Bayesian borrowing for meaningful change in small samples

Bayram, M[1], Fromy, P[1], Terluin, B[2]

[1] SeeingTheta [2] Amsterdam UMC, Amsterdam, The Netherlands

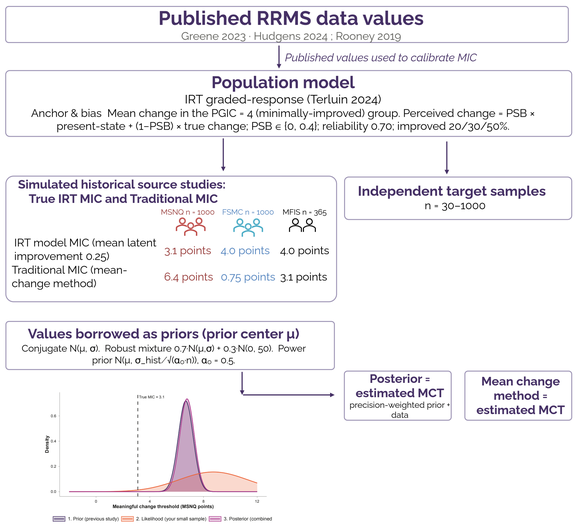
Introduction

Background. Meaningful change thresholds (MCTs) make PRO scores interpretable, but need sample sizes rare-disease & CNS studies rarely reach.

Objective. Can Bayesian historical borrowing sharpen small-sample MCT estimation, and how does the benefit depend on the borrowed source?

Methods

Design. IRT-simulated MSNQ/FSMC data; N = 30–1000; traditional mean-change vs. conjugate, power & robust-mixture Bayesian updates; 500 reps.



Results

Borrowing reduces small-sample size estimation error: benefits are greatest at smaller Ns and diminish as the sample grows.

Relevance of the source (proximity of the borrowed threshold to the target) and the quality of cross-instrument/cross-study linking.

Borrowing sharpens precision and not validity: the anchor-based method's inherent bias persists unless the prior is derived from a less biased source.

Discussion

→ Borrow at small N, prioritize robust priors to guard against a potentially poor source.

Small Sample Meaningful Change Estimation Is More Precise When You Borrow From Prior Work

But Bias From Previous Methods Can Carry Over Too

Take a picture for more details

SeeingTheta

Same-instrument borrowing

MSNQ prior on an MSNQ target

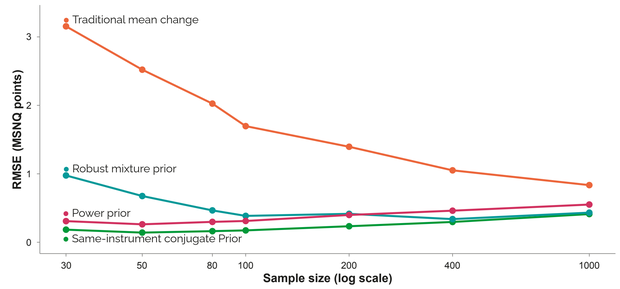


Figure 1: With only 50 patients, the usual method lands far from the answer a large study would give. Borrowing a threshold from an earlier study of the same questionnaire cut that error by about 94%, while still keeping an honest 95% uncertainty range.

Cross-instrument, same construct borrowing

FSMC prior on an MSNQ target

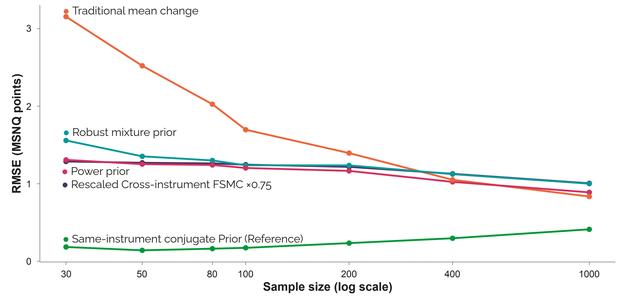


Figure 2: Borrowing from a different but related questionnaire helps, but a range-rescaled cross-instrument prior helps, but does not reach the same-instrument level.

How far off the "true" MIC?

In a scenario where the truth is set (N = 50)

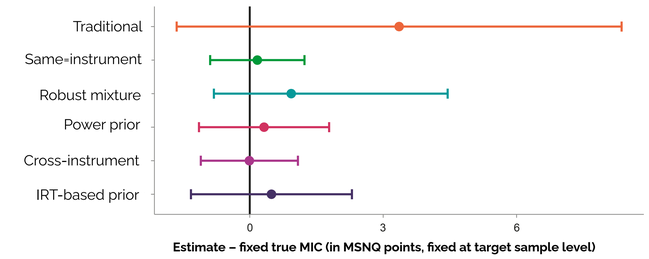
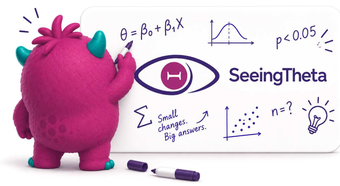


Figure 3: With the truth fixed in advance and every prior aimed at it, borrowing methods sit near zero., whereas the traditional method remains biased.

Contact
Mariam Bayram
maryam@seeingtheta.com
seeingtheta.com

References
1. Greene M, et al. Journal of Patient-Reported Outcomes, 2023; 7(1), 61.
2. Hudgens S, et al. Value in Health. 2024; 27(5), 598-606.
3. Ibrahim, J. G, & Chen, M. H. Statistical Science, 2000; 46-60.
4. Rooney S, et al. Multiple Sclerosis and Related disorders, 2019; 35, 158-163.
5. Terluin B., et al. Quality of Life Research, 2024; 33(11), 2963-2973.



The four borrowing scenarios

Same estimand and same update throughout; only the prior's origin (and, across instruments, a linear range-ratio rescale of the borrowed value) changes.

Scenario	Target	Prior source	Scale	Benchmark
A · Same-instrument	new MSNQ	historical MSNQ traditional MIC	—	large-sample MSNQ traditional
B · Cross-instrument	new MSNQ	historical FSMC MIC (raw & ×0.75)	0.75	large-sample MSNQ traditional
C · IRT ceiling	new MSNQ	IRT true MIC (SD 1.0)	—	IRT true MIC (ground truth)
D · Cross-study	new FSMC	independent MFIS MIC (×0.952)	0.952	large-sample FSMC traditional

Simulation engine

The IRT engine generates data; estimation is separate. Each replication:
1. Latent traits — correlated baseline & change (r = −0.30); change shifted so a set % improve; follow-up = baseline + change.
2. IRT scoring — graded-response model (each instrument's item parameters) → item responses → totals; observed change = follow-up − baseline.
3. Anchor (PGIC) — present-state-bias-weighted perceived change + noise (reliability 0.70) → 5-level PGIC; PGIC = 4 = "minimally improved."
4. Two MCTs — Traditional = mean change among PGIC = 4; True = read from the IRT model (known only to the simulation).

Parameters

Instruments (historical N)

MSNQ 15 items · 0–60 · true MIC ≈ 3.1 · N = 1000
FSMC 20 items · 20–100 · true MIC ≈ 4.0 · N = 1000
MFIS 21 items · 0–84 · true MIC ≈ 4 · N = 365

Simulation grid (new samples)

Sample sizes 30, 50, 80, 100, 200, 400, 1000 (log axis)
Proportion improved 0.20 / 0.30 / 0.50
Present-state bias (PSB) 0 / 0.4
Mean individual MIC / SD 0.25 / 0.15 (latent units)
Baseline–change correlation −0.30 · Transition reliability 0.70
200 replications per cell

Metrics & benchmarks

Bias — mean(estimate − benchmark).
RMSE — root-mean-square of that error; the headline metric.
Coverage — fraction of 95% intervals containing the benchmark; the calibration analogue of a false-positive rate.
Interval width — mean width of the 95% interval; precision.

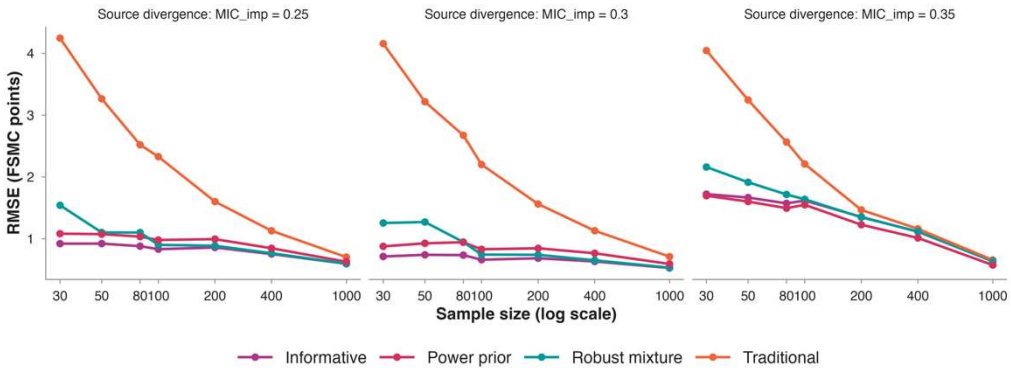
Two benchmarks. Primary = the large-sample *traditional* MIC — the estimand practitioners report, so "error" = distance from what a large study using the same method would find. Secondary = the IRT *true* MIC (Scenario C), the only place the mean-change method's own bias shows.

Prior forms

vague N(0,50) · weakly informative · correctly informative N(MIC, SE) · overconfident (too-tight SD) · misspecified (2× MIC) · robust mixture (0.7 informative + 0.3 vague) · power prior ($\alpha_0 = 0.5$).

Results not shown on the poster

Cross-study, different population (mean improvement different)



Coverage & interval width. Borrowing narrows the 95% interval most at small N while coverage stays near nominal 95%

