6: Sequential Testing: ROPE vs Bayes Factor vs LOO

Understanding Divergences in Bayesian Decision Making

Job Schepens

2026-01-21

Table of contents

1 Introduction

In this example, we will simulate a scenario where we collect data sequentially ($N = 1, ..., 100$) and track how three different Bayesian decision metrics behave:

1. **ROPE (Region of Practical Equivalence)**: Does the effect magnitude matter?
2. **Bayes Factor (via Savage-Dickey)**: Is the null hypothesis ($H_0 : \beta = 0$) less likely than the alternative?
3. **LOO (Leave-One-Out CV)**: Does including the parameter improve predictive accuracy?

We will also see how **prior width** affects the Bayes Factor but has less impact on ROPE and LOO (once N is moderate).

2 Part 1: The Simulation Study

2.1 Data Generation

We generate data where there is a **real effect** (raw effect 0.12, approx $d \approx 0.45$ ($d = 0.12/0.27$)). This is a "medium" effect size.

$$SD_{total} = \sqrt{0.20^2 + 0.15^2 + 0.10^2} \approx 0.27$$

2.2 The Loop

We will loop through sample sizes. At each step, we fit three models:

1. **Null Model (H_0)**: `y ~ 1`
2. **Wide Prior Model (H_1 Wide)**: `y ~ condition`, prior `normal(0, 1.0)`.
   - **Scale Context**: Since we are analyzing **log-reaction times** (where $y = \log(RT)$), a coefficient of β roughly corresponds to a fractional change in raw seconds (e^{β}).
   - **Why this width?** A prior of $\sigma = 1.0$ implies that we consider effects of magnitude 1.0 ($e^1 \approx 2.7\times$ change) to be plausible. This is effectively "uninformative" because it is orders of magnitude larger than typical Stroop/Priming effects (which are usually 5-20%). A naive prior of `normal(0, 5)` would be even worse, implying standard deviations of $e^5 \approx 148\times$, which is physically impossible for human reaction times.
3. **Narrow Prior Model (H_1 Narrow)**: `y ~ condition`, prior `normal(0, 0.1)`.
   - **Why this width?** This is "scientifically informed" by the literature. We expect effects in the range of 5-15%, which corresponds to $\beta \approx 0.05 - 0.15$ on the log scale. A standard deviation of $\sigma = 0.1$ places 95% of the probability mass between ± 0.2 ($\pm 22\%$ effect), penalizing absurdly large effects while remaining open to robust psychological phenomena.

Deep Dive: From Milliseconds to Log-Scale (and back)

To understand why `normal(0, 1)` is "wide" and `normal(0, 0.1)` is "narrow", we must look at the transformation.

1. **Baseline**: Suppose a typical reaction time is **400 ms**.
   - Log-Scale: $\log(400) \approx 5.99$. (This explains why our Intercept prior is around 6).
2. **Effect Size (The Prior)**:
   - A coefficient β on the log scale represents a **multiplicative shift** on the raw scale: $RT_{new} = RT_{base} \times e^{\beta}$.
   - For $\beta = 0.1$: Multiplier is $e^{0.1} \approx 1.105$.

  - Effect: A **10.5% increase**.
  - Calculation: 400 ms $\times$ 1.105 $\approx$ 442 ms. (A +42 ms effect).
- For $\beta = 1.0$ (The "Wide" Prior σ): Multiplier is $e^{1.0} \approx 2.718$.
  - Effect: A **171% increase** (nearly tripling the time).
  - Calculation: 400 ms $\times$ 2.718 $\approx$ 1087 ms. (A +687 ms effect).

3. **Conclusion**: A prior of `normal(0, 1)` says "I expect the effect of the condition to potentially **triple** the reaction time". This is extremely unlikely in a standard cognitive task. A prior of `normal(0, 0.1)` says "I expect the effect to be around 10% (e.g., 40ms)", which is a reasonable effect size for a cognitive constraint.

2.3 Results Table

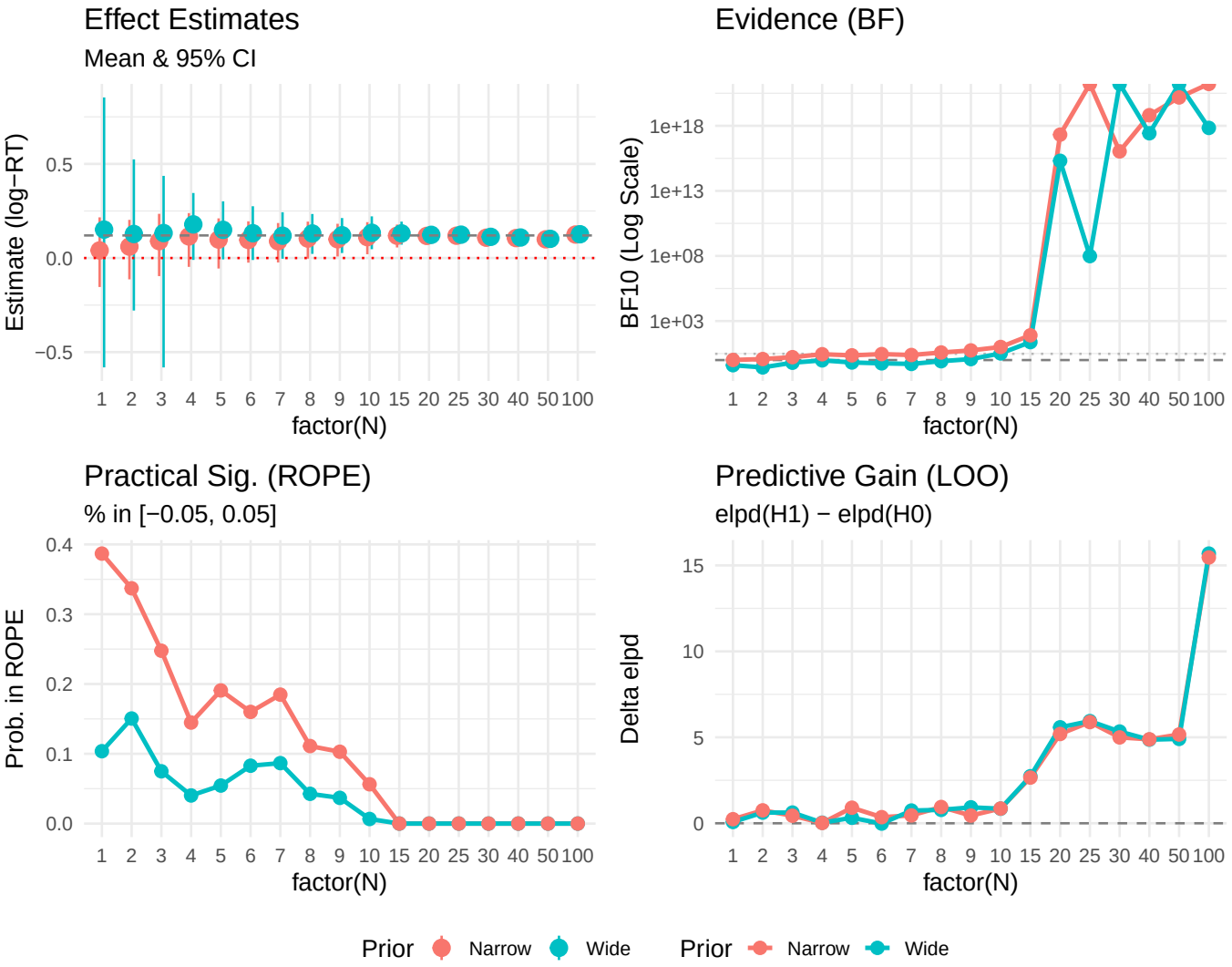
Table 1: Comparison of Decision Metrics by Sample Size and Prior Sensitivity

N	Prior	Estimate_CI	BF10	ROPE_Prob	LOO_Gain
1	Wide	0.15 [-0.58, 0.85]	4.030000e-01	0.104	0.076
1	Narrow	0.04 [-0.15, 0.22]	1.017000e+00	0.387	0.236
2	Wide	0.13 [-0.28, 0.52]	2.720000e-01	0.151	0.620
2	Narrow	0.06 [-0.11, 0.20]	1.195000e+00	0.337	0.750
3	Wide	0.13 [-0.58, 0.44]	6.250000e-01	0.075	0.622
3	Narrow	0.09 [-0.10, 0.23]	1.667000e+00	0.248	0.438
4	Wide	0.18 [-0.01, 0.35]	9.490000e-01	0.040	0.033
4	Narrow	0.11 [-0.05, 0.24]	2.805000e+00	0.145	0.017
5	Wide	0.15 [-0.01, 0.30]	6.530000e-01	0.054	0.319
5	Narrow	0.10 [-0.06, 0.21]	2.303000e+00	0.191	0.904
6	Wide	0.13 [-0.01, 0.28]	5.600000e-01	0.083	-0.018
6	Narrow	0.10 [-0.02, 0.20]	2.872000e+00	0.160	0.354
7	Wide	0.12 [-0.00, 0.24]	5.040000e-01	0.087	0.739
7	Narrow	0.09 [-0.02, 0.19]	2.444000e+00	0.185	0.457
8	Wide	0.13 [0.02, 0.23]	8.170000e-01	0.043	0.779
8	Narrow	0.10 [-0.00, 0.19]	3.818000e+00	0.111	0.946
9	Wide	0.12 [0.03, 0.21]	1.187000e+00	0.037	0.926
9	Narrow	0.10 [0.01, 0.18]	5.508000e+00	0.103	0.447
10	Wide	0.13 [0.05, 0.22]	3.236000e+00	0.007	0.847
10	Narrow	0.11 [0.02, 0.19]	1.012300e+01	0.056	0.857
15	Wide	0.13 [0.07, 0.19]	2.394000e+01	0.000	2.735
15	Narrow	0.12 [0.06, 0.18]	8.004300e+01	0.000	2.657
20	Wide	0.12 [0.08, 0.17]	2.040489e+15	0.000	5.582
20	Narrow	0.12 [0.07, 0.16]	2.093883e+17	0.000	5.193
25	Wide	0.12 [0.08, 0.16]	9.728796e+07	0.000	5.943
25	Narrow	0.12 [0.08, 0.16]	Inf	0.000	5.885
30	Wide	0.11 [0.07, 0.15]	Inf	0.000	5.335
30	Narrow	0.11 [0.07, 0.15]	1.081383e+16	0.000	4.994
40	Wide	0.11 [0.07, 0.14]	2.675363e+17	0.000	4.863
40	Narrow	0.11 [0.07, 0.14]	6.526253e+18	0.000	4.884
50	Wide	0.10 [0.07, 0.13]	Inf	0.000	4.907
50	Narrow	0.10 [0.07, 0.13]	1.511938e+20	0.000	5.163

N	Prior	Estimate_CI	BF10	ROPE_Prob	LOO_Gain
100	Wide	0.13 [0.10, 0.15]	6.898743e+17	0.000	15.687
100	Narrow	0.12 [0.10, 0.15]	Inf	0.000	15.463

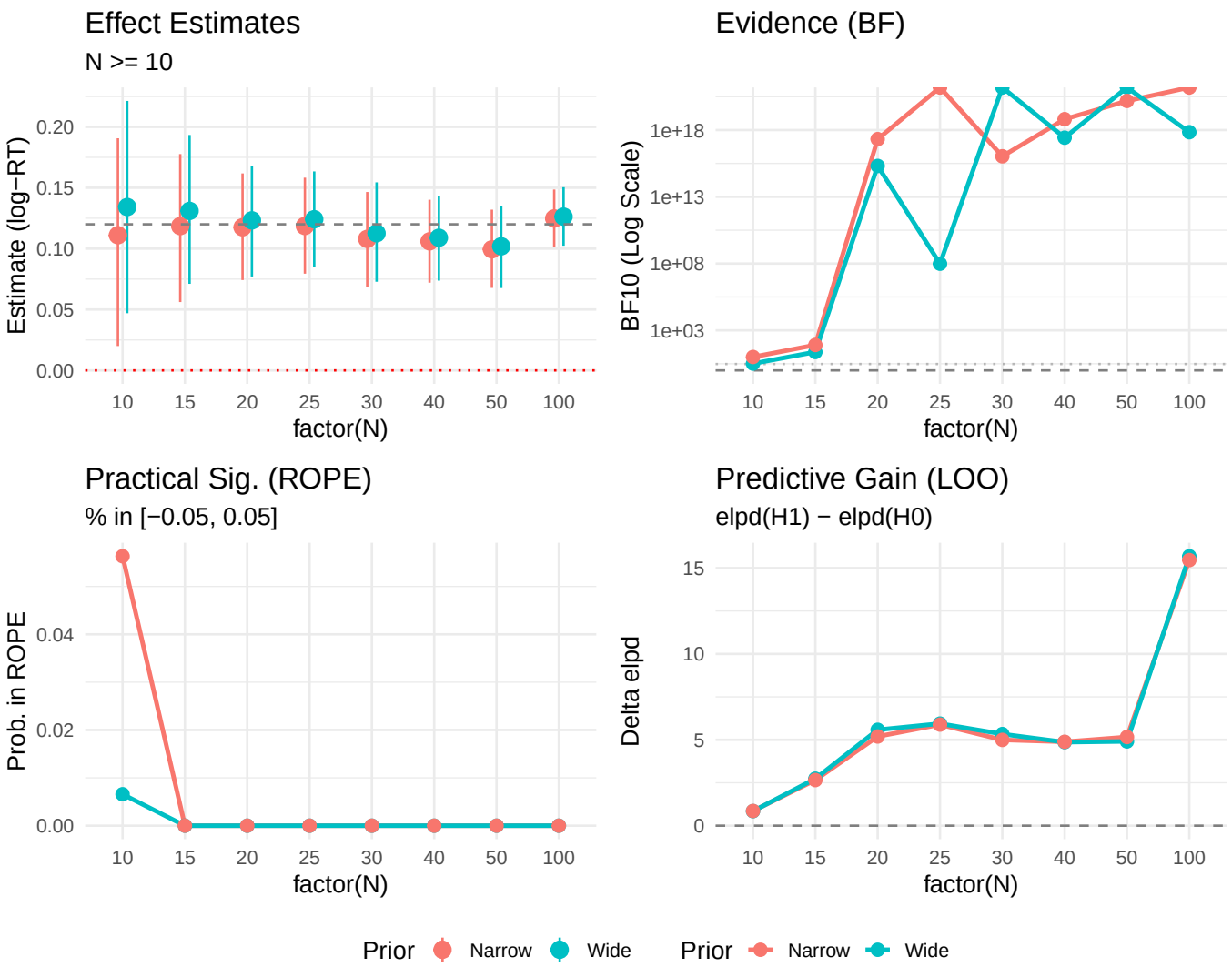
2.4 Visualization of Divergence

Let's visualize how these metrics evolve as N increases.



2.5 Visualization (Zoomed: N >= 10)

Here is the same plot, but filtering out the very small sample sizes (N < 10) to focus on the stability and convergence of the metrics.

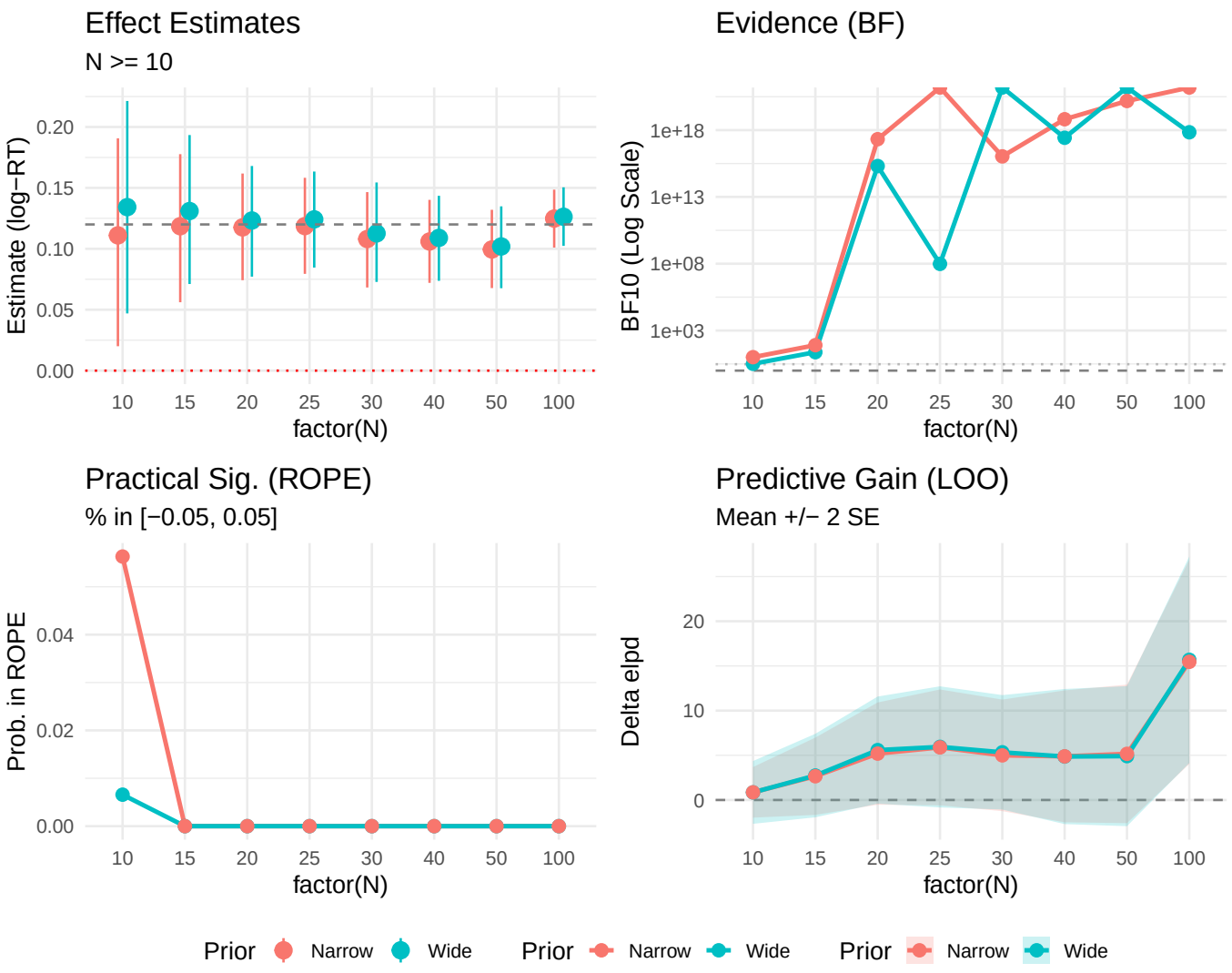


Plot Reference Lines: * **Red Dotted Line** ($y = 0$): Represents the **Null Hypothesis** (no effect). * **Gray Dashed Line** ($y = 0.12$): Represents the **True Effect Size** used in data generation. * **Gray Dashed/Dotted Lines (BF Panel)**: Thresholds for evidence ($BF = 1$ is neutral, $BF = 3$ is moderate evidence).

2.6 Visualization (Uncertainty: N >= 10)

This third version includes **uncertainty bounds** where readily available (specifically for LOO).

- **LOO Panel**: Shaded areas represent $\pm 2 \times SE_{diff}$.
- **BF/ROPE Panels**: Displayed as lines (Bayesian posterior summaries, not frequentist estimates).



3 Statistical Interpretation of LOO

3.1 What does `2 * SE` represent?

In the LOO plots above, we visualized the difference in expected log predictive density (`elpd_diff`) along with its standard error (`se_diff`).

3.1.1 1. Crossing the 0 line

If the `2 * SE` band crosses the zero line (dashed), it means **we cannot statistically distinguish the predictive performance of the two models**. * Even if the mean difference is positive (suggesting H_1 is better), the uncertainty is large enough that the difference could arguably be due to noise. * In the plot, you might see the bands overlapping 0 for smaller sample sizes, indicating that we don't yet have enough data to confidently claim the "effect" model predicts better than the null.

3.1.2 2. Why doesn't the error band get narrower?

You might expect the error bars to shrink as N increases (like standard errors of mean estimates, which scale as $1/\sqrt{N}$). **However, for ELPD, the standard error typically grows or stays constant.** * **Reason**: ELPD is a **sum** over all N data points, not an average. * $ELPD = \sum_{i=1}^{N} \log p(y_i|y_{-i})$ * As you calculate a sum over more items, the total uncertainty accumulates. * The signal (difference in ELPD) scales with N. * The noise (SE of the difference) scales roughly with $\sqrt{N}$. * **Result**: The *relative* error shrinks (the signal-to-noise ratio improves), but the *absolute* width of the `2 * SE` band on the plot will actually expand as N grows. This is a feature of summing predictive densities, not a bug.

3.2 Alternatives: Bayesian Stacking

Instead of using LOO solely for **model selection** (picking the single "best" model), a more robust Bayesian approach is **model averaging** or **stacking**.

Bayesian Stacking combines the posterior predictive distributions of multiple models using weights (w_k) that maximize the leave-one-out predictive density of the combined distribution. This avoids the "all-or-nothing" decision of selection and preserves uncertainty across models.

4 Part 3: Why do the metrics disagree?

You might notice that **Bayes Factors** (BF_{10}) and **Posterior Estimates** (95% CI) often scream "Effect!" (BF > 100, CI excludes 0) while **LOO** remains cautious (bands crossing 0 or small gains). This is not a contradiction; they are asking different questions.

4.1 1. Parameter Estimation vs. Prediction (The Core Difference)

- **Bayes Factor / Posterior Estimates**: Ask *"Is the effect non-zero?"* (In-Sample / Parameter Focus)
  - As sample size (N) grows, our certainty about the location of the *mean* parameter increases rapidly (Standard Error of Mean $\propto 1/\sqrt{N}$).
  - Even a tiny true effect (e.g., 1ms difference) becomes "highly statistically significant" with enough data. The BF effectively measures evidence for the *existence* of an effect, no matter how small.
- **LOO (Leave-One-Out)**: Asks *"Does knowing this effect help me predict the next data point?"* (Out-of-Sample / Prediction Focus)
  - Prediction is limited by the **residual noise** (σ) in individual data points.
  - **Signal-to-Noise Ratio**: If the effect size is small (e.g., shifts the mean by 0.1 SD) but the data is noisy, knowing the effect gives you almost no advantage in predicting an individual person's score compared to just guessing the grand mean.
  - Therefore, an effect can be **statistically "real"** (high BF, CI excludes 0) but **predictively "negligible"** (low LOO gain).

4.2 2. The Accumulation of Evidence

- **BF** grows exponentially with N because evidence multiplies. It's a measure of *relative* probability.
- **LOO difference** grows linearly with N (sum of log-likelihoods), but the **uncertainty** (SE) also grows (as $\approx \sqrt{N}$). This keeps the "distinguishability" lower for longer, reflecting the inherent difficulty of prediction in noisy data.

5 Part 4: Interpretation Cheat Sheet

Here is a quick reference for interpreting the values from the plots above.

5.1 1. Bayes Factor (BF_{10})

Evidence for the Alternative Hypothesis (H_1) over the Null (H_0).

BF_{10} Value	Strength of Evidence
1 - 3	**Anecdotal** (Barely worth mentioning)
3 - 10	**Moderate**
10 - 30	**Strong**
> 30	**Very Strong**

5.2 2. LOO (`elpd_diff`)

Predictive performance difference. Ratio = `|elpd_diff| / se_diff`.

`elpd_diff`	Ratio	Interpretation	Action
< 4	< 2	**Equivalent models**	Pick simpler model (Occam's razor)
4 - 10	2 - 4	**Moderate difference**	Consider model with larger elpd
> 10	> 4	**Clear winner**	Prefer model with larger elpd

5.3 3. ROPE (Region of Practical Equivalence)

Based on the 95% Highest Density Interval (HDI) of the posterior distribution relative to the ROPE (e.g., $[-0.05, 0.05]$).

- **Reject** H_0 (Effect is meaningful):
  - The entire 95% HDI is **outside** the ROPE.
- **Accept** H_0 (Effect is negligible):
  - The entire 95% HDI is **inside** the ROPE.
- **UNDECIDED** (Insufficient precision):
  - The 95% HDI **overlaps** with the ROPE.

> Note: The ROPE Probability plot shows what % of the posterior is inside the ROPE. A value of ~10-20% usually corresponds to the "Undecided" or "Reject H0" cases depending on where the bulk of the density lies.

5.4 4. Posterior Estimates (95% CI)

Estimate of the effect size (log-RT difference) with uncertainty.

- **Credible Effect**:
  - The **95% Credible Interval (CI)** (the error bar) **excludes 0**.

  - Interpretation: We are 95% confident that the true parameter is not zero.
- **Indistinguishable from Null**:
  - The **95% CI includes 0**.
  - Interpretation: Zero is a credible value for the parameter; we cannot rule out the null hypothesis.

6 Part 5: The Danger of Optional Stopping (P-Hacking)

An advantage of Bayesian methods is their robustness to **Optional Stopping**—the practice of checking your results as data comes in and deciding whether to stop or continue collecting data.

6.0.1 1. The Likelihood Principle

The reason Bayesian methods handle optional stopping gracefully is the **Likelihood Principle**: *All the evidence from the data relevant to model parameters is contained in the likelihood function.* In simple terms, **the intention of the experimenter does not change the evidence**. Whether you planned to stop at $N = 50$ or you just happened to stop there because the result looked good, the data (D) is the same, and therefore $P(D|H_0)$ and $P(D|H_1)$ are the same.

- **Frequentist P-values** violate this. They calculate the probability of observing data **as extreme or more extreme** than observed, which depends on hypothetical outcomes that did not occur ("unobserved data").
  - **The Sampling Plan Matters**: To define "what is possible", you need a sampling plan. If you plan to stop at $N = 50$, the space of possible outcomes is different than if you plan to stop "when significant".
  - **Example**: Two researchers both observe 5 Heads in a row. Researcher A planned to flip 5 times; for them, this is rare ($p < .05$). Researcher B planned to flip until they saw Tails; for them, 5 Heads is just one of many stopping points, so it is less surprising and often not significant.
  - Because p-values depend on this "Plan B" (what you *would have done* if data were different), checking results repeatedly changes the sampling plan to a sequential one, inflating the false positive rate (often $> 20\%$).
- **Bayes Factors** respect this. They simply update the odds. If H_0 is true, adding more data generally pushes the BF towards 0 (evidence *for* null). It does not "drift" toward a significance threshold in the same way p-values do (random walk vs. convergence).

6.0.2 2. Assumptions & Validity

While robust, optional stopping with Bayes Factors is not magic. It relies on: 1. **Likelihood Validity**: Your model (distribution, noise) must reasonably approximate the data-generating process. * **Garbage In, Garbage Out**: The Bayes Factor measures how much *better* one model predicts the data than another. If both models are fundamentally wrong (e.g., assuming a Normal distribution for data that is actually Log-Normal or has heavy outliers), the "evidence" is meaningless. * **Outlier Sensitivity**: If your likelihood does not account for outliers (e.g., using a pure Gaussian likelihood), a single extreme data point can dominate the likelihood calculation. The BF might sway wildly based on which model accidentally "captures" that outlier better, rather than reflecting the true effect. * **Check Your Residuals**: You must still perform posterior predictive checks. Optional stopping doesn't save you from a bad model. 2. **Prior Sensitivity**: If you use a Prior that is "too wide" (the Dilution Effect seen in our earlier plots), you artificially penalize H_1, making it harder to find evidence for an effect even if it exists. 3. **Finite Horizons**: In theory, if you sample *forever*, a Bayes Factor can transiently cross a threshold (e.g., $BF > 3$) even if H_0 is true, though the probability of this is much lower than for p-values.

6.0.3 3. Practical Rules for Psycholinguistics

For real experiments, **"continuous monitoring" is valid**—meaning you can check your Bayes Factor after every subject without "breaking" the stats. * **Why?** Because the Bayes Factor at $N = 50$ answers: *"Given the data I represent right now, what are the relative odds?"* It doesn't care if you also looked at $N = 49$. * **Contrast**: If you check a p-value at $N = 50$, you are technically asking: *"What is the probability of this data occurring given that I planned to stop at N=50?"*. If you peeked at $N = 49$, you weren't planning to stop at 50 (you were planning to stop at 49 OR 50), so the p-value calculation is wrong.

However, even though it IS valid, **structured approaches (Sequential Bayes Factor Design)** are better for practical planning:

- **Define Thresholds**: Pre-register a strict evidence threshold (e.g., Stop if $BF_{10} > 10$ or $BF_{10} < 1/10$). $BF > 3$ is often considered "anecdotal" or too weak for stopping in high-noise fields like linguistics.
- **Set a Maximum N**: Resources are finite. Define a hard stopping point (e.g., $N = 100$) where you stop regardless of evidence (usually concluding "Inconclusive").
- **Minimum N**: Collect a decent initial sample (e.g., $N = 20$) before the first check to stabilize the priors and avoid early-stage noise.

6.1 Simulation: Testing Repeatedly

We simulate 500 experiments. In each, we collect data sequentially up to $N = 100$, checking the results every 5 subjects.

- **H1 True**: Effect = 0.5 (Moderate effect).
- **H0 True**: Effect = 0.
- **Stopping Rules**:
  1. **P-Value**: Stop if $p < .05$.
  2. **Bayes Factor**: Stop if $BF_{10} > 3$ (Moderate Evidence) or $BF_{10} > 10$ (Strong Evidence).

6.2 Visualization: Error vs. Power

How to Read this Plot:

This visualization demonstrates the fundamental trade-off in sequential testing:

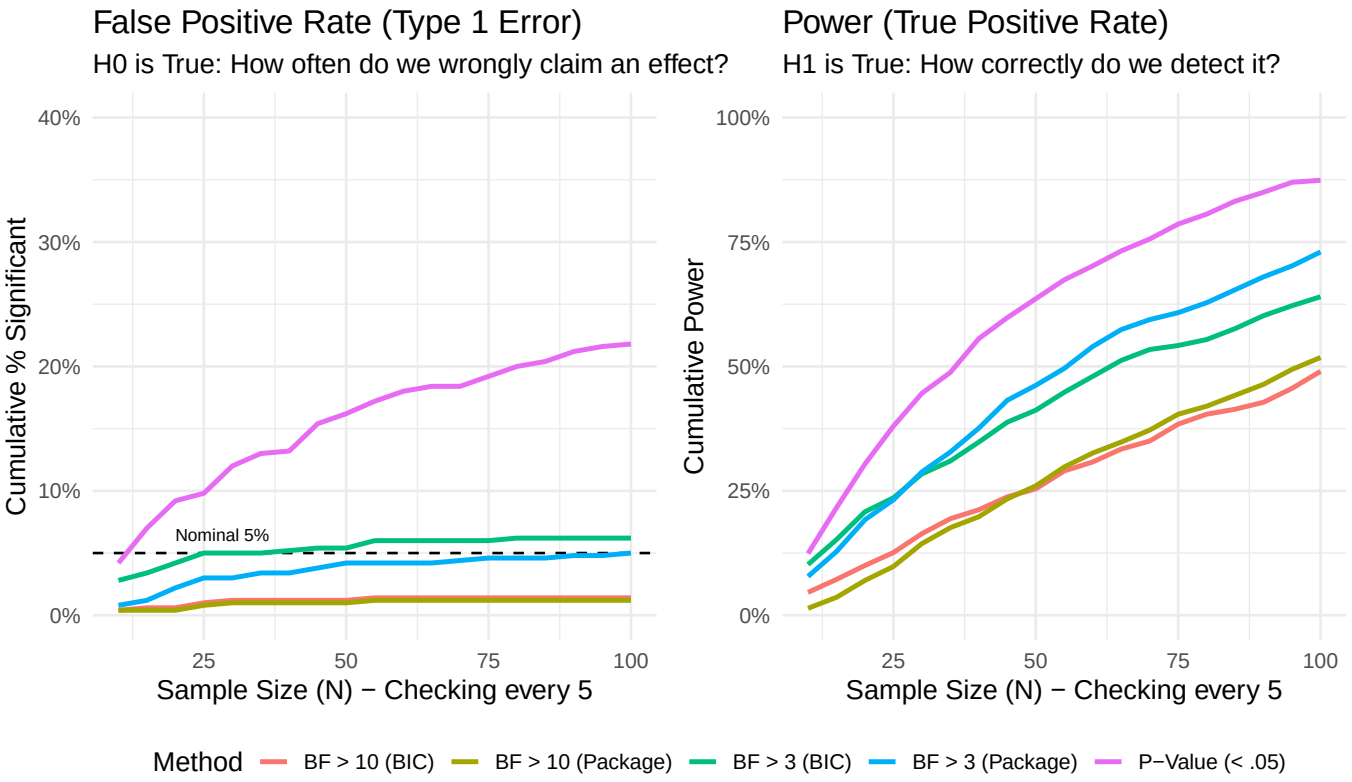
- **X-Axis (Sample Size)**: Represents the progress of the experiment as we add more subjects and "peek" at the data (every 5 subjects).
- **Y-Axis (Cumulative Rate)**: The percentage of experiments that stopped early because they found a "significant" result.

The Panels:

1. **Left Column (False Positive Rate)**:
   - **Scenario**: **H0 is True** (No real effect). Ideally, these lines should be flat at 0% (or max 5%).
   - **Result**: The Red Line (**P-value**) climbs steadily. This is **P-Hacking**: by looking repeatedly, we inevitably find "lucky" noise and call it an effect. The Green/Purple Lines (**Bayes**) stay flat, showing they are robust to this error.
2. **Right Column (Power)**:
   - **Scenario**: **H1 is True** (Real effect exists). Ideally, these lines should go to 100% quickly.

- **Result**: The Red Line rises fastest—P-values are "trigger happy" and detect effects quickly. The Bayesian lines rise more slowly (the curve). This is the cost of safety: to be sure it's not a false alarm, Bayes Factors require **more data** to reach the evidential threshold ($BF > 10$).

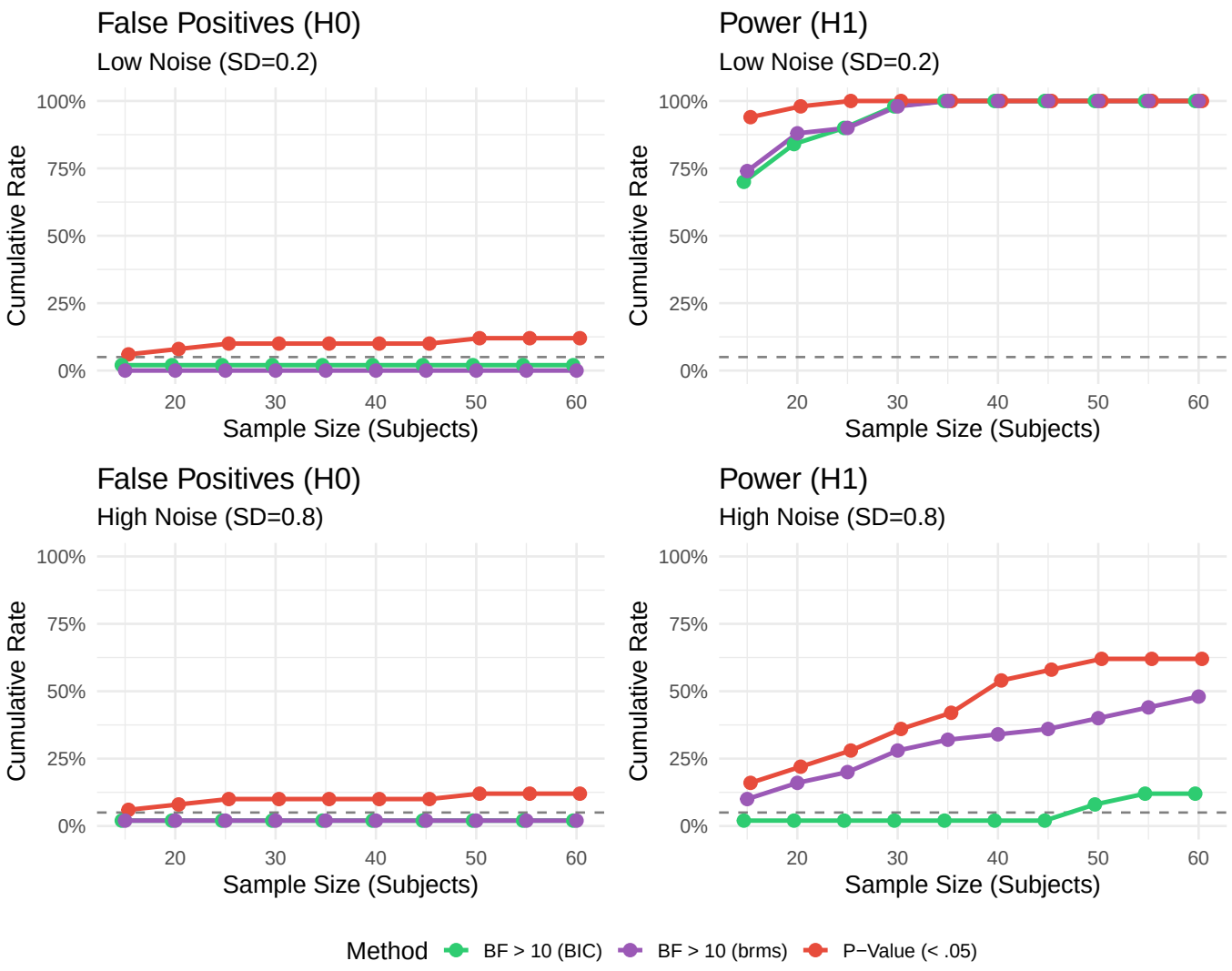
Conclusion: P-values offer high power but dangerous false positives when peeking. Bayes Factors offer safety against false positives but require larger samples to reach strict thresholds.



6.3 Simulation 2: Psycholinguistic Experiment (Mixed Models)

In this second simulation, we replicate the structure of a real psycholinguistic experiment with **Subjects** and **Items**. This is computationally expensive, so we run fewer simulations ($Sims = 20$) but use the appropriate **Linear Mixed Models (LMM)**.

- **Structure**: 20 Items, Random Intercepts for Subjects and Items.
- **Sequential Test**: We add subjects in batches (N=10 to 60, step 5).
- **Comparison**:
  1. **Likelihood Ratio Test ($p < .05$)**: `lmer(y ~ cond)` vs `lmer(y ~ 1)`.
  2. **Bayes Factor (BIC Approx)**: Using the BIC differences of the LMMs.



6.4 Simulation Results Table

Table 4: Sequential Testing Results: Binary Condition Simulation

Noise	Scenario	check_n	P-Value (<.05)	Bayes Factor (BIC) > 10	Bayes Factor (brms) > 10
High Noise (SD=0.8)	H0 (No Effect)	15	6.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	20	8.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	25	10.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	30	10.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	35	10.0%	2.0%	2.0%

Noise	Scenario	check_n	P-Value (<.05)	Bayes Factor (BIC) > 10	Bayes Factor (brms) > 10
High Noise (SD=0.8)	H0 (No Effect)	40	10.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	45	10.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	50	12.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	55	12.0%	2.0%	2.0%
High Noise (SD=0.8)	H0 (No Effect)	60	12.0%	2.0%	2.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	15	16.0%	2.0%	10.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	20	22.0%	2.0%	16.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	25	28.0%	2.0%	20.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	30	36.0%	2.0%	28.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	35	42.0%	2.0%	32.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	40	54.0%	2.0%	34.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	45	58.0%	2.0%	36.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	50	62.0%	8.0%	40.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	55	62.0%	12.0%	44.0%
High Noise (SD=0.8)	H1 (Effect d=0.06)	60	62.0%	12.0%	48.0%
Low Noise (SD=0.2)	H0 (No Effect)	15	6.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	20	8.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	25	10.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	30	10.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	35	10.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	40	10.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	45	10.0%	2.0%	0.0%

Noise	Scenario	check_n	P-Value (<.05)	Bayes Factor (BIC) > 10	Bayes Factor (brms) > 10
Low Noise (SD=0.2)	H0 (No Effect)	50	12.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	55	12.0%	2.0%	0.0%
Low Noise (SD=0.2)	H0 (No Effect)	60	12.0%	2.0%	0.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	15	94.0%	70.0%	74.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	20	98.0%	84.0%	88.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	25	100.0%	90.0%	90.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	30	100.0%	98.0%	98.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	35	100.0%	100.0%	100.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	40	100.0%	100.0%	100.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	45	100.0%	100.0%	100.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	50	100.0%	100.0%	100.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	55	100.0%	100.0%	100.0%
Low Noise (SD=0.2)	H1 (Effect d=0.06)	60	100.0%	100.0%	100.0%