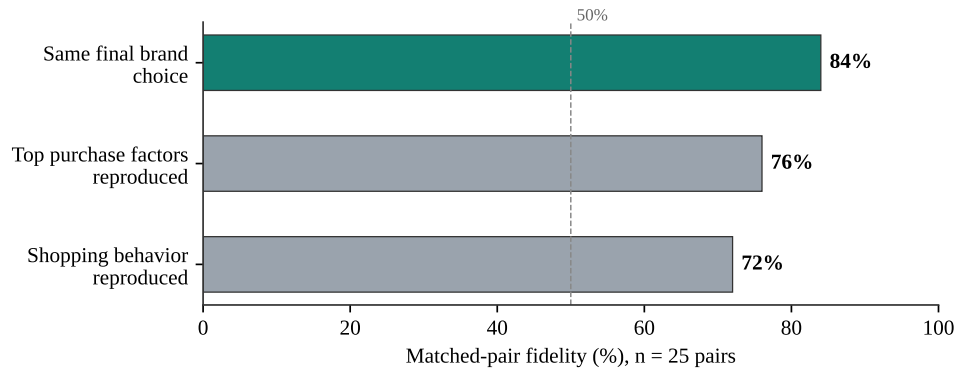


Figure 2: Matched-pair fidelity. Brand choice shows the strongest agreement.

Figure 3 shows every pair. Thirteen pairs both chose Shark and eight both chose Dyson. Of the four pairs that disagreed, three involved a human choosing Dyson and their Synthetic Tester choosing Shark, and one the reverse.



Figure 3: Final brand choice for all 25 matched pairs, sorted by outcome. Circles are human shoppers; squares are their Synthetic Testers. Dashed boxes mark the four pairs that disagreed.

6.2 Group outcome

Both studies chose the same winning brand. Shark was selected by 56% of human shoppers and 64% of Synthetic Testers, a difference of only two shoppers out of 25 (Figure 4). Average purchase confidence was almost identical between the two groups.

Read together, the individual and group results tell a consistent story. Individually, 84% of Synthetic Testers

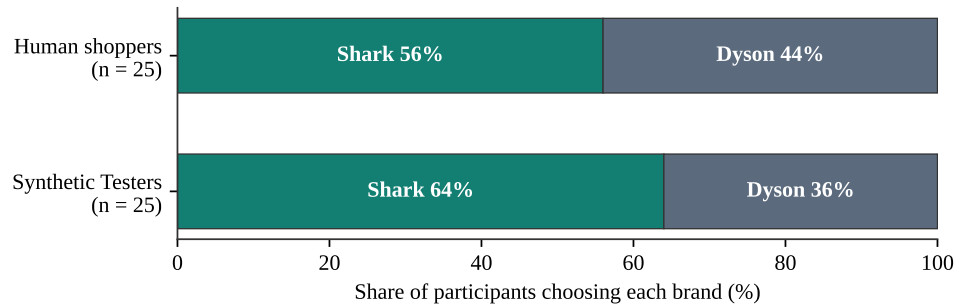


Figure 4: Final brand split by group. Each participant represents four percentage points.

chose the same brand as the person they were built from. Collectively, the synthetic study produced a similar overall purchase result.

6.3 Major findings

We identified 10 major findings in the human study and 10 in the synthetic study. Synthetic Testers reproduced 50% of the human findings directly and 70% when partial matches are included (Figure 5).

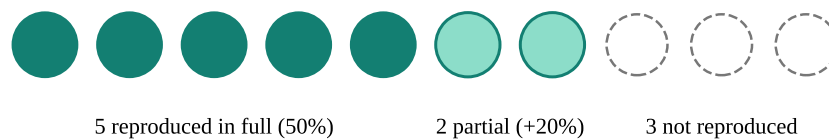


Figure 5: Coverage of the 10 major human findings by the synthetic study.

Both groups independently surfaced the same major themes:

- **Price and value** drive the decision.
- **Customer reviews** build or undermine confidence.
- **Pet-hair performance** is a deciding feature for many buyers.
- **Brand trust** shapes how options are judged.
- **Negative reviews** can change the final decision.

Not every finding was identical, but the major story of the research was the same.

6.4 Purchase priorities

Price and value were the number one purchase factor for both groups, and reviews and pet-hair performance were important to both. The two groups agreed on what mattered most but weighted some factors differently (Table 3).

Table 3: Shared priorities and differences in weighting.

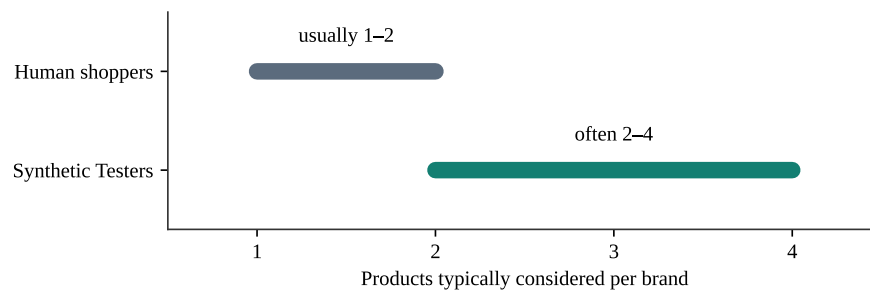
Purchase factor pattern	Humans	Synthetic Testers
Price and value as the #1 factor	✓	✓
Reviews as a major factor	✓	✓
Pet-hair performance as a major factor	✓	✓
Greater weight on suction and cleaning performance	✓	
Greater weight on pet-hair features		✓
Heavy reliance on the number of reviews		✓
Design as a main reason for choosing	✓	

7 Research Behavior

The clearest difference between the groups was not what they valued but how they shopped. Humans ranged from deep researchers to quick, “good enough” shoppers. Synthetic Testers were consistently thorough and methodical (Table 4, Figure 6).

Table 4: Shopping behavior by group.

Behavior	Human shoppers	Synthetic Testers
Products considered per brand	Usually 1–2	Often 2–4
Use of reviews	Many relied on the star rating	Read review summaries and rating breakdowns
Stopping point	Often stopped at the first acceptable option	Rarely stopped early
Changed brand preference	12%	32%
Evidence gathered	Uneven, from a glance to a deep dive	Consistently extensive
Typical session length	About 15 minutes	About 30 minutes

**Figure 6:** Breadth of comparison. Synthetic Testers typically evaluated more products per brand.

In everyday terms, Synthetic Testers shop like very diligent shoppers. Humans are messier: some research deeply, while others skim, get distracted or stop as soon as something looks good enough. This is not a weakness of either group. It tells research teams what kind of behavior Synthetic Testers simulate particularly well (deliberate, evidence-driven evaluation) and where human participants remain different (shortcuts, impulses and personal history). The higher rate of brand-preference change among Synthetic Testers is consistent with this pattern: gathering more evidence creates more opportunities to revise an initial view.

8 Complementary Insights

Each group surfaced findings the other did not.

Human-only perspectives. Humans brought more real-life feeling, habit and experience into the decision. They were more likely to carry strong opinions about Dyson or Shark from past ownership, refuse to sign in or distrust Prime-only pricing, stop at the first good-enough option, care about looks and color, trust video reviews over Amazon’s AI review summary, and feel nervous, frustrated or uncertain about the purchase.

Synthetic-only perspectives. The strongest Synthetic-only insight was that shoppers may want more specific product proof. Synthetic Testers repeatedly asked for battery runtime by power mode, suction numbers and warranty length, and were skeptical of vague marketing claims. Because Synthetic Testers research deeply, they can surface useful additional hypotheses like this one. As with any single-source finding, Synthetic-only findings are best treated as hypotheses to validate with human participants.

9 Business Implications

A research study is ultimately judged by the decisions it supports. At the big-picture level, both studies would give a vacuum brand the same direction (Table 5).

Table 5: Recommendations supported by each study.

Recommendation	Human study	Synthetic study
Focus on pet-hair and anti-tangle benefits	✓	✓
Make value comparisons easy	✓	✓
Strengthen review proof	✓	✓
Fix confusing product listings	✓	✓
Push harder on Prime pricing, design, brand loyalty and quick “good enough” buyers	✓	
Push harder on spec details and review evidence		✓

Both studies also reached the same overall winner. Synthetic Testers got the overall direction right, while humans added details that could change where a company spends its money. The strongest research approach is therefore not human *or* synthetic, but human *and* synthetic.

10 Discussion

10.1 What fidelity means in practice

This study offers two complementary ways to describe fidelity. The first asks whether a Synthetic Tester can reproduce an individual shopper: 84% of Synthetic Testers chose the same brand as their human counterpart, with 76% overlap in top purchase factors. The second asks whether a group of Synthetic Testers can reproduce the conclusions of an entire study: the synthetic study recovered 70% of the major human findings, chose the same winner and supported the same core recommendations.

10.2 Why results were not identical

Two factors explain most of the differences. First, Synthetic Testers do not behave exactly like people. They research more deeply, read more reviews and give more weight to signals such as review count, while humans are more likely to decide quickly or respond to softer cues such as appearance. Second, the two groups did not always see the same store. Amazon personalizes the shopping experience, and participants were sometimes shown different products, listings, prices or purchase restrictions. Two shoppers using similar decision logic can therefore see different options and reach different outcomes.

10.3 Why video grounding matters

User research is not only about what people say. Text captures what participants said; images capture what they saw; video captures what they saw, what they did, what they said and why. Video also helps separate participants who are simply reading a review aloud from those reacting to it and explaining why it matters. These can look nearly identical in a transcript but mean very different things to a researcher. Building Synthetic Testers from video sessions lets the persona inputs reflect behavior, not just stated opinions.

10.4 Scope

This is a single-category study of 25 matched pairs, designed to show directional agreement rather than statistical equivalence. Amazon listings, prices and offers were not held constant across sessions. A temporary Amazon outage affected two sessions, and those pairs showed lower agreement. Some personas were generated from transcripts and survey answers alone when analysis features were temporarily unavailable. Because factor rankings informed persona inputs, we place more weight on participants' open-ended reasoning and observed behavior than on the ranking question alone. Speed and cost were not formally measured. Additional studies in other consumer categories will test how these results generalize.

11 Conclusion

Synthetic Testers built from real human behavior reproduced the overall story of a consumer research study. They chose the same brand as their human counterpart in 84% of pairs, reproduced 76% of top purchase factors, recovered 70% of the major human findings, selected the same winning brand and supported the same four business recommendations. Their main difference from humans was research style: they behaved like consistently diligent shoppers, and that diligence surfaced an additional hypothesis about shoppers' demand for hard product proof. Humans, in turn, contributed emotion, taste and lived experience. Used together, the two give research teams both the big picture and the real-world nuance.

A Post-Task Questionnaire

Purchase decision

1. Which brand did you choose? (Dyson / Shark)
2. Which specific vacuum did you choose? (open text)
3. Select and rank the top five factors that mattered most in your final decision: price; suction and cleaning performance; battery life and runtime; pet-hair performance; weight and ease of handling; attachments and features; customer ratings; customer reviews; brand reputation and trust; warranty and reliability; ease of maintenance, emptying and cleaning; design and appearance; other.
4. Why did you ultimately choose this vacuum? (open text)
5. How confident are you in your choice? (very confident to very uncertain, 5-point scale)

Background (for segmentation and analysis only)

6. What type of vacuum do you currently use most often?
7. What brand is your primary vacuum?
8. Before this study, how familiar were you with Dyson cordless vacuums?
9. Before this study, how familiar were you with Shark cordless vacuums?
10. Do you have pets in your home?
11. What type of flooring do you clean most often?
12. Which best describes your home?
13. Before today, did you already have a preference between Dyson and Shark?

B Pilot Synthetic Tester Persona

The technical pilot used a broad, neutral US household shopper persona.

Motivation. Needs to buy a cordless vacuum for their home with a budget of up to \$300. Has narrowed the search to Dyson and Shark and wants to find the option that best fits their everyday cleaning needs. Wants to compare available models on Amazon, review product details and customer feedback, and make a confident purchase decision.

Worldview. A typical US household shopper who is practical, open-minded and not automatically loyal to Dyson or Shark. Cares about getting good value for the money, but is also influenced by convenience, product quality, ease of use, brand reputation and customer experiences. Willing to pay more within the budget if the benefits feel worthwhile.

Knowledge. Has general consumer-level knowledge of vacuums but is not a vacuum expert. Understands common buying factors such as price, suction and cleaning performance, battery life, weight, attachments, pet-hair performance, maintenance, reliability, ratings and customer reviews. Is familiar with Dyson and Shark as major vacuum brands but has no strong assumed preference for either.

C Reproducing This Workflow Without a Research Platform

Teams sometimes ask whether this study could be rebuilt with a general-purpose AI model. It can, but because such models do not currently take native video as input, a video-based study requires substantial supporting work. A do-it-yourself version involves at least the following steps:

1. Export the 25 human session videos.
2. Transcribe every session.
3. Extract screenshots or key frames so the model can see what participants saw.

4. Reconstruct the relationship between transcript, screen content, clicks, navigation and reactions.
5. Structure each participant into an analyzable record.
6. Infer each participant's motivations, worldview, knowledge, preferences, biases and decision rules.
7. Review and correct the generated personas.
8. Configure a browser or computer-use environment.
9. Write prompts that keep each AI participant consistent with its persona.
10. Recreate the research task and moderation instructions.
11. Run the AI participants through the website.
12. Capture what they search, click, compare, say and choose.
13. Handle failed sessions, navigation errors, blocked pages and inconsistent agent behavior.
14. Match each AI session back to its human counterpart.
15. Normalize human and AI findings into comparable themes.
16. Separate original participant reasoning from content read off the page.
17. Measure factor overlap, behavioral similarity, final-choice match and finding coverage.
18. Review matches manually to avoid false equivalencies.
19. Build the final analysis and report.
20. Select the strongest human and AI moments and produce highlight reels.

None of these steps is impossible. Together, they describe the research system that TheySaid provides as one workflow: from understanding human behavior, to creating Synthetic Testers, to running the research, capturing sessions, analyzing results and showing the evidence.