FitFuel

Usability Test Plan · New User Onboarding Flow

Version 4.0 · Tim St Clair, Senior UX Lead · July 2026
Flow: Account Creation → Profile Setup → Plan Delivery · 12 Screens
Includes edge-case supplement: User with hyperthyroidism

01 Prototype Flow Reference

The 12-screen onboarding sequence as observed in Figma Make (July 2026).

Screen	Step	Key Design Decision
App Landing	Entry	Value proposition — does the user feel addressed before they commit?
Create Account	Pre-profile	Account creation before any personalization payoff is visible
Password Validation	Pre-profile	Credential setup; biometric opt-in framing
About You	1 / 6	Physical data (height, weight, age) collected first in the profile sequence
Fitness Background	2 / 6	Self-assessed experience — 4 options; must cover lapsed athletes without forcing a beginner frame
Your Goals	3 / 6	Multi-select, priority-ordered — vocabulary must match how users actually describe intent
Health & Access	4 / 6	Sensitive condition and equipment disclosure mid-flow, before trust is established
Your Schedule	5 / 6	Days, time of day, session length — realistic vs. aspirational input
Activity Preferences	6 / 6	Weighted preferences as last input before plan generation
Building Plan	Loading	AI transparency display — what it claims to be doing with the data
Plan Home Page	Delivery	Personalized plan output — does it feel earned or templated?

02 Method

Recommended method: Moderated remote think-aloud

Session decisions are primarily perceptual and affective — does the physical data request feel invasive? Does the AI rationale feel earned? Click-through metrics alone will capture where people stop, not why. You need spoken reaction. Unmoderated tools (Maze, etc.) will miss the interpretive layer entirely.

Session length: 45–60 minutes per participant

Participants: 5–6 total. Use 6 if the Access Needs participant is recruited.

Protocol: Think-aloud throughout all tasks. Researcher guides; dedicated note-taker or Otter.ai/Fireflies transcript runs in parallel.

Recommended tool: Lookback (screen + face camera + audio). Zoom with screen share is an acceptable budget alternative.

Recording: Obtain explicit consent before each session. Store recordings per your data handling policy.

03 Participant Criteria

Segment A — Returning Routine Builder (3 of 5 participants)

- Has used at least one fitness app in the past 3 years, currently or lapsed
- Has prior fitness history — not a first-time exerciser
- Currently inconsistent or not exercising regularly
- Has stopped a routine at least once due to injury, burnout, schedule, or life disruption
- Exclude: anyone in an uninterrupted active routine for 3+ months

Segment B — Time-Constrained Professional (2 of 5 participants)

- Works 45+ hours per week, carries primary caregiving responsibility, or both
- Has attempted and dropped a fitness routine due to schedule pressure, variable energy, or life demands
- Fitness app experience preferred, not required

Segment C — Access Needs / Edge Case (1 participant — optional but recommended)

Recruit a participant with hyperthyroidism or a comparable chronic condition involving energy variability, heat sensitivity, or heart rate unreliability as a fitness metric. This participant runs the same five tasks as all other participants. Edge-case probes are integrated at Tasks 2, 3, and 5 — flagged clearly in this document. Analyze separately; do not fold findings into the primary cohort summary. Note as a qualifier in the synthesis where findings overlap or diverge.

- Do not brief the participant on the edge-case framing before the session — their unprompted reactions are the data
- If the participant's condition is not represented on the Health & Access screen, that is a primary finding, not an edge observation

Screener disqualifiers — all segments

- Currently employed as a fitness professional
- Works in UX, product design, or app development (will critique the prototype rather than use it naturally)
- Has previously seen or used FitFuel

04 Participant Warm-Up Questions

Administer verbally before showing any part of the prototype. Purpose: establish each participant's real-life context so task behavior can be interpreted against it. Budget 10–12 minutes. Do not show the app during this section.

Open with: "Before we look at anything, I want to hear a bit about your experience with fitness apps and where you're at with your routine right now. There are no right or wrong answers — I'm just trying to understand your world before we get into the app."

Q1 — App history

"What fitness or health apps have you used — either now or at any point in the past? Take a moment — include anything, even if you only used it briefly."

If they name apps, follow up: "Were there any you used seriously for a while — and then stopped?"

Q2 — Frequency and drop-off

"At your most consistent, how often were you using them? And for the ones you stopped — what made you stop?"

Listen for: shame or guilt language, plan-life mismatch, notification pressure, cost, an injury or life event.

Q3 — Current routine

"Walk me through what your fitness life looks like right now — a typical week, if you have one. If it's not consistent, that's fine — just describe where things actually are."

Note: don't push for a tidy answer. "It's kind of all over the place" is more useful than a polished response.

Q4 — Current frustrations

"What are the biggest frustrations you've had — either keeping a routine going, or with the apps themselves? What has let you down?"

Listen for: plans that didn't flex to life, judgmental language, schedule tools that assumed an ideal week, intrusive data requests.

Q5 — Personalization expectations

"When you think about a fitness app that was actually built around your life — not a generic plan — what would that look like? What would it need to know about you that apps don't usually ask?"

Purpose: seeds expectations before the prototype. At Task 5, compare what they said they needed against what the plan surfaces. Gaps are your most direct personalization signal.

⚠ EDGE CASE — Hyperthyroidism Participant Only

After Q3 (current routine), add for the hyperthyroidism participant:

"How does your condition affect your energy day to day — and how does that show up when you try to keep a fitness routine?"

After Q4 (frustrations), add:

"Have you ever felt like a fitness app was giving you guidance that didn't account for your condition — recommendations that felt off or potentially unsafe for where you were that day?"

These responses establish the participant's lived baseline. Cross-reference against Health & Access screen reactions in Task 3 and plan delivery reactions in Task 5.

05 Research Questions

These are the questions the test is designed to answer — not what you ask participants.

#	Research Question

RQ1	Does asking for physical data (height, weight, age) before any personalization payoff creates friction or distrust — particularly for users who have felt judged by fitness products before?
RQ2	Does the Fitness Background screen allow returning or lapsed users to accurately and comfortably place themselves — or does the vocabulary force them into a beginner frame they've outgrown, or an 'experienced' label they don't feel they've earned?
RQ3	Does the Health & Access screen feel safe enough for users with relevant conditions to disclose honestly — or does it read as clinical, gatekeeping, or anxiety-inducing mid-flow?
RQ4	Does the goal vocabulary on the Your Goals screen match how participants actually describe their intent — and is anything meaningfully missing from their real fitness goals?
RQ5	Does the plan delivery feel genuinely personalized to what the participant entered — or does it feel like a template with their name on it? And does it reflect what they said mattered to them in the warm-up?

⚠ **EDGE CASE — Hyperthyroidism Participant Only**

Edge-case research question (hyperthyroidism participant — informs RQ3 and RQ5, not a standalone finding):

RQ-EC: Does the onboarding flow give a user with hyperthyroidism adequate means to communicate their condition's effect on energy variability, heart rate unreliability, and heat sensitivity — and does the plan output reflect those constraints in a way that feels clinically plausible?

This question feeds into synthesis under RQ3 (health disclosure adequacy) and RQ5 (plan personalization). It is not a separate objective.

06 Tasks

All tasks use situational scenario framing, not instruction framing. "Find the health section" is an instruction — it tells the user where to go. "You're setting up the app after a busy week" is a scenario — it lets you observe what they notice, skip, or hesitate on. The distinction is what makes task behavior interpretable as design signal.

Task 1 — First Contact: Landing and Account Creation

Screens: App Landing → Create Account → Password Validation

Situational Frame (read aloud):

"It's a Tuesday afternoon. You've had a busy couple of weeks — work has been relentless, your usual routine has slipped, and you've been meaning to do something about it. A colleague mentioned FitFuel. You've got about ten minutes before your next meeting, so you download it. Go ahead and open it — talk me through what you're seeing and what you're thinking."

Why this frame: mirrors the exact entry condition of both primary segments — a real window of intent after a period of disruption, with a time constraint. 'Ten minutes before a meeting' tests whether account creation feels proportionate to the available moment, without you asking that directly.

What this tests:

- Whether the landing value proposition feels addressed to this user's situation (not a peak athlete's)
- Whether account creation before profile setup registers as a barrier or a natural first step

- Whether the privacy notice ('Your health data is encrypted and never shared') is noticed and believed

Success criteria:

- Participant understands what FitFuel offers without reading every line on the landing screen
- Participant proceeds through account creation without expressing that it feels premature or disconnected
- Privacy statement is noticed or recalled when probed — doesn't need to be first thing read
- Face ID opt-in reads as a convenience feature, not a data-collection mechanism

Failure signals:

- Participant questions why they're creating an account before understanding what the app will do with their data
- Landing imagery or headline doesn't feel relevant to their situation (note for bias audit — do not task this)
- Participant reads the privacy statement skeptically or dismisses it

Follow-up probes:

- "That first screen — before you tapped anything — did it feel like it was talking to you? What made you think that, or not?"
- "At what point did you feel like you understood what this app was actually going to do for you?"
- "Creating the account right at the start — did that feel like the right moment, or did it feel early? Why?"
- "The note about your health data at the bottom — did you notice it? Do you believe it?"

Task 2 — Physical Data and Fitness Background

Screens: About You (1/6) → Fitness Background (2/6)

Situational Frame (read aloud — continue from Task 1, no break):

"You're in — the app is asking you to set up your profile. Go ahead and fill it in as you actually would. If there's anything you'd skip or that gives you pause, say it out loud."

Why this frame: 'as you actually would' gives permission to hesitate or skip without feeling like failure, which surfaces honest behavior rather than compliant behavior. Participants who would not enter their real weight in a real app will tell you — that is exactly the signal you need.

What this tests:

- RQ1 — whether physical data upfront feels justified before any personalization payoff is visible
- RQ2 — whether Fitness Background options map to how this participant honestly describes themselves
- Whether 'Experienced: Consistent training history, on a break' covers the lapsed-athlete without condescension

Success criteria:

- Participant completes 'About You' without pausing to question why height/weight/age is needed at this point
- On Fitness Background, participant selects an option without expressing that none of the four fit
- Sub-labels are understood without confusion — participant doesn't need to re-read them
- Participant does not express discomfort about the physical data request

Failure signals:

- Participant says they would not enter their real weight or age — context for the data request isn't established clearly enough
- Participant hovers between two experience options and says 'I'm not sure which one I am' — most likely failure point for Returning Routine Builders
- Participant selects 'Starting fresh' when their warm-up suggested meaningful prior history — labels being misread

Follow-up probes:

- "When you entered your height and weight — what did you think the app was going to do with that? Did you feel comfortable entering it?"
- "On that experience screen — walk me through how you landed on the option you picked. Was there a moment of uncertainty?"
- "If none of those options felt quite right — what would you have written if there was a text field instead?"

Personalization gap probe:

"Two screens in — what's your sense of how well the app is going to understand your situation? What's it missing so far?"

⚠ **EDGE CASE — Hyperthyroidism Participant Only**

After the Fitness Background screen — hyperthyroidism participant only:

"On that experience screen — did any of those categories account for the fact that your condition "
"affects what 'experienced' actually means for you right now? Where did you land, and does it feel accurate?"

Listen for: whether energy variability or condition-related detraining is factored into their self-assessment, or whether they're forced to choose a category based on training history alone without accounting for current functional capacity. Note whether this is a label problem or a missing option.

Task 3 — Goals and Health Disclosure

Screens: Your Goals (3/6) → Health & Access (4/6)

Situational Frame (read aloud — continue from Task 2):

"Keep going — you're still building your profile."

Why intentionally minimal: Goals and health are the two highest personal-stakes screens. Over-framing will prime responses. The task is to observe cold reaction — what they select, whether they hesitate, and what they say unprompted. The probes do the targeted work afterward.

What this tests:

- RQ4 — whether goal vocabulary matches how participants actually describe their intent, and what's missing
- RQ3 — whether the Health & Access screen feels safe enough for honest disclosure mid-flow
- Whether 'routes PT review when relevant' is understood without clarification
- Whether equipment options are sufficient across participant contexts

Success criteria — Goals:

- Participant selects at least one goal without saying their real goal isn't represented
- Selection order as priority-setter is understood — 'your plan prioritizes the items you choose first' is noticed or understood when probed
- 'Reclaim athletic identity' and 'Manage a health condition' are understood without needing definition
- 'Build consistent habits' is not selected as a fallback because nothing else fit

Success criteria — Health & Access:

- Participant with a relevant condition does not hold back a disclosure they would make to a human trainer
- 'Routes PT review when relevant' is understood without clarification from the researcher
- Equipment options are sufficient — participant finds their context represented
- The screen does not feel like a medical intake form

Failure signals — Goals:

- Participant selects the closest option and says 'that's not quite it' — vocabulary gap
- Participant asks what 'Reclaim athletic identity' means — copy isn't doing enough work
- Multiple participants gravitate to 'Build consistent habits' as a safe default

Failure signals — Health & Access:

- Any participant with a condition says they'd leave it blank on a real app, or discloses less than they would to a trainer
- Participant expresses concern about what the app does with health information, despite the earlier privacy statement
- A participant's specific condition is absent and they note it — not represented is a design gap, not a recruiting anomaly

Follow-up probes — Goals:

- "Looking at those goal options — did any of them feel like they were written for someone else's life, not yours?"
- "Was there a goal you have that wasn't on the list? What would you have added?"
- "Did you understand that the order you selected things in would affect your plan? How did that change how you approached it?"

Follow-up probes — Health & Access:

- "On that health screen — did you feel comfortable sharing what you shared? Was there anything you held back?"
- "Is there something about your body or health situation that you'd want a fitness app to know — that this screen didn't ask about?"
- "What did you think the app was going to do with that information?"

Personalization gap probe:

"You've just told the app your goals and your health situation. Does it feel like it has what it needs to build something genuinely useful — or is there something it still doesn't know about you?"

⚠ **EDGE CASE — Hyperthyroidism Participant Only**

After Health & Access screen — hyperthyroidism participant only:

"Looking at those health options — did you find a way to communicate what your condition actually means "
"for your fitness? Things like energy that varies day to day, or how heat affects you, or the fact that "
"your heart rate might not be a reliable signal?"

"If you left something out — what was it, and why?"

Listen for: whether 'Chronic fatigue' is the closest available option and whether it feels accurate to a hyperthyroidism experience. Note whether heart rate unreliability, heat sensitivity, or medication effects have no disclosure path. These are candidate gaps in the condition taxonomy that would affect plan safety for this population.

Synthesize under RQ3 — health disclosure adequacy. Flag as edge-case qualifier in findings.

Task 4 — Schedule and Activity Preferences

Screens: Your Schedule (5/6) → Activity Preferences (6/6)

Situational Frame (read aloud — replace generic 'keep going' entirely):

"Before you fill this in — take a moment and think about your actual week right now. Not the week you wish you had, but the one that's actually happening. Then set up your schedule based on that."

Why this is the most important instruction in the test: your research documents that users systematically input aspirational availability rather than real availability, and every app treats the plan-reality gap as a

motivation failure when it's a scheduling design failure. 'Actual week right now' is the pressure test. If participants still select three weekday mornings when their warm-up described chaos, the on-screen framing isn't doing enough to pull realistic input.

What this tests:

- Whether 'Built around the time you actually have — not ideal time' shifts behavior toward realistic scheduling
- Whether session length options (20 / 30–45 / 45–60 min) are understood as a preference or a commitment
- Whether non-standard patterns (weekends only, variable shifts) are accommodated by the available options
- Whether 'you can adjust any time' on activity preferences is believed, or dismissed as a disclaimer

Success criteria:

- Participant's schedule selections are consistent with what they described in the warm-up
- Session length is chosen based on real available time, not aspirational intent
- A participant with a compressed or irregular pattern can configure something that reflects their actual life
- Activity preferences are understood as weighted input — not a locked or final selection

Failure signals:

- Participant selects a session length longer than anything described as available in the warm-up — aspirational input despite the framing
- A participant with a non-standard pattern cannot find a configuration that fits and settles for the closest approximation
- 'You can adjust any time' is dismissed — participant says 'but I probably won't'

Follow-up probes:

- "The days and times you picked — is that what your week actually looks like, or what you'd want it to look like?"
- "If your schedule is unpredictable week to week — does this setup handle that? What's missing?"
- "Is there something about how you actually live your life that you couldn't tell the app here?"
- "The 'adjust any time' note on activities — do you believe you'd actually go back and do that?"

Personalization gap probe:

"You've set your schedule and your preferred activities. Is there anything else about your life — work, family, energy levels, travel — that you'd want it to factor in, that there was no place to put?"

Task 5 — Plan Delivery and First Action

Screens: Building Plan (loading) → Plan Home Page

Situational Frame (read aloud — continue from Task 4):

"Go ahead and build your plan. Before you look at what comes up — hold onto one thought: think about what you've told this app about yourself. Then see if the plan reflects it."

Why this frame: 'Think about what you told this app' primes participants to evaluate the plan as personalization — not just as an interface. It doesn't tell them what to look for; it keeps the comparison between input and output active. That's the direct test of RQ5.

What this tests:

- RQ5 — whether the plan feels genuinely derived from inputs or templated
- Whether the AI loading screen builds appropriate credibility or feels performative
- Whether 'Fit Fuel Intelligence' rationale is specific enough to be believed
- Whether 'PT-reviewed,' 'Load-managed,' and 'RPE-guided' labels add confidence or generate confusion
- Whether the first-session CTA is the natural next action

Success criteria:

- Participant reads the 'Fit Fuel Intelligence' rationale and finds it credible — connected to what they entered, not generic
- Participant can identify at least one plan element they trace back to something they told the app
- 'PT-reviewed,' 'Load-managed,' 'RPE-guided' generate curiosity questions, not confused ones
- 'Begin today's baseline session' is the natural next tap — participant doesn't go looking for something else first
- 'Phase 1 of 4' reads as a progression structure, not as more setup steps

Failure signals:

- Participant says 'this could be anyone's plan' — personalization signals aren't landing
- Participant asks what 'RPE-guided' or 'Load-managed' means — jargon the brief prohibits, appearing in the prototype
- Participant looks for their goals, schedule, or health context in the plan and can't find where they've been reflected
- 'Starting Metrics · Baseline — set today' reads as another setup task, not a future tracking anchor

Follow-up probes:

- "Looking at this plan — where do you see yourself in it? Point to anything that feels like it came from what you told the app."
- "Is there anything about your situation — from what you shared earlier — that you don't see reflected here?"
- "The explanation the app gives for the plan — does that feel accurate to what you entered? Does it feel earned, or generic?"
- "If you could add one thing the app should have asked you to make this plan more useful — what would it be?"

- "What would need to be different about this plan for you to trust it enough to actually start today?"

Personalization gap probe — closing (most important of the session):

"Now that you've seen the full setup and the plan it built — think back to what you said at the start about what a fitness app would need to know about you to actually work. Did this app get there? What's the gap?"

⚠ **EDGE CASE — Hyperthyroidism Participant Only**

After plan delivery — hyperthyroidism participant only:

"Looking at this plan — does anything here concern you given your condition? Intensity levels, "
"the session types, the way progress is described?"

"When the app says your plan is built around your health profile — do you believe it actually "
"accounted for what hyperthyroidism means for your body day to day? Or does it feel like it "
"treated your condition as a checkbox?"

"If this plan was real and you started it tomorrow — what would you be watching for or worried about?"

Listen for: whether the plan's intensity framing accounts for energy variability; whether heart rate as an implicit metric (RPE-guided is better, but check) conflicts with condition-related HR irregularity; whether heat sensitivity affects any of the session types surfaced. These observations feed into RQ5 and RQ3 findings under the edge-case qualifier. Do not generalize to the primary cohort.

07 Synthesis Structure

Organize findings from all sessions into the structure below. Core issues are primary — reported across all eligible participants. Edge-case observations from the hyperthyroidism participant are noted as qualifiers under relevant core issues, not as a separate findings section.

Core Issues (primary synthesis)

A core issue is any finding that appears across 3 or more participants, or represents a clear failure against a defined success criterion. Report these first.

ID	Issue	Core Question	Maps to
CI-01	Account creation sequence	Does upfront account creation before personalization create measurable friction or abandonment intent?	Task 1 / RQ1
CI-02	Physical data trust	Is the request for height, weight, and age before any plan payoff perceived as justified or intrusive?	Task 2 / RQ1
CI-03	Fitness background self-placement	Do the four experience options accommodate lapsed athletes without forcing a beginner or overachiever frame?	Task 2 / RQ2
CI-04	Health disclosure safety	Does the Health & Access screen generate honest disclosure — or self-censorship?	Task 3 / RQ3
CI-05	Goal vocabulary adequacy	Does the goal taxonomy match how participants describe their real fitness intent? What's missing?	Task 3 / RQ4
CI-06	Schedule realism	Do participants input aspirational or realistic availability —	Task 4 / RQ2

		and does the screen framing influence this?	
CI-07	Plan personalization credibility	Does the plan output feel earned by the inputs — or templated?	Task 5 / RQ5
CI-08	Personalization gap inventory	Across all tasks — what did participants say the app needed to know that it never asked?	All tasks

Edge-Case Qualifier (hyperthyroidism participant — note under relevant CI)

Do not create a separate edge-case findings section. Instead, annotate relevant core issues with an 'EC note' where the hyperthyroidism participant's findings diverge from or amplify the primary cohort. The following core issues are most likely to carry an EC note:

- CI-03 (Fitness Background) — does the experience taxonomy account for condition-related detraining?
- CI-04 (Health & Access) — are hyperthyroidism-specific constraints (energy variability, heat sensitivity, HR unreliability) expressible on screen?
- CI-07 (Plan credibility) — does the plan account for condition constraints or treat the condition as a checkbox?
- CI-08 (Personalization gaps) — what condition-specific context had no disclosure path?

Synthesis Output Format

For each core issue, record:

- Finding — what participants did or said, in plain language
- Evidence — verbatim quotes, number of participants who expressed it, task and screen reference
- Pass / Fail — against the success criterion defined in the task
- EC note — if the hyperthyroidism participant's finding differs meaningfully from the primary cohort, note it here
- Design implication — one sentence: what needs to change and why
- Priority — High / Medium / Low based on frequency, severity, and proximity to a defined failure signal

Personalization Gap Inventory (dedicated output)

Compile across all warm-up Q5 responses and all task-level gap probes. Organize by input type:

- Context the app never asked for (e.g., stress load, caregiving demands, travel frequency)
- Health or body information with no disclosure path (e.g., medication effects, energy variability, condition-specific constraints)
- Schedule or life patterns the setup couldn't accommodate (e.g., weekends only, shift work, unpredictable weeks)
- Goal language that was absent from the taxonomy

Cross-reference warm-up Q5 responses against Task 5 gap probe responses for each participant. If a participant named something in the warm-up that the app never asked — that is a confirmed gap, not a hypothesis.

08 What This Test Cannot Tell You

- Whether the plan output is clinically appropriate. Participants will react to perceived credibility. Whether loads and progressions are medically sound is a clinical review question.

- Whether onboarding completion rates are acceptable. This is a qualitative study of 5–6 people. It tells you why people would drop and where — not how many. Quantitative dropout data requires a larger unmoderated study.
- Whether aspirational vs. realistic schedule input is a design problem or a behavior problem. If participants enter aspirational schedules despite the framing, you'll know the framing isn't working — but whether the fix is copy, sequencing, or a different interaction model requires a follow-on sprint.
- Whether the hyperthyroidism findings generalize to the broader Access Needs population. One participant is a signal, not a representative sample. Use findings to form hypotheses; validate with a larger Access Needs cohort before any build commitment.
- Long-term retention. This test covers first contact through plan delivery. Day 7 and week 3 behavior are separate studies.

09 Expected Outputs

- Per-session observation notes with screen-level timestamps
- Verbatim quotes tied to each task and specific screen
- Per-task pass / fail table mapped against success criteria
- Synthesis table organized by Core Issues (CI-01 through CI-08) with EC notes where applicable
- Personalization Gap Inventory — compiled from warm-up Q5 and all task-level gap probes, organized by input type
- Prioritized revision list for prototype iteration before build handoff

---

Source basis: Prototype screens (Figma Make, July 2026, 12 screens observed) · Personas v4 · Interview Thematic Clusters (June 2026) · Market Landscape — Opportunity 02 context-aware scheduling (June 2026) · Design Assumptions & Qualifiers (June 2026) · Design Brief Opportunity 01 v1.1. All research questions are working hypotheses pending this test.