48 Predictions and Counting: How to Test a Theory

Jon Smirl

2026-03-01

The Replication Problem

Economics has a credibility problem. Studies that report clean, statistically significant results often fail to replicate when other researchers try them with new data (Ioannidis2005). A careful audit of top economics journals found systematic bunching of p-values just below the 0.05 threshold – the telltale signature of researchers trying many specifications until one "works" (Brodeur2020). When 21 experimental results from *Nature* and *Science* were subjected to independent replication, only 62% held up (CamererDreber2018).

The root cause is well understood: when you test a theory *after* seeing the data, it is too easy to find patterns that look real but are not. The theory adapts to fit the noise. The solution is also well understood, at least in principle: commit to your predictions *before* looking at the data. If the theory is specific enough to be wrong, and it survives honest testing, that is meaningful evidence.

The CES framework takes this approach seriously. The theoretical articles — from *Emergent CES* through *information friction*, *dynamics*, and *hierarchical architecture* — derive 48 falsifiable predictions from the mathematics *before* any empirical testing begins. The empirical program then tests every single one, reports the results in full, and discusses the failures openly.

What Counts as a Prediction?

A good prediction has three properties: (1) **Specificity** – it says something precise enough to be wrong; (2) **Independence** – it is not used to calibrate the model; (3) **Uniqueness** – no standard alternative model generates the same prediction. "The economy will fluctuate" fails all three. "Sectors with lower substitution elasticity σ will enter recessions earlier" passes all three.

The 48 predictions in the CES inventory span a range. Some are strong (uniquely implied, no standard alternative). Some are moderate (shared with other models but with a specific quantitative twist). The *scorecard* tracks all of them.

The Current Scorecard

As of March 2026, the 48 predictions break down as follows:

Category	Count	Meaning
Confirmed / Consistent	23	Statistical support at conventional levels

Category	Count	Meaning
Directional / Ambiguous	9	Predicted sign holds but significance is marginal or data is limited
Not confirmed / Inconsistent	2	Predicted pattern not detected or contradicted
Pending	14	Not yet tested or awaiting data

The honest ratio – 23 out of 34 tested predictions confirmed, with 2 failures – is the kind of scorecard that lets readers judge for themselves.

Five Highlights

1. Damping Cancellation (158 Countries)

The CES hierarchy predicts that tightening regulation at one level of the economy should produce a *transient* effect that decays over time, not a permanent shift. This is the *damping cancellation* prediction: regulatory shocks are absorbed by the system's own adjustment dynamics.

Testing this requires a natural experiment. The Basel III banking reforms (2010-2013) provide one: a large, synchronized regulatory shock applied across many countries at roughly the same time. Using Barth-Caprio-Levine banking regulation indices (Barth2013) across 158 countries and five survey waves, the difference-in-differences estimate is $\hat{\beta}_3 = -0.003$ with $p = 0.95$. That near-zero coefficient is *exactly* the prediction: the regulatory shock had no detectable long-run effect on growth, because the system damped it out.

This is a case where the absence of an effect is the finding. Most regulation studies look for persistent impacts. The CES framework says you should not find them – and you do not.

2. Dispersion as a Leading Indicator

The *dispersion indicator test* checks whether cross-segment variation in semiconductor shipments leads aggregate regime shifts. The theory predicts that when the dispersion across subsectors starts rising, a downturn is coming – because rising dispersion signals that the balanced allocation supporting the CES superadditivity premium is breaking down.

The VAR impulse response peaks at a 3-quarter lead: dispersion today predicts aggregate semiconductor growth three quarters from now. The mechanism works because complementary inputs amplify imbalances. When one semiconductor segment stumbles, the CES aggregate penalizes the resulting asymmetry more than a linear model would.

3. Business Cycle Ordering by ρ

Different manufacturing sectors enter recessions at different times. The CES framework predicts the order: sectors with lower σ (stronger complementarity) should enter recessions *first*, because their steeper CES curvature makes them more sensitive to the common shock.

The *ordering test* confirms this with Kendall $\tau = -0.40$ ($p < 0.001$) and a slope of -14.65 months per unit ρ. Sectors with strong complementarity lead the downturn by over a year relative to sectors

with substitutable inputs. The cross-sectional regression explains $R^2 = 0.51$ of the variation in recession entry timing – from a single parameter.

4. CPI Dispersion Leads Turning Points

The *CPI dispersion test* is one of the cleanest results. The theory says that when the cross-category dispersion of CPI components rises, an aggregate turning point is approaching. In 18 out of 18 observed turning points, the dispersion measure leads the aggregate, with a median lead of 12 months and a correlation of $r = +0.37$.

Eighteen for eighteen is hard to dismiss as coincidence. The mechanism is the same as the semiconductor dispersion test: CES curvature means that rising imbalance across components erodes the aggregate before the aggregate itself moves.

5. Overidentification: Three Roads to ρ

Perhaps the most important test is about internal consistency. The substitution parameter ρ can be estimated three independent ways: NLS (fitting the production function directly), variance filtering (idiosyncratic-to-aggregate variance ratio), and equicorrelation (cross-sectional correlation structure). If ρ is a real structural parameter, these three methods should agree.

The Hansen-style overidentification statistic is $\chi^2 = 6.37$ with $p = 0.27$ – comfortably above the rejection threshold. This is like measuring the speed of light three different ways and getting the same answer. If the three estimates of ρ agree, ρ is probably real.

What the Failures Reveal

The two inconsistent results are discussed openly because they are informative:

V-shape ρ at Perez turning points. The theory predicts that the aggregate substitution parameter should trace a V-shape across major technological transitions. The data shows no V-shape ($p = 1.000$). The diagnosis is *proxy failure* – the available time series are too aggregated to isolate turning-point dynamics from the long-run trend.

Beveridge curve slope vs. task complementarity. The theory predicts that industries with more complementary tasks should have steeper Beveridge curves. The estimated slope has the wrong sign ($p = 0.16$). This suggests that production complementarity may not map directly to labor market matching frictions – a genuine model boundary.

Neither failure is statistically significant. Both point to places where the theory's assumptions stop holding, rather than fundamental contradictions.

Why This Approach Matters

Most economic theories are tested with a handful of carefully chosen empirical exercises. The CES framework inverts this: derive *everything* the theory implies, test *all* of it, and let the scorecard speak for itself.

This matters for three reasons:

Transparency. Readers can see exactly which predictions work and which do not. There is no file drawer of suppressed failures.

Discipline. Pre-specifying predictions prevents the theory from being retrofitted to match the data. Every prediction in the inventory was derived from the mathematics before any test was run.

Cumulation. As new data becomes available, the 14 pending predictions will be tested and added to the scorecard. The theory gets more constrained over time, not less.

The Lean formalization (0 sorry, 0 unjustified axioms across 97 files) provides a further layer of discipline: every mathematical step from CES axioms to empirical predictions is machine-verified. There is no gap between what the theory says and what it actually implies.

A theory that cannot be wrong is not useful. A theory that commits to 48 ways it could be wrong, tests 34 of them, and survives 23 with 2 honest failures – that is a theory worth taking seriously.

References