FULL PRODUCT EVALUATION OF THISYS.COM, RELEASE 1.1.0005

Every one of Thisys's 19 product areas was scored, and they average 9.45 out of 10

Seventy-three AI agents used the live product the way a demanding survey professional, a respondent and an attacker would, and worked through the AI-runnable part of our 93-case human test plan. They wrote down every defect with steps to reproduce it, and a second agent re-checked every finding of medium severity or worse. This report sets out what was tested, how it was scored, what was found, and what the evaluation could not cover.

AVERAGE OF ALL 19 AREAS

9.45 / 10

Each of the 19 areas is scored from 0 to 10; this is their weighted average. Sixteen areas scored 9.5 and three scored 9.0; none scored below 9. On our scale, 9 to 10 means best in class compared with Qualtrics, SurveyMonkey and Typeform. Combined with the human test plan (9.49), the release score is also 9.45.

0 critical defects	**1** high defect (AI quiz key check), fixed and verified
9 medium defects	**9.20** average of 8 specialist tracks
220 / 247 feature claims proved true	**302 / 318** user journeys completed end to end

Corrected 29 October 2026: this page first said release 1.1.0005 “ships to thisys.com on 31 October 2026”. It was released on 16 October 2026.

1 What was evaluated

The product

Thisys is an AI survey platform: build a survey, collect responses, analyse them, and get an AI decision report. The evaluation covered release 1.1.0005, the version that was released to thisys.com on 16 October 2026.

The environment

The release ran on our pre-production copy of thisys.com. It runs the same build, sits behind the same Cloudflare network and uses the same live AI model. Payments run in Stripe's test mode, and mail to test addresses is recorded but not delivered.

PART	COUNT	QUESTION IT ANSWERS
Core product areas	19	Is every page and journey good? These are scored and weighted, and make up the headline score.
Specialist tracks	8	Imports and exports, plan limits and billing states, email, authorisation and abuse, AI honesty, admin controls, operations, and browsers and devices. Scored separately.
Feature claims	247	Does the product do what the public Features page says? Each claim was found and used on the live site.
Independent re-checks	every medium+	Is the finding real, and is its severity right? Critical and high findings get two re-checks and a tie-break.
Human test plan	93 cases	Our own manual test plan for a release. AI agents ran 80 of its cases; the rest need a person (real devices, real payments, real sign-in providers).
Coverage sweep	1	Which pages did nobody visit? A final agent walks the routes the others missed.

2 How it was evaluated

Who did the work

The evaluators are AI agents built on Anthropic's Claude models. Each one takes a written brief for one area and works in real browsers (Chromium, Firefox and WebKit, desktop and phone sizes). Each signs in as one of five test accounts covering the Free, Pro, Business and Enterprise plans and a read-only admin.

This is Thisys's own evaluation, run with our own method. It is not a third-party audit or a certification.

What every finding must carry

- The exact page and state, account and screen size
- Numbered steps that reproduce it
- What was expected and what actually happened
- Screenshot or request evidence
- A severity on the scale below

Anything an agent could not test is listed with the reason, never guessed.

How findings are checked

A finding of medium severity or worse goes to a second agent that has not seen the first agent's reasoning. It reproduces the finding from scratch and judges its true severity. A critical or high finding goes to two such agents, and a third, stronger model settles any disagreement. In this cycle no finding was refuted; one reported high was downgraded to medium and one medium to low on re-check.

Rules the agents work under

- They create and change only their own test content
- They never complete a payment or create an account
- Admin access is read-only
- They use at most two real AI actions each
- No personal data goes into any finding

3 The scales

Each area gets a score from 0 to 10 in half points. 9–10 is best in class against Qualtrics, SurveyMonkey and Typeform; 7–8 is solid with minor gaps; 5–6 is usable with notable friction; 3–4 is broken in parts; 0–2 is unusable. The headline is the average of the 19 area scores, weighted by how much each area matters to a survey customer (weights below).

CRITICAL
0
Blocks release: wrong numbers, data loss, a security or privacy breach, or a core journey that cannot be done.

HIGH
1
A core journey badly impaired or misleading.

MEDIUM
9
Noticeable friction or a gap against market-standard tools; a workaround exists.

LOW
85
Polish, wording or a minor accessibility point.

OBSERVATION
81
Not a defect: a competitive gap, an idea, or a detail done well.

Our release bar for thisys.com is 0 critical and 0 high, with a written justification for every medium that ships.

4 Scores for all 19 areas

Every area of the product was evaluated and scored on the same 0 to 10 scale. The headline 9.45 is the average of these 19 scores, weighted by how much each area matters to a survey customer (the weights add up to 100).

AREA	WEIGHT	SCORE ON A SCALE FROM 8.0 TO 10.0	SCORE
Builder core & question types	10		9.5



The scale starts at 8.0 so the differences between areas are visible; every area scored 9.0 or 9.5. Weights add up to 100. Amber dots mark areas below 9.0.

ALL 19 CORE AREAS, CYCLE 46. CORE AREAS ONLY: THE SPECIALIST TRACKS AND THE HUMAN TEST PLAN FOUND THE REST OF THE HIGH AND MEDIUM FINDINGS LISTED IN SECTION 6. COUNTED AFTER RE-CHECKS, BEFORE THE FIXES.

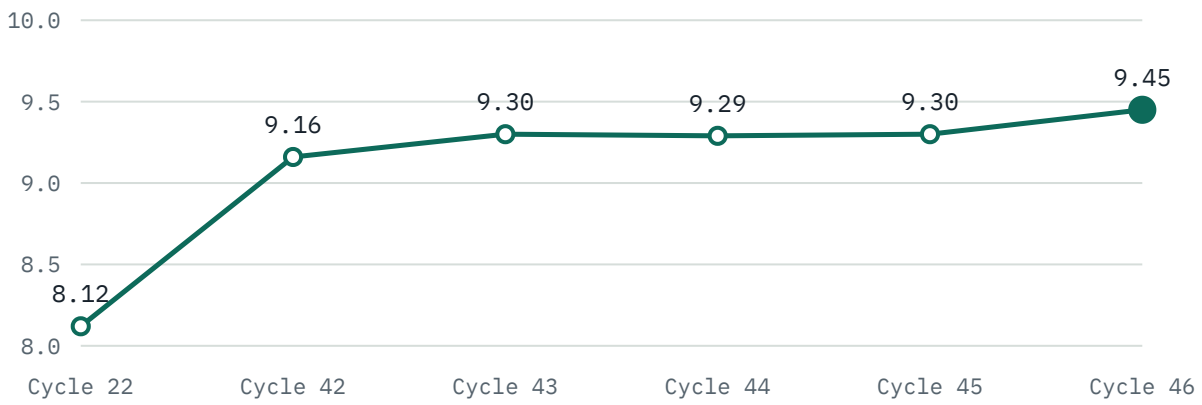
AREA	WEIGHT	SCORE / 10	MEDIUM+	LOW	OBSERVATIONS	JOURNEYS WALKED
Builder core & question types	10	**9.5**	0	3	5	7
Respondent experience & data capture	10	**9.5**	0	1	3	9
Analyse: charts, segments & exports	10	**9.5**	0	2	1	8
Logic, piping, randomisation & quotas	9	**9.5**	0	0	4	9
Collect & fieldwork	8	**9.5**	0	3	3	13
AI Decision Report & sharing	8	**9.5**	0	4	0	10
Create: templates, AI builder, import	6	**9.0**	1	2	0	8
Act: alerts, follow-ups, API	5	**9.5**	0	3	4	9
Protect: consent, privacy, retention	5	**9.5**	0	1	5	10
Languages, translation & accessibility	4	**9.5**	0	2	0	5
Assess: quizzes, grading & certificates	4	**9.5**	0	3	5	12
Account, billing & plan gating	4	**9.5**	0	0	2	9

AREA	WEIGHT	SCORE / 10	MEDIUM+	LOW	OBSERVATIONS	JOURNEYS WALKED
Home, navigation, help & themes	4	**9.0**	1	1	4	8
Sign-up, sign-in & account access	3	**9.5**	0	1	3	8
Team & collaboration	3	**9.5**	0	1	3	10
Security, performance & SEO	3	**9.5**	0	1	6	10
Marketing site, claims & legal	2	**9.5**	0	4	2	6
Public demo experience	1	**9.0**	0	5	4	9
Admin console	1	**9.5**	0	2	4	9
All 19 areas (score: weighted average; counts: totals)	100	**9.45**	2	39	58	169

5 Score across recent cycles

We run a full evaluation after every significant round of change, fix what it finds, and run it again. The first comparable cycle, cycle 22 in September 2026, scored 8.12. Cycle 46 is the first to pass 9.4.

Corrected 29 October 2026: the quiz topic was removed from the high finding below, so the example names no subject area.



Scale 8.0 to 10.0. Cycle 43 ran on a local copy with a stand-in AI model; cycles 44 to 46 ran on the pre-production site with the live model. Cycle 42 was a focused re-check of the previous release.

6 What was found

After re-checks, the evaluation confirmed one high finding and nine medium findings: eight distinct defects, because two evaluators independently reported the same header problem. None affected survey data, calculations or account security. Every one was fixed, passed our full automated test suite, and was re-checked on the live pre-production site on 16 October before the release.

SEVERITY	FINDING	STATUS
HIGH	**AI-drafted quiz with two right answers.** In one AI-drafted practice quiz, a question had two defensible correct options while only one was marked correct, and the answer-key check did not warn about it. A second question had the same issue (medium).	Fixed · verified 16 Oct
MEDIUM	**Large imports refused.** Import files over about 2 MB were refused because of a new limit meant only for small requests.	Fixed · verified 16 Oct
MEDIUM	**Formatted Excel cells.** A cell with partly formatted text was imported as "[object Object]".	Fixed · verified 16 Oct

SEVERITY	FINDING	STATUS
MEDIUM	**Names on a public report.** A shared report could show a name a respondent typed into a name question.	Fixed · verified 16 Oct
MEDIUM	**AI lead form and short briefs.** The AI lead-form drafter refused a short, ordinary brief as too large.	Fixed · verified 16 Oct
MEDIUM	**Header width on some phones.** Signed in, the public header ran off the screen on phones 376 to 412 px wide.	Fixed · verified 16 Oct
MEDIUM	**Two smaller workflow gaps.** The share check missed one quiz warning, and the team access picker could not tell apart two surveys with the same name.	Fixed · verified 16 Oct

The low findings are mostly wording, layout at particular screen sizes and minor accessibility points. Those a user would notice are fixed for the release; the rest are logged for the next one.

7 Specialist tracks

TRACK	SCORE	WHAT IT PROBES
Imports, exports & files	8.5	Spreadsheet and document imports, every export format, file uploads
Plans, limits, credits & billing states	9.5	What each plan allows, AI credits, upgrade paths, plan statements
Email, notifications & suppression	9.5	Invitations, reminders, unsubscribe and suppression lists
Authorisation & abuse	9.5	Access across accounts and team roles, input handling, rate limits
AI failure, abuse & honesty	9.0	AI answers checked against the data, prompt injection, cost limits

TRACK	SCORE	WHAT IT PROBES
Admin write paths	9.5	What staff can change, and whether each change is confirmed and logged
Operations, recovery & deploy health	9.0	Health checks, error pages, backups and deploy safety, from outside
Engines, devices & assistive technology	9.0	Firefox and WebKit, phone sizes, keyboard use, accessibility checks

8 What this evaluation does not cover

- Completed card payments, refunds and plan cancellations: Stripe runs in test mode, and agents never finish a payment.
- Delivery to real inboxes: the pre-production site records mail to test addresses without sending it, so mail was judged from its content and records.
- Completing Google or Microsoft sign-in, and changes to two-factor settings.
- Real phones, screen readers and zoom: browser engines at phone sizes stand in for devices.
- Load and stress testing, which runs separately and only in an approved window.
- The parts of the human test plan that need a person: 35 cases have a step for a real device, a real payment or a real Google or Microsoft sign-in. Until those are done, the human-test score is provisional.
- Twenty-four feature claims that cannot be proved on a test environment, such as dedicated Enterprise infrastructure or a plan's grace period after it ends. They are counted as untestable, not as true.

The evaluators are AI agents. They are thorough and consistent, but they can misjudge a design choice; that is why every serious finding is re-checked and every finding carries steps anyone can follow.

9 How to check this report

Try it yourself

Every claim tested here is on the public Features page. You can try the product without signing up through the demo on thisys.com.

See the evidence

The results files, step-by-step findings, scripts and screenshots behind this report are kept for every cycle. Prospective clients can ask Thisys Technologies for them, with test-account details removed.

Thisys Technologies Pty Ltd · Evaluation cycle 46 · Release 1.1.0005 · Evaluated 16 October 2026

Method: 19 weighted areas, 8 tracks, 247 feature claims, 80 human-test-plan cases, independent re-check of every finding of medium severity or worse. Scores are Thisys's own assessment.