Survey Data Quality Audit for Brand Tracking

159 / 160

planted issues caught by five transparent screening rules in a 2,000-response benchmark wave, at the cost of 85 valid responses sent to review.

99.4%	65.2%	12.2%	2.4 to 0.3
recall on planted issues	precision of a flag (159 of 244)	of the wave enters the review queue	points of headline error, before and after screening

EXHIBIT 1

Flags against the truth



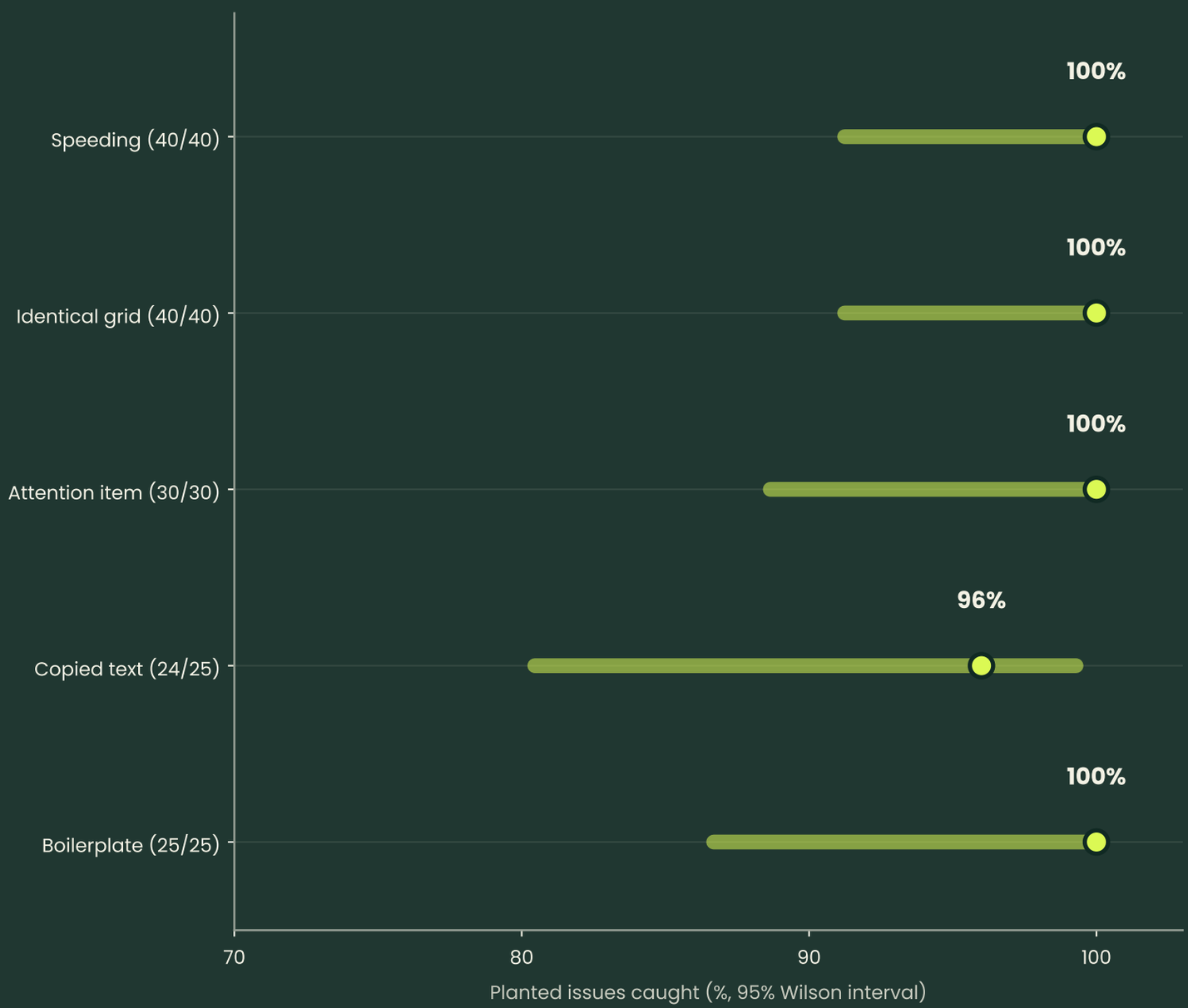
Recall 99.4% (159 of 160), specificity 95.4% (1,755 of 1,840). One planted issue slips through; 85 valid responses are queued.

What the analysis shows

- Unscreened, the wave reads 38.0% likely to recommend. The known clean reference is 35.6%, so low-quality answers inflate the headline by 2.4 points.
- Screening removes almost all of that: the screened wave reads 35.3%, within 0.3 points of the clean reference and well inside its interval.
- In the original records, 104/160 planted answers (65.0%) and 655/1,840 valid answers (35.6%) chose a top-two rating.
- A flag is a prompt to review, not a verdict: 85 of 244 flagged responses (35%) were valid, so automatic deletion would discard good interviews.
- Four of five rules caught every planted case. The copied-text rule missed one of 25.

EXHIBIT 2

Recall by screening rule

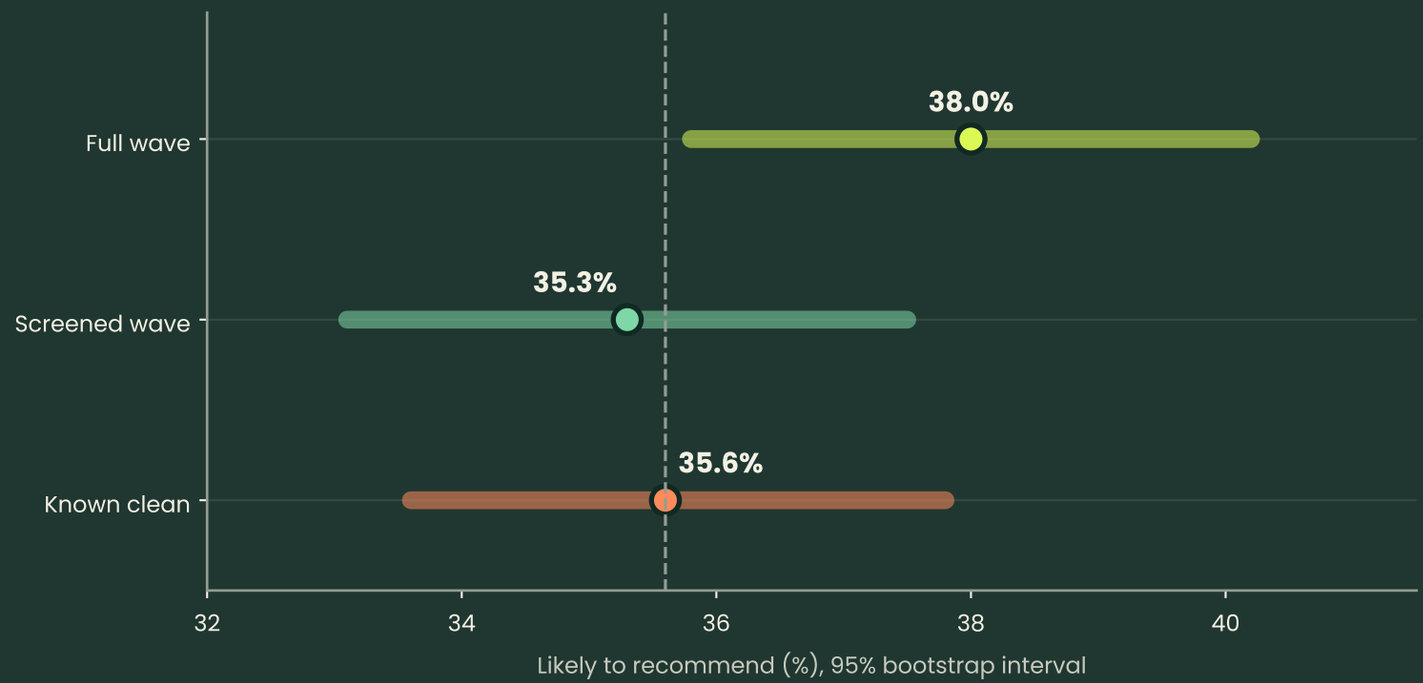


Speeding, identical-grid, attention-item and boilerplate rules caught every planted case. Small counts keep the lower bounds near 85–90%.

Inside the numbers

EXHIBIT 3

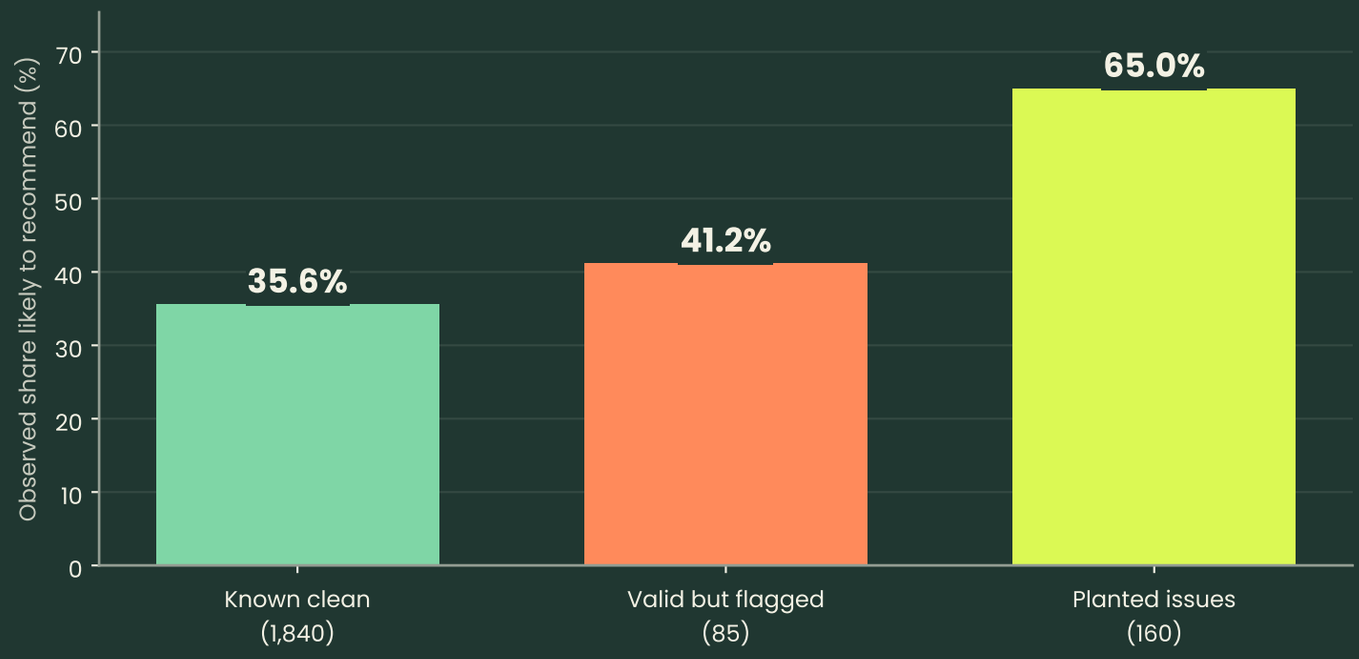
The headline before and after screening



The full-wave interval (35.8–40.2) sits above the clean estimate of 35.6%; the screened wave lands on it.

EXHIBIT 4

Who inflates the score

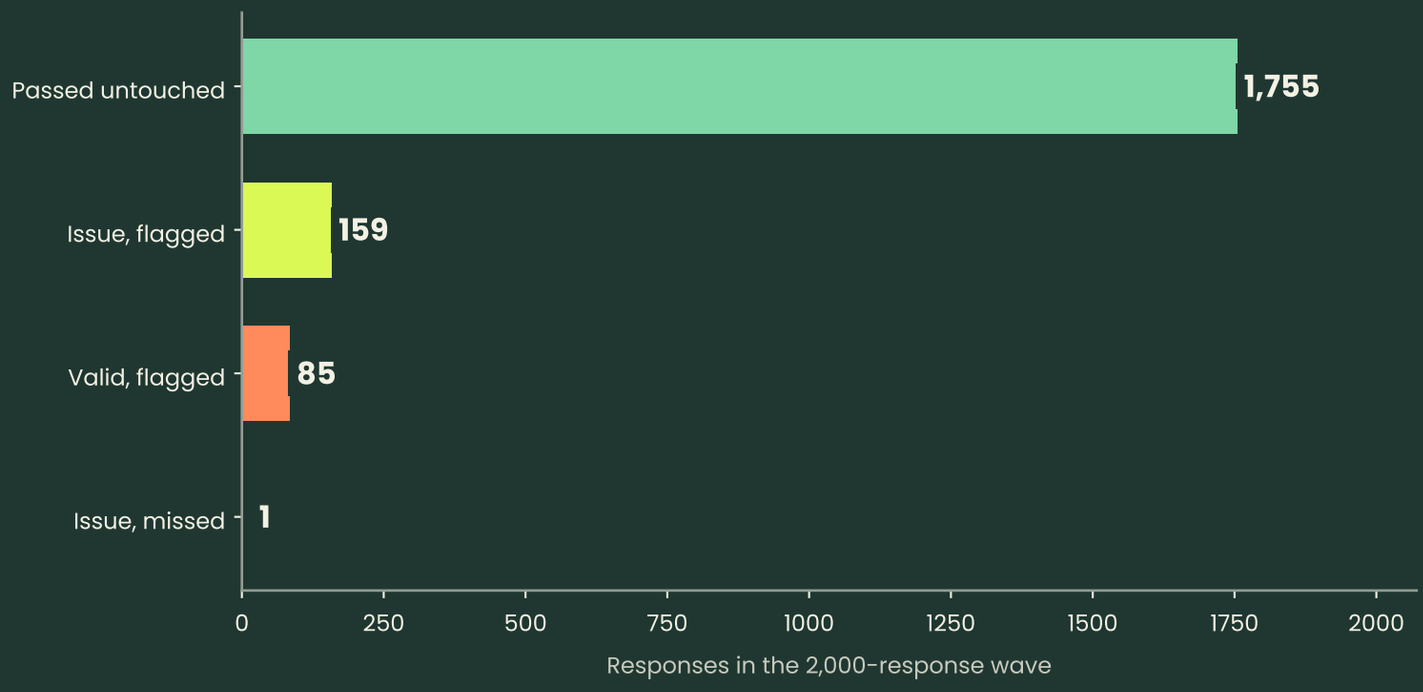


Top-two counts were 104/160 planted, 655/1,840 valid and 35/85 valid-but-flagged. The last group recommends at 41.2%.

What changes the answer

EXHIBIT 5

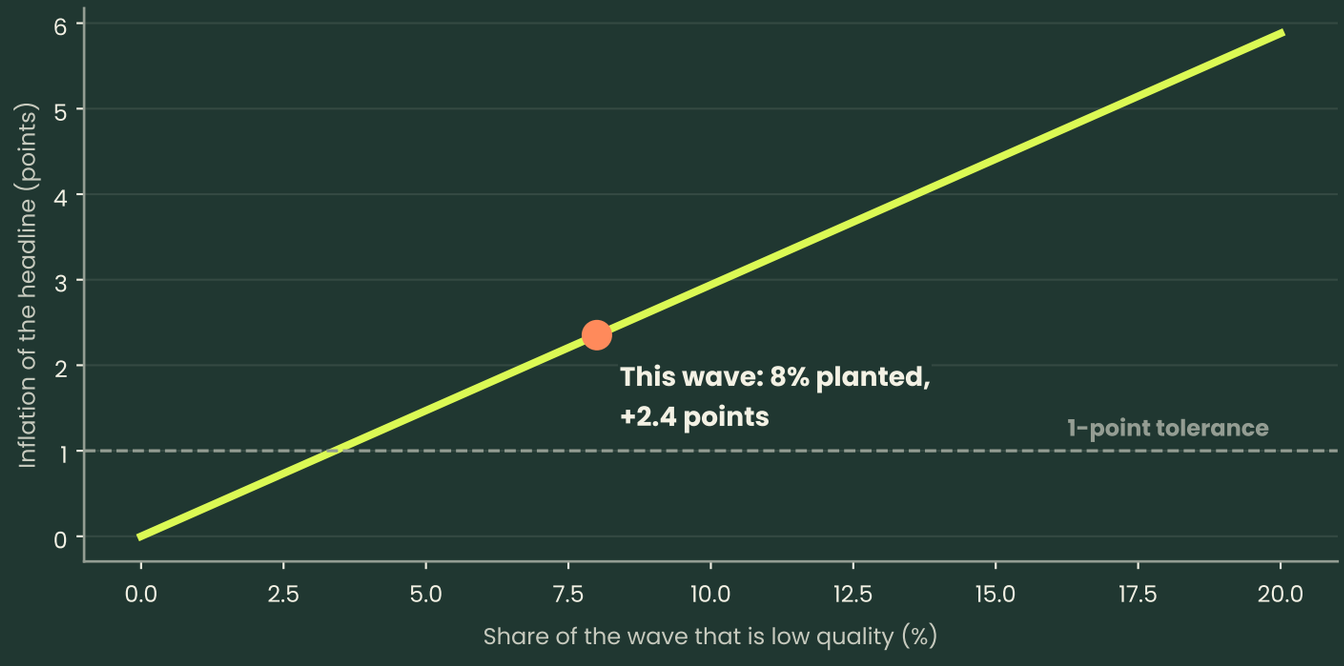
Where 2,000 responses end up



The review queue is 244 responses, 12.2% of the wave. Two thirds of the queue is real issues.

EXHIBIT 6

How much contamination a tracker can absorb



Holding group rates fixed, each 3.4% of planted responses adds about one point. At 8% contamination the increase is 2.35 points.

How it was built and how far it goes

RECOMMENDED USE

- Send all 244 flagged responses to a reviewer; do not auto-delete. Expect about one in three to be reinstated.
- Publish the screened figure and state the unscreened one beside it, so readers see what exclusion changed.
- Re-run the benchmark whenever a rule or threshold changes; the planted cases give a fixed yardstick.

METHOD

- Constructed benchmark: 2,000 responses, 160 of them planted with five known issue types (40, 40, 30, 25, 25).
- Five rule-based checks flag responses; flags are compared with the known truth to give recall, precision and specificity.
- The replay joins original response rows and five-rule flags by ID. Top-two means 4 or 5. Headline intervals use 2,000 bootstrap draws; rule intervals use Wilson.

LIMITS

- The benchmark is constructed, so real fraud may look different from the planted patterns.
- The sensitivity model varies only contamination prevalence and holds both observed group recommendation rates fixed.
- One wave; no test yet of how the rules behave across waves or markets.

NUMBERS RECONCILED

Quantity	Value	How it was checked
Planted issues caught	159 of 160	40+40+30+24+25 from the rule table
Flags raised	244	159 true + 85 valid
Valid untouched	1,755	2,000 - 160 - 85
Precision	65.2%	159 ÷ 244
Headline gap	2.35 pts	759/2,000 minus 655/1,840
After screening	-0.3 pts	35.3 - 35.6