N=1 Experimentation and Personal Science

Steady Practice Applied Science Series — SP-9

Steady Practice Research · 2026

Abstract

Personal science — the systematic self-study of one's own behavior, physiology, and cognition using experimental methods — is the intellectual foundation of the Steady Practice platform. This survey covers the methodology of within-person experimentation: why individuals cannot reliably use population-level research to predict their own responses, the statistical and design principles that make self-experiments valid, and what can and cannot be concluded from N=1 data. Key findings: individual responses to behavioral interventions are highly heterogeneous, with effect sizes at the population level often masking a distribution of positive, null, and negative individual responses; within-person crossover designs are far more statistically efficient than between-person designs for detecting individual effects; washout periods, randomization of condition order, and blinding are achievable in everyday self-experimentation and substantially improve validity; Bayesian inference is better suited to N=1 data than frequentist null hypothesis testing; and the personal science community has produced a methodological literature — largely outside academic journals — that deserves serious attention. We cover the case for N=1 over population inference, crossover design, statistical methods, effect size estimation, common threats to validity (confounding, carryover, regression to the mean), and design principles for a platform whose core product is structured self-experimentation.

Contents

1. Why Population Research Cannot Tell You What Will Work for You

1.1 The Heterogeneity Problem

Every randomized controlled trial reports an average treatment effect (ATE) — the mean outcome difference between treatment and control groups. This average conceals a distribution of individual treatment effects (ITEs) that may be wide.

Kravitz et al. (2004) made this point forcefully in a landmark *Annals of Internal Medicine* paper: RCT results describe what happens on average across a population, but the clinical question is what happens to *this patient*. The population average may be a poor guide to individual response when:

- The intervention has heterogeneous effects across subgroups (some benefit, some don't)
- The outcome is continuous and individual variation is large relative to the mean effect

- The individual differs from the trial population in characteristics that moderate the effect

Example from exercise science: The average VO_2max response to aerobic training in intervention studies is approximately 15–20% improvement. But Bouchard et al. (1999) — the HERITAGE Family Study (N=481) — showed that individual training responses ranged from −10% to +100% improvement, with a coefficient of variation of ~75%. The average is almost uninformative for predicting any individual's response. Approximately 20% of participants show minimal response ("non-responders"). Genetic factors account for ~47% of the variance in response.

This is not an isolated finding. Similar response heterogeneity has been documented for: - Dietary interventions and weight change (Zeevi et al., 2015 — glycemic response to identical foods varies 4-fold across individuals) - Sleep need (7–9 hours for most; genuinely shorter for 1–3%) - Caffeine response (fast vs. slow CYP1A2 metabolizers) - Mindfulness and anxiety (some improve markedly; some worsen) - Antidepressants (response rates ~50–60% for any given drug)

1.2 The External Validity Problem

Even if a population-level study identifies a real average effect, it may not apply to you because:

You differ from the study population: Most behavioral intervention trials recruit motivated, health-conscious adults from university communities. Their baseline behaviors, health status, socioeconomic context, and self-selection into trials make them unrepresentative of the general population — and certainly of any specific individual.

Your context differs: An intervention that works in a controlled lab setting or highly supervised program may not produce the same effect when self-administered at home.

Moderators are rarely reported: Even when moderator analyses exist, they typically explain 10–20% of variance in individual responses. The remaining variance is unexplained — and from your individual perspective, it is the entire question.

1.3 The Personal Science Argument

The argument for personal science is not that population research is wrong. It is that population research is *the wrong unit of analysis* for individual decision-making about individual behaviors.

If you want to know whether magnesium supplementation improves your sleep quality, the relevant evidence is not the population-average effect from a meta-analysis (which may be small or null) — it is what happens to *your sleep quality* when you take magnesium vs. when you don't, with confounders controlled, in your specific context. Personal science provides a framework for generating that evidence.

Schork (2015), writing in *Nature*, called for a systematic shift toward "participant-centric trials" — n-of-1 designs where the individual is both researcher and subject — precisely because population averages are an inadequate guide to individual treatment decisions. Such trials are increasingly used in precision medicine to resolve exactly the question population RCTs cannot answer: what happens to *this person*?

---

2. The N=1 Design Toolkit

2.1 The Basic Crossover

The fundamental N=1 design is the crossover: the individual alternates between treatment and control conditions across multiple periods, with their own baseline as the control.

Why crossover is efficient: In a parallel-group RCT, between-person variance (the fact that people differ) is a source of noise that must be statistically controlled. In a crossover, the same person serves as their own control — between-person variance is completely eliminated from the treatment effect estimate. This makes crossover designs dramatically more statistically efficient.

Senn (2002) showed that for outcomes with high between-person variance relative to within-person variance (the intraclass correlation is low), the efficiency gain from crossover can be 5–10× — meaning a crossover with 10 observations can have the same statistical power as a parallel-group trial with 50–100 participants.

The basic N=1 crossover: - Define the treatment (e.g., magnesium glycinate 400 mg before bed) and control (placebo or usual behavior) - Define the outcome (e.g., sleep quality score, HRV) - Alternate between conditions in blocks (e.g., week on, week off) for 4–8 cycles - Randomly assign condition order within blocks - Analyze: compare the within-person mean outcome under treatment vs. control

2.2 Randomization of Condition Order

Random assignment of condition order is as important in N=1 trials as in RCTs. Without randomization:

- Time trends (seasonal variation, life events, progressive adaptation) can be confounded with treatment
- Expectation effects (knowing what comes next) can bias outcomes
- The experimenter (who is also the participant) may unconsciously favor certain orderings

Block randomization: Randomize within blocks (e.g., in each 2-week block, randomly assign which week is treatment and which is control). This balances time trends within each block while maintaining some randomization structure.

Coin-flip protocols: For daily alternation, a simple coin flip for each day works if the intervention has no carryover (see Section 3.2). Apps or random number generators can automate this.

2.3 Washout Periods

A washout period is a gap between conditions during which the effects of the previous condition dissipate before the new condition begins. Washout is essential when:

- The treatment has residual effects (e.g., supplements take days to clear; exercise adaptations persist for weeks)
- The previous condition changes the baseline state in ways that would confound the next condition (carryover)

Washout length: Determined by the half-life of the treatment effect, not the half-life of the substance. For supplements, washout of 2–3 biological half-lives of the active compound is

typically sufficient. For behavioral interventions (exercise habits, dietary patterns), washout may need to be 1–2 weeks to allow adaptation effects to decay.

Practical washout for common self-experiments: - Caffeine dosing: 2–3 days (half-life ~6 hours; 4–5 half-lives = ~30 hours + buffer) - Magnesium supplementation: 5–7 days - New exercise protocol: 2 weeks (adaptation effects persist longer than substance clearance) - Dietary pattern change: 2–4 weeks (gut microbiome, metabolic adaptation) - Mindfulness practice: 1–2 weeks

2.4 Blinding

Single-blind N=1 designs (where the participant does not know which condition they are in) are achievable for supplement experiments using identical-looking capsules prepared in advance. This eliminates expectation effects — a major threat to validity for subjective outcomes (mood, energy, focus).

For behavioral interventions (exercise, diet, sleep habits), blinding is not achievable. In this case, pre-specifying outcomes and analysis plans before running the trial reduces the risk of post-hoc outcome selection.

2.5 Ecological Momentary Assessment

Ecological momentary assessment (EMA) is a methodology for repeated within-person measurement in real time and in real-world contexts — as opposed to retrospective recall at the end of a day or week. It is the methodological backbone of mobile logging apps, and its principles apply directly to self-experimentation.

Shiffman, Stone & Hufford (2008) provide the definitive review. Key EMA principles relevant to personal science:

- **Signal-contingent sampling**: prompt at random intervals (e.g., 3 times per day at random times) to capture outcomes without reactivity to a fixed schedule
- **Event-contingent sampling**: log immediately after a specified event (e.g., after every meal, after every exercise session) for outcomes closely tied to specific behaviors
- **Retrospective bias**: end-of-day recall systematically underweights emotionally neutral

events and overweights the peak and end of the day (Kahneman's peak-end rule); real-time or near-real-time logging is substantially more accurate for mood and energy ratings

- **Burden calibration**: more frequent prompts yield richer data but reduce compliance; 3–4 prompts per day is the empirical sweet spot for most outcomes

For platform design, EMA principles argue for prompting users close to events of interest (just after waking for sleep quality; just after eating for satiety and energy) rather than at a fixed evening check-in.

2.6 Multiple Replications

The fundamental source of statistical power in N=1 designs is not sample size (N of people) but number of replications (number of treatment-control cycles). More cycles increase power to detect the within-person effect.

Schork (2015) showed that an N=1 crossover with 8–10 treatment-control cycles can achieve 80% power to detect a medium effect size ($d \approx 0.5$) at a two-tailed $\alpha = 0.05$ — comparable to a parallel-group RCT with ~30 participants per group. Twelve or more cycles extends power to detect smaller effects ($d \approx 0.4$) or provides additional confidence at the same effect size.

Practical implication: A 12-week self-experiment with weekly alternation (6 treatment, 6 control weeks) provides adequate power for most personally meaningful effect sizes. Longer experiments (16–20 weeks) are worth the investment when the expected effect is small or the outcome is noisy.

---

3. Threats to N=1 Validity

3.1 Confounding

Confounding occurs when a third variable co-varies with both the treatment and the outcome, creating a spurious association.

Time-varying confounders: In N=1 experiments, the main confounders vary over time: work stress, travel, illness, social events, seasonal variation, life events. These can create apparent treatment effects if they happen to correlate with condition assignment.

Control strategies: - Random condition assignment disrupts systematic confounding - Tracking known confounders as covariates (logged in the same check-in as outcomes) - Long experiments with multiple replications average out random confounders - Restricting experiments to stable life periods (no travel, no major events)

Example: A user running a caffeine-free experiment during a week that also involves a stressful work deadline will likely see worse sleep regardless of caffeine status. Without tracking stress as a covariate, this confound is invisible.

3.2 Carryover Effects

Carryover occurs when the effect of one condition persists into the next condition period, contaminating the comparison.

Biological carryover: Supplements, dietary changes, and exercise adaptations create biological states that persist beyond the condition period. A 1-week creatine loading protocol followed immediately by a control week is not a clean control — creatine stores remain elevated for 2–4 weeks.

Psychological carryover: Learning, habituation, and adaptation to a behavioral intervention may persist. A user who practiced meditation for 2 weeks may retain some attentional benefits during the subsequent control period.

Detection: Plot treatment effects by cycle number. If later cycles show systematically different effects than early cycles, carryover is likely.

Mitigation: Adequate washout (Section 2.3), avoid short conditions for slow-clearing interventions, analyze cycle order as a covariate.

3.3 Regression to the Mean

If a user begins an experiment during a period of unusually poor performance (e.g., poor sleep), the subsequent treatment period will appear to show improvement simply due to regression to the mean — performance naturally moves toward the individual's average regardless of the treatment.

Detection: Compare baseline periods in treatment vs. control blocks. If treatment weeks

systematically follow worse baseline periods, regression to the mean may explain apparent effects.

Mitigation: Random assignment of condition order; adequate baseline measurement before starting; including pre-period values as covariates.

3.4 Observer Effects and Reactivity

The act of measurement changes what is being measured (see SP-2). In N=1 experiments, the participant knows they are being measured, which may itself change behavior. This is not a threat to internal validity if it affects both conditions equally, but it reduces generalizability to unmonitored behavior.

Practical rule: If tracking behavior during an experiment changes the behavior substantially, the experimental result reflects "behavior while being tracked" not baseline behavior. For some outcomes (mood, subjective performance ratings), this effect is minimal. For behaviors like dietary tracking, it can be substantial.

3.5 Expectation and Placebo Effects

For subjective outcomes (energy, mood, focus), expectation is a strong driver of apparent treatment effects. A user who believes magnesium improves sleep will likely report better sleep on magnesium nights regardless of pharmacological effect.

Mitigation: Blinding (see Section 2.4); pre-specifying what size effect would constitute a meaningful result before running the trial; using objective outcomes when available (wearable sleep data rather than self-reported sleep quality).

---

4. Statistical Methods for N=1 Data

4.1 Why Frequentist NHST Is a Poor Fit

Null hypothesis significance testing (NHST) — the p-value framework — was designed for between-person population research. Its application to N=1 data creates several problems:

- **The null hypothesis is rarely the right question**: In a self-experiment, the question is not "does this intervention have a non-zero effect at the population level?" but "how large is the effect for me, and how confident should I be in that estimate?"
- **Power is often low**: With 8–12 observations per condition, p-values are unstable and $p < 0.05$ may require implausibly large effect sizes to achieve
- **Binary thinking**: $p < 0.05$ vs. $p > 0.05$ is a poor decision framework for personal decisions where effect size and uncertainty are both relevant

4.2 Bayesian Inference for N=1

Bayesian inference is better suited to N=1 experimentation because it: - Produces probability distributions over effect sizes rather than binary decisions - Allows incorporation of prior knowledge (what does the literature say about typical effect sizes for this intervention?) - Updates naturally as more data accumulates - Outputs that directly answer the user's question: "What is the probability that this intervention has a meaningful effect for me?"

Basic Bayesian N=1 analysis:

Let μ_T be the mean outcome under treatment and μ_C be the mean outcome under control. The quantity of interest is $\delta = \mu\_T - \mu\_C$.

1. Specify a prior on δ: for supplement experiments, a weakly informative prior centered at 0 with $\sigma \approx 0.3$ standard deviations is reasonable
2. Observe the data (n_T treatment periods, n_C control periods)
3. Update the prior via Bayes' theorem to obtain the posterior distribution on δ
4. Summarize: posterior mean (best estimate of effect), 95% credible interval (uncertainty), and $P(\delta > \text{threshold})$ — the probability the effect exceeds a personally meaningful threshold

Worked example: 12 weeks of alternating 1-week magnesium / 1-week control. Sleep quality measured daily (0–10 scale). Treatment mean = 6.8; control mean = 6.2; within-person SD = 1.1; SE of difference = 0.45.

Posterior (with weakly informative prior): $\delta \approx 0.6$ (95% CI: -0.3 to 1.5). $P(\delta > 0.5) \approx 52\%$. Interpretation: there is weak evidence of a personally meaningful effect — replication with more cycles or an objective measure is warranted before concluding the intervention works for

this individual.

4.3 Time Series Methods

When data are collected continuously (daily HRV, sleep scores) rather than as condition-period averages, time series methods account for autocorrelation — the fact that today's outcome is correlated with yesterday's.

Interrupted time series (ITS): Model the outcome as a time series with an intervention indicator. This is appropriate when the experiment has a single treatment period followed by a single control period (not ideal) but can be extended to multiple crossover periods.

ARIMA with intervention terms: Autoregressive integrated moving average models can include condition assignment as a covariate while modeling within-person autocorrelation structure. These require more data (20+ observations) than simpler approaches.

Practical recommendation: For most self-experiments with weekly condition blocks, averaging the daily values within each week and treating the weekly averages as independent observations is adequate. Autocorrelation within weeks is averaged out; across-week autocorrelation is typically small.

4.4 Effect Size Estimation

For personal decision-making, the most useful statistic is the within-person standardized effect size:

$d\_w = (\mu\_T - \mu\_C) / \sigma\_within$

where σ_within is the standard deviation of the within-person outcome over time. This differs from the between-person Cohen's d reported in population research.

Interpreting within-person effect sizes: - $d\_w \approx 0.2$: small, likely not noticeable in daily life - $d\_w \approx 0.5$: medium, noticeable with attention; probably worth the effort for a significant behavior change - $d\_w \approx 0.8$: large, reliably noticeable in daily life; likely worth the effort for any behavior change - $d\_w > 1.0$: very large, extremely noticeable; rare for behavioral interventions in non-clinical populations

Minimum detectable effect: With 12 observation cycles (6 treatment, 6 control), an experiment has 80% power to detect d_w $\approx$ 0.5 (medium effect) at a two-tailed $\alpha = 0.05$ — consistent with the Schork (2015) estimate in Section 2.5. This is the practical lower bound for most self-experiments; experiments targeting smaller effects (d_w < 0.3) require 20+ cycles to achieve adequate power.

---

5. What Can and Cannot Be Concluded from N=1 Data

5.1 What N=1 Can Show

Within-person causal effects: If the design is valid (randomization, adequate washout, controlled confounders), a well-designed N=1 trial provides strong evidence of the causal effect of an intervention on an outcome *for that individual in that context.*

Personalized dose-response: Multiple experiments varying the dose can characterize the individual's dose-response curve for a given intervention.

Context dependencies: Multiple experiments under different conditions (e.g., with vs. without exercise, during high-stress vs. low-stress periods) can identify moderators of the individual's response.

5.2 What N=1 Cannot Show

Generalization across contexts: A result from a 3-month winter experiment may not replicate in summer. A result during a period of high work stress may not replicate during vacation. N=1 results are context-specific.

Mechanism: Observing that magnesium improves your sleep does not reveal why. Multiple plausible mechanisms (GABA-A agonism, NMDA antagonism, muscle relaxation, placebo) are consistent with the observation.

Generalization across individuals: Your N=1 result tells you nothing about whether the intervention will work for other individuals — though it does tell you that individual variation exists and that population means are not universal.

Long-term effects: Most self-experiments run 4–12 weeks. Long-term effects (months, years) cannot be assessed in this window.

5.3 The Replication Principle

Replication is as important in personal science as in academic science. A result that replicates across multiple independent experiments — different time periods, different contexts, different conditions — is substantially more reliable than a single well-designed experiment.

Internal replication: Running the same experiment twice, at different time points, and observing consistent results is strong evidence.

Cross-domain replication: If an intervention improves sleep quality AND HRV AND next-day energy, the convergence of independent outcomes strengthens the inference.

---

6. The Personal Science Community

6.1 Quantified Self and Self-Experimentation Culture

The Quantified Self movement (Gary Wolf, Kevin Kelly; Wired magazine, ~2007) explicitly framed self-tracking as personal science. QS Meetups around the world have since 2010 hosted thousands of "show and tell" presentations of individual self-experiments, creating a practitioner literature of N=1 methodology.

Prominent self-experimenters have published detailed methodology: - **Seth Roberts**: Self-experimented systematically with sleep timing, mood, and weight for decades. His 2004 paper in *Behavioral and Brain Sciences* on self-experimentation methodology is one of the most rigorous treatments in the literature. - **Tim Ferris**: Popular science dissemination of self-experimentation, lower methodological rigor but high influence on the QS community. - **Peter Attia**: Systematic self-experimentation with metabolic health, sleep, and longevity interventions with increasing methodological sophistication.

6.2 N-of-1 Trials in Clinical Research

The N-of-1 clinical trial literature predates the QS movement. Guyatt et al. (1986, *New England Journal of Medicine*) established formal criteria for clinical N-of-1 trials and demonstrated their usefulness for guiding individual patient treatment decisions (e.g., which of two asthma medications works better for this patient?).

Key methodological contributions from clinical N-of-1 literature: - Preference for crossover over single-period designs (Guyatt et al., 1990) - Importance of blinding for subjective outcomes (Nikles et al., 2006) - Role of washout in chronic condition management (Zucker et al., 2010) - Aggregation of multiple N-of-1 trials to estimate population effects (Zucker et al., 1997)

6.3 Precision Medicine and Personalized Health

The precision medicine movement explicitly aims to predict individual treatment responses rather than average responses. The 2015 NIH *All of Us* program (N=1 million+) is designed in part to build individual-level prediction models.

For behavioral interventions, personalized response prediction is less mature than for pharmacological treatments, but the principle is the same: population averages are insufficient for individual decision-making.

---

7. Practical Self-Experiment Design: A Framework

7.1 The Question-First Principle

Start with a specific, answerable question, not a vague curiosity. "Does X improve Y?" is answerable if X is a discrete, manipulable variable and Y is a measurable outcome.

Good questions for self-experimentation: - Does taking magnesium glycinate (400 mg) before bed improve my Oura sleep score? - Does a 20-minute walk after lunch reduce my afternoon energy dip (self-rated, 1–10)? - Does cold shower exposure (2 minutes) improve my

morning HRV vs. no cold exposure? - Does no alcohol for 4 days before a night improve my next-day cognitive performance rating?

Poor questions for self-experimentation: - Does magnesium improve my health? (Too vague; which outcome? what dose?) - Is intermittent fasting good for me? (Too broad; which outcomes? what window?) - Am I a morning person? (Not a treatment-comparison question)

7.2 The Minimum Viable Experiment

For a valid self-experiment: - **Treatment**: defined precisely (what, how much, when) - **Outcome**: measured consistently (same time, same method, each period) - **Duration**: at least 6 treatment + 6 control periods (longer for smaller expected effects) - **Randomization**: condition order randomized within blocks - **Washout**: appropriate to the intervention (at least 2–3 days for most supplements) - **Confounders tracked**: the 2–3 most likely confounders logged alongside outcomes

7.3 Interpreting Results

After running a self-experiment, the question is: what should I conclude and what should I do?

Decision framework: 1. **Estimate effect size**: δ = mean(treatment) − mean(control), in outcome units and in within-person standard deviations 2. **Assess uncertainty**: Is the credible interval narrow enough to be actionable? (If CI includes both +0.8 and −0.2, the experiment is uninformative) 3. **Compare to threshold**: Does the estimated effect exceed a personally meaningful threshold? (e.g., >0.5 points on a 10-point sleep scale) 4. **Assess plausibility**: Does the result make mechanistic sense? Does it align with prior evidence? 5. **Decide and plan**: Adopt, reject, or replicate. If the result is positive and above threshold, adopt and retest in 6 months. If negative, reject and move on. If uncertain, replicate with more cycles.

7.4 Stopping Rules

Self-experiments do not always need to run to a pre-specified endpoint. Stopping rules — criteria for ending an experiment before the planned duration — prevent wasted effort and

protect against harm.

Stop for harm: If an objective marker worsens substantially from personal baseline during a treatment condition (e.g., HRV drops >25% below the 30-day average, or sleep duration falls >60 minutes below average for 3+ consecutive days), pause the experiment regardless of the planned endpoint. The self-experiment should never override clear physiological safety signals.

Stop for futility: After 6+ treatment-control cycles, if the estimated effect size is <0.1 within-person standard deviations with a narrow credible interval, the experiment is informative: the treatment has a negligible effect for this individual. Additional cycles will not change the conclusion.

Stop for clear success: If after 4–6 cycles the posterior probability $P(\delta > \text{threshold}) > 90\%$, the result is practically conclusive. Continuing only adds marginal precision; the adoption decision is already well-supported.

When not to stop: If effect estimates vary widely across cycles, or the credible interval is wide, more data genuinely increases precision. For noisy outcomes (subjective mood, energy) where within-person variance is high, stay with the design until the CI narrows enough to be actionable.

7.5 Sequential Experimentation: Building a Personal Protocol

Most users will run multiple experiments over time. Sequencing matters.

One experiment at a time. Running simultaneous experiments on overlapping outcomes makes attribution impossible. If sleep quality improves while testing magnesium AND a consistent wake time simultaneously, neither can be attributed. Sequential testing is the scientific requirement.

Prioritize by expected value. Test interventions with the highest prior probability of meaningful effect first — for most people, sleep timing before supplements before advanced protocols. The highest population-level evidence predicts the highest probability of a detectable individual signal.

Use prior results as priors. If a magnesium experiment showed a plausible but uncertain effect ($P(\delta > \text{threshold}) = 60\%$), this prior should inform the replication design — wider uncertainty means more cycles are needed to resolve it.

Sequence recovery before performance. Attempting to optimize cognitive or physical performance when sleep is unresolved produces experiments contaminated by the primary unfixed variable. Establish recovery baseline first (sleep, HRV, stress); test performance interventions second.

Build toward a personal protocol. The output of sequential experimentation is not a list of individual results but a coherent personal protocol: interventions confirmed effective for this individual, at this dose, in this context. Experiments confirmed negative are as valuable as positive ones — they remove candidates efficiently.

7.6 Complete Worked Example: Setup to Decision

Question: Does magnesium glycinate (400 mg, 30 min before bed) improve my Oura sleep score?

Prior: Meta-analytic evidence shows small positive effects in magnesium-deficient populations (d $\approx$ 0.15–0.3). As a healthy adult with varied diet, deficiency is possible but not assumed. Prior: δ centered at 0 with weak positive pull (mean 0.2 SD, $\sigma = 0.3$ SD).

Design: - Treatment: 400 mg magnesium glycinate, 30 min pre-bed - Control: identical-looking placebo capsule (gelatine), prepared by a friend using block randomization - Outcome: Oura sleep score (0–100), auto-captured each morning - Duration: 12 weeks, alternating weekly (6 treatment, 6 control weeks) - Washout: first 2 days of each week excluded from analysis to allow clearance between conditions (magnesium half-life ~3–5 days; weekly alternation with 2-day buffer is adequate) - Confounders tracked daily: alcohol units, bedtime, exercise, perceived stress (1–10)

Data summary (12 weeks): Treatment mean: 72.8 (SD across 6 weeks: 3.1). Control mean: 69.1 (SD: 3.3). Raw difference: 3.7 points. Within-person SD across all weeks: 6.4 points.

Analysis: - Within-person effect size: d_w = 3.7 / 6.4 = 0.58 (medium) - Bayesian update: posterior mean 3.4 (slight pull toward prior), 95% CI [0.8, 6.0] - P(δ > 2 points — the pre-specified meaningful threshold): 82% - Confounder check: three high-stress weeks distributed across both conditions (not systematically correlated with treatment) — confounding not detected

Decision: Moderately strong evidence that magnesium meaningfully improves sleep for this

individual (P = 82% of exceeding the personal threshold). Decision: adopt for 3 months, recheck sleep baseline to confirm sustained effect. Schedule replication if baseline drifts.

Key lessons from this example: - The 2-day washout exclusion reduced usable observations but substantially improved validity — without it, carryover from condition transitions would contaminate estimates - Blinding was achievable because the treatment was a supplement; for behavioral interventions it is not, making pre-specified outcomes more important - The prior pulled the estimate modestly downward ($3.7 \rightarrow 3.4$), reflecting appropriate scientific humility — the data alone would overstate certainty - Logging stress as a covariate enabled the confounder check that validates the result; without it, three bad-sleep stress weeks could have appeared to favor one condition

---

8. Design Principles for Steady Practice

The experiment as the product. The platform's core value is not tracking — it is valid self-experimentation. This means making crossover design, randomization, washout, and confounder tracking as easy to set up as a basic todo app.

Show the user what they're learning, not just what they're doing. A Bayesian posterior on their personal effect size — updated each week — is more valuable than a streak counter. "Based on your 6 cycles, the probability that magnesium meaningfully improves your sleep is 74%" is the product.

Set design expectations explicitly. Most users will not naturally think about washout, carryover, or confounders. The onboarding for a new experiment should surface these: "We recommend a 5-day washout before starting" is a design suggestion, not a constraint.

Calibrate effect size thresholds individually. What counts as a meaningful effect differs by outcome and user. Prompt users to set their minimum effect threshold before running the experiment. This prevents post-hoc rationalization of marginal effects.

Build the replication norm. Single experiments are suggestive; replicated experiments are reliable. After a positive result, the platform should prompt: "Your first experiment suggests magnesium improves your sleep. Want to replicate it to be more confident?" This builds the scientific culture the platform depends on.

Pool carefully. Aggregating results across users can increase precision on average effects, but the whole motivation for N=1 experimentation is that individual effects differ from population averages. Pooling should augment, not replace, individual inference. Present pooled results as: "Most users who tried this saw a small improvement — but responses vary widely. Your own experiment is the best guide."

Individual Variation

Personal science methodology interacts with individual characteristics to produce large differences in experiment quality, interpretation accuracy, and protocol completion. Understanding these sources of variation allows practitioners to adapt the standard framework to their own profile rather than assuming one design fits all.

Scientific training predicts design quality but not engagement or outcome validity. Lay experimenters with no formal training produce valid self-experiments when given structured protocols. Plsek & Greenhalgh (2001) and subsequent N-of-1 methodology work (Nikles et al., 2006) show that the limiting factor in self-experimentation is design discipline — following a protocol consistently — not domain knowledge. However, individuals with analytical training do show measurably more reliable interpretation of ambiguous results. Without training, confirmatory bias inflates the probability of a false positive conclusion by an estimated 30–40% in unblinded self-experiments (Mosconi et al., 2010).

Interoceptive accuracy determines whether subjective ratings are valid outcome measures. Interoception — the ability to accurately perceive internal body states — varies substantially across individuals. High-interoceptive individuals produce outcome ratings with lower test-retest variability and higher correlation with objective physiological markers. Low-interoceptive individuals (a trait measurable via heartbeat detection tasks; Garfinkel et al., 2015) often generate inconsistent subjective ratings that produce noisy outcomes and reduced statistical power. The practical adaptation is not to exclude subjective measures but to supplement them: individuals who notice high variability in their self-ratings should add objective proxies — HRV, wearable sleep scores, reaction time tests — as primary or co-primary outcomes.

Cognitive style determines the most important design control. Analytical thinkers

show more reliable interpretation of ambiguous crossover data; intuitive thinkers are prone to confirmatory interpretation, particularly when results are marginally consistent with prior belief (Epstein et al., 1996). For intuitive-style individuals, the single most effective design control is pre-specifying the decision criterion in writing before running the experiment: "I will adopt this intervention if $P(\delta > 0.5 \text{ SD}) > 80\%$." This prevents the subjective reinterpretation of thresholds after results are visible.

Time horizon patience directly determines minimum viable experiment duration. Some individuals complete 3–4 week crossover periods without difficulty; others abandon experiments mid-protocol due to novelty-seeking traits, high external time pressure, or low delay discounting. Research on protocol adherence in self-quantification contexts (Swan et al., 2013) suggests dropout risk rises sharply after two weeks for individuals with high novelty preference. Attempting to match experiment duration to ideal statistical power rather than personal patience reliably produces incomplete experiments that yield no information. A one-week crossover completed on both conditions is more valuable than a four-week design that ends on day 11.

Genetic and physiological predictors of self-experiment success. Trait impulsivity (measurable via BIS-11 scale) predicts protocol abandonment rate; high-impulsivity individuals are better served by daily check-ins and shorter periods. Trait openness to experience predicts willingness to engage with quantitative results and replication. Individuals with lower baseline resting HRV — who tend to have lower allostatic buffer — show higher within-person outcome variability, requiring more measurement periods to reach the same inferential precision as higher-HRV peers.

Practical self-experiment implication. Before designing your first experiment, assess two things: (1) run a brief interoception check — rate your mood and energy three times in one day without looking at prior ratings, then check consistency; high variability means you need an objective co-primary outcome; (2) complete a 7-day pre-experiment baseline tracking period and note whether you follow through reliably. If you miss more than two days, shorten your planned experiment design until the commitment matches your demonstrated follow-through rate.

References

Bouchard, C., An, P., Rice, T., Skinner, J. S., Wilmore, J. H., Gagnon, J., … & Rao, D. C. (1999). Familial aggregation of VO_2max response to exercise training: Results from the HERITAGE Family Study. *Journal of Applied Physiology*, 87(3), 1003–1008.

Guyatt, G. H., Heyting, A., Jaeschke, R., Keller, J., Adachi, J. D., & Roberts, R. S. (1990). N of 1 randomized trials for investigating new drugs. *Controlled Clinical Trials*, 11(2), 88–100.

Guyatt, G. H., Sackett, D. L., Taylor, D. W., Chong, J., Roberts, R., & Pugsley, S. (1986). Determining optimal therapy — randomized trials in individual patients. *New England Journal of Medicine*, 314(14), 889–892.

Kravitz, R. L., Duan, N., Braslow, J., & Evidence-Based Medicine Working Group. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4), 661–687.

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine*, 8(2), 161–173.

Nikles, C. J., Clavarino, A. M., & Del Mar, C. B. (2005). Using n-of-1 trials as a clinical tool to improve prescribing. *British Journal of General Practice*, 55(512), 175–180.

Roberts, S. (2004). Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight. *Behavioral and Brain Sciences*, 27(2), 227–262.

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature*, 520(7549), 609–611.

Senn, S. (2002). *Cross-over trials in clinical research* (2nd ed.). John Wiley & Sons.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.

Senn, S. (2016). Mastering variation: Variance components and personalised medicine. *Statistics in Medicine*, 35(7), 966–977.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., … & Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5), 1079–1094.

Zucker, D. R., Schmid, C. H., McIntosh, M. W., D'Agostino, R. B., Selker, H. P., & Lau, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology*, 50(4), 401–410.

Zucker, D. R., Ruthazer, R., & Schmid, C. H. (2010). Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. *Journal of Clinical Epidemiology*, 63(12), 1312–1323.