More variables raise prediction skill, yet most targets stall far below their noise ceiling

A preregistered test across trees, mood, music ratings, forecasters and a similarity task, with shuffled-target controls and two toy systems with known answers

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F47 in the findings ledger. Strong signal (tier B*).

Abstract

A common hope says that measuring enough things about a system lets you predict it. We tested that hope on public data against the best score each target's own measurement noise allows. Variables were added in random orders and the model refitted and scored on held-out cases, for six targets with a measurable ceiling: tree growth, next-beep mood, music valence and arousal ratings, forecaster accuracy and a similarity task. Held-out skill rose with added variables in every target and beat a shuffled-target control in every target. Only the similarity task, whose variables were learned from the same task, reached its ceiling (0.985 of the way); mood, valence, arousal and forecaster accuracy stopped at 0.65, 0.54, 0.46 and 0.18, and the tree ceiling estimator failed. The largest gains came from a different kind of variable: past accuracy lifted forecasters from 0.131 to 0.360 and prior growth lifted trees from 0.396 to 0.639. Both preregistered forecasts for the claim failed. Two toys with every variable known set hard limits: chaos returns forecast time only logarithmically in precision, and a published forecast that people react to can defeat itself.

1 What this means for a reader

Adding more of the same kind of measurement helps, then stops helping long before the data's own noise would stop it. The gap closes when something different arrives, most often a record of how the same thing behaved before. If a team says it can predict an outcome once it measures enough, ask how close its current model sits to the outcome's measurement noise, and what kind of variable it has not yet measured.

2 Introduction and motivation

How predictable an outcome is, and how much of that predictability a model captures, are two separate questions. Salganik and colleagues [1] gave 160 teams a rich birth-cohort dataset and asked them to predict six life outcomes; the best predictions were only slightly better than a simple benchmark, and error depended more on which family was predicted than on the method. That study measured how far models got. It did not set the result against a ceiling from the outcome's own noise.

In neuroscience the noise ceiling is routine: the best held-out score any model could reach on a target, given how far repeated measurements of that same target disagree. Lage-Castellanos and colleagues [2] compare ways to estimate it for brain responses. Hebart and colleagues [3] reported that their similarity model explained most of the explainable variance in human odd-one-out choices, against a ceiling set by repeated triplets.

For forecasters, Himmelstein, Atanasov and Budescu [5] found that psychometric traits predict future accuracy when little track record exists and that past accuracy takes over as it accumulates. That points to one kind of variable that may close a gap others leave open.

We asked whether held-out skill reaches the noise ceiling when enough measured variables are added, across unrelated domains in one design. Searches run on 9 November 2026 found the work above, and a preprint

on the PISA 2022 school assessment [10] that reads predictive R2 against a ceiling set by measurement noise inside one domain. A search for a cross-domain comparison of skill curves against noise ceilings returned nothing closer; that is what the search returned, and it does not show that no such study exists. The decision this informs is where effort goes when a prediction stalls: more of the same variables, a different kind of variable, or acceptance of a hard limit.

3 Preregistration and design

The preregistration, a dated plan hash-locked before any result is seen, was written on 29 October 2026 before any model in the study was fitted to real data. Data caches were built first and printed row counts only. Several of the datasets had been used in earlier studies of ours, which the plan states so that it is not mistaken for blindness.

The plan fixed the targets, variables, splits, models, ceilings and the rule for each verdict, the pass or fail of one preregistered rule set read off by code. The predictions were:

- P1, ceiling reached: a target is at ceiling when its all-variable held-out skill above chance is at least 0.8 of its ceiling above chance. Pass if at least 4 of the 6 targets with a ceiling are at ceiling.
- P2, unmeasured variables: a target's plateau moves when its named strong variable adds at least 0.02 skill over all base variables and the result exceeds the upper 97.5th percentile of the fitted plateau. Pass if at least 3 of 7 targets move.
- P3, chaos: in the Lorenz-63 toy, the forecast horizon is linear in the log of precision (linear R2 at least 0.98), its slope is within 25 percent of theory, and the last three decades gain no more than the first three.
- P4, reflexivity: five rules on a toy crowd, listed in Appendix A.
- P5, overfitting control: every target beats its shuffled-target control, and a deliberately overfitted fit behaves as expected (Appendix A).

The plan also held an identification test on a private image collection. That test, and the seventh P2 target built on the same collection, are left out of this paper; the P2 count below includes that target because the verdict was computed over seven.

The corrected rerun

The first run, on 29 October 2026, gave forecaster accuracy a held-out R2 of -1.686. The cause was a pipeline fault: 84 of the 1,349 forecaster variables were nearly constant among the training forecasters, so their scaled values in held-out forecasters became enormous. A second preregistration file, hashed after the first run had been read, made one change: every scaled variable is clipped to the range -5 to 5. Domains, splits, seed, models and verdict rules stayed the same. The second run is the primary result. It is a corrected rerun and not a blind one. Outside forecasters, the two runs agree within 0.01.

4 Data and provenance

Five public sources supply six targets with a ceiling and one without:

- Tree growth and tree death: the USDA Forest Service Forest Inventory and Analysis [8], remeasured trees in Washington and Oregon, a US government dataset. Growth is annual diameter growth of survivors from the second to the third measurement (73,553 trees); death is death between them (90,427 trees, 18.6 percent died). 42 variables measured at the second visit cover size, crown, damage, site, stand, location and species.
- Mood: openESM dataset 0062 [9], experience sampling with repeated short questionnaires (beeps) during the day, licensed CC BY 4.0. Target: rated happiness at the next beep of the same day, from 31 items at the current beep (8,141 cases).
- Music valence and arousal: DEAM static ratings [6], 1,744 clips, licensed CC BY-NC (non-commercial use

with attribution); this paper reports statistics only. 22 variables: sound measures, genre flags and set flags.

- Forecaster accuracy: Good Judgment Project data [7], licensed CC0. Target: each forecaster's mean per-question z-scored Brier score, the mean squared gap between a probability forecast and the 0 or 1 outcome, on the even half of their questions (2,120 forecasters with at least 10 scored questions in each half). 1,349 individual-difference variables.
- Similarity triplets: THINGS odd-one-out choices with the 66 published object dimensions [4], used under the publisher's terms. In each triplet a person picks the image least like the other two; chance is one in three.

5 Methods

Curves

For each target the variables were added in 8 random orders and one informed order (by absolute training-set correlation with the target), at about 14 counts from one variable to all of them. At each count the model was refitted on training cases and scored on held-out cases. Held-out skill is R2 for numeric targets, AUC (the chance a random positive case is ranked above a random negative one) for tree death, and accuracy for triplets.

Models were ridge regression with a leave-one-out penalty for numeric targets, L2 logistic regression for tree death and a conditional logit for triplets. Medians, scaling and clipping came from the training cases only. Held-out cases were whole plots for trees, each person's last 30 percent of beeps for mood, 30 percent of artists for music, 30 percent of forecasters, and between-subject repeated triplets for the similarity task.

Ceilings and shares

Each ceiling comes from the target's own repeats:

- Mood, by parallel forms: the correlation of happy with cheerful at the same beep, over 11,472 beeps. If the two items differ in what they measure, this is a lower bound, which makes the ceiling easier to reach.
- Forecasters, by split-half reliability: the correlation of accuracy on odd and even questions.
- Similarity: leave-one-out majority agreement on 1,000 repeated triplets (36,401 trials).
- Music, from rater spread, with sd the rater standard deviation of a clip, n the number of raters and m the clip's mean rating over clips:

$$c = 1 - \frac{mean(sd^2/n)}{var(m)} \tag{1}$$

The number of raters was not stated in the files we had; $n = 10$ was assumed, with $n = 5$ reported beside it.

- Tree growth: the error variance of a remeasured increment was estimated from the shrinking increments of survivors, on the assumption that true growth is never negative.

The share of ceiling puts targets on one scale, with s the held-out skill, s_0 its chance level and c the ceiling:

$$q = \frac{s - s_0}{c - s_0} \tag{2}$$

For the four numeric targets with a valid ceiling, one minus R2 splits into measurement noise, $1 - c$, and reliable variance the variables do not reach, $c - s$.

Plateau fit

A saturating curve was fitted to skill above chance against the number of variables k:

$$s(k) - s_0 = \frac{ak}{k + h} \tag{3}$$

It was fitted to the mean random-order curve, and to each of the 8 orders separately; the spread of a over the orders gives the plateau's interval. The fit bound was 2.0, and for curves that had not flattened the upper end

of the interval reached it.

Strong variables

Each target had a strong variable named before the run, of a different kind from the base set: prior diameter growth for trees, the person's mean training happiness for mood, the other dimension's human rating for music, and accuracy on the other half of the questions for forecasters.

Toys with known answers

The Lorenz-63 system was run with an exact model from 100 states on its attractor, with the starting state measured with error δ per coordinate from 10^{-1} to 10^{-12}. The forecast horizon is the first time the error norm exceeds 1. Theory gives the slope per tenfold gain in precision as

$$\frac{\ln 10}{\lambda} \approx 2.54 \tag{4}$$

with $\lambda = 0.906$ the largest Lyapunov exponent.

The crowd toy tests a reflexive forecast, one that the forecast population reads and reacts to. Each of N members attends when $u_i + \beta(c - F) > 0$, where u_i is the member's own taste, c a fixed threshold and F the published forecast of the attending share; β sets how strongly members react. The forecaster measures k members' tastes exactly. An unpublished forecast is never seen; a naive one is the unpublished forecast published anyway; a fixed-point forecast solves for the share that would attend given the forecast itself.

6 Controls and planted checks

Every target was refitted on shuffled training labels 20 times and scored on the real held-out labels. This control removes the link between variables and target. The 95th percentile of those 20 draws, the top of the null distribution, is the bar each real score must clear. The similarity control starts its weights at 1, so a fit that stops early keeps the raw embedding similarity, and its bar sits above chance for that reason.

The Lorenz toy is a planted check for the chaos limit, since its answer is known from theory. A deliberate overfit forms a second check: 150 forecasters fitted by plain least squares with up to all 1,349 variables, which should fit the training data almost perfectly and fail held out.

The tree-growth ceiling estimator failed. It returned -3.96, because a few gross shrinkages dominate the mean of squared negative increments. It is kept as preregistered, and the tree target counts as not at ceiling.

7 Results

Held-out skill rose with added variables in every target and beat a shuffled-target control in every target. The highest control bar among the numeric targets was 0.013 R2; tree death's was 0.531 AUC against 0.686, and the similarity task's was 0.439 accuracy against 0.642 on its validation triplets.

Table 1 and Figure 1 give the curves against their ceilings. Shares follow eq. 2; the music ceilings assume 10 raters; the last column is the shuffled control's 95th percentile. Only the similarity task, whose variables were learned from the same task, reached its ceiling (0.985 of the way); mood, valence, arousal and forecaster accuracy stopped at 0.65, 0.54, 0.46 and 0.18, and the tree ceiling estimator failed.

For mood, the first informed variable, current happiness, gives 0.501 of the 0.521. For forecasters the curve has flattened: the plateau fit gives 0.137 with interval 0.118 to 0.155, far below the ceiling of 0.714. With 5 raters assumed instead of 10, the music ceilings fall to 0.642 for valence and 0.725 for arousal, so the music shares depend on a count we do not have.

Figure 2 shows where the variance left after the full model sits. Outside the similarity task, the larger part is

Table 1. Held-out skill with all variables against each target's ceiling, second run.

Target	Variables	Skill	Ceiling	Share	With strong	Shuffled 95th
Similarity triplets (accuracy)	66	0.655	0.660	0.985	none named	0.439
Mood at next beep (R2)	31	0.521	0.798	0.653	0.553	0.005
Music valence (R2)	22	0.445	0.821	0.542	0.527	0.005
Music arousal (R2)	22	0.393	0.862	0.455	0.484	0.013
Forecaster accuracy (R2)	1,349	0.131	0.714	0.183	0.360	0.004
Tree growth (R2)	42	0.396	failed	none	0.639	0.005
Tree death (AUC)	42	0.686	none	none	0.702	0.531

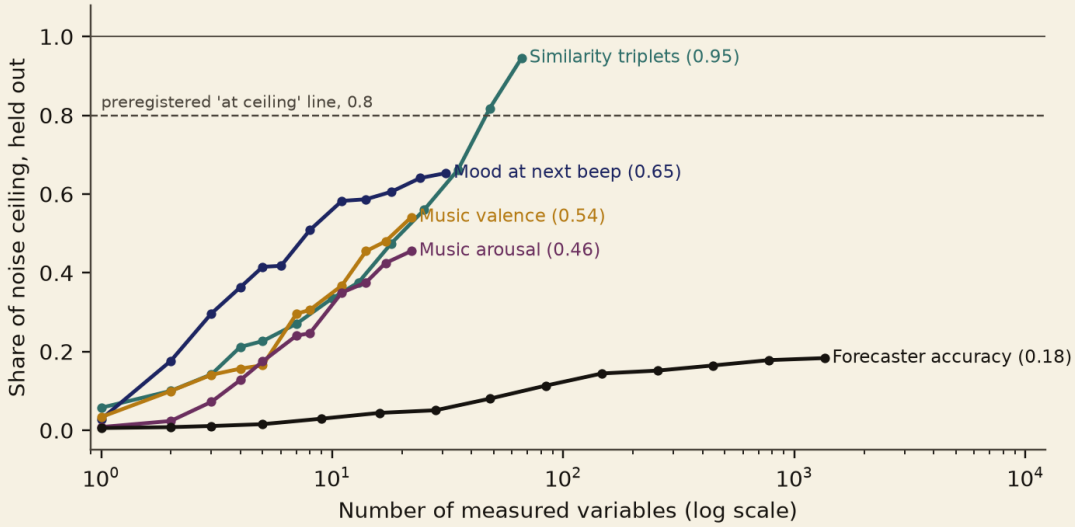


Figure 1. Share of the noise ceiling reached, mean of 8 random orders, against the number of variables. The similarity curve is scored on validation triplets (0.95 at all 66), while its headline share of 0.985 is scored on the repeated test triplets. Drawn from the result file.

reliable signal that the measured variables do not reach, not noise.

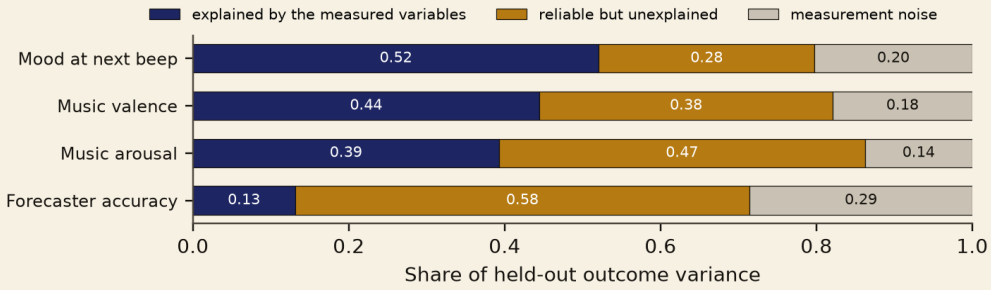


Figure 2. Held-out outcome variance split into the part the variables explain, reliable variance they do not reach and measurement noise, for the four numeric targets with a valid ceiling. Drawn from the result file.

The largest gains came from a different kind of variable: past accuracy lifted forecasters from 0.131 to 0.360 and prior growth lifted trees from 0.396 to 0.639. The other dimension's rating added 0.082 to valence and 0.092 to arousal; mood's training mean added 0.032 and tree death's prior growth 0.015 AUC. Only forecasters moved by the P2 rule. For trees the gain was 0.243, but the plateau interval for tree growth ran to the fit bound

of 2.0, so the rule could not pass; the preregistration's addendum records that defect and keeps the rule.

One of six targets reached 0.8 of its ceiling, against a forecast of at least four, and one of seven plateaus moved, against a forecast of at least three. Both are honest nulls of the claim's forecasts, reported with the same weight as the passes.

The toys

In the Lorenz toy the forecast horizon grew 2.51 time units per tenfold gain in precision, against 2.54 from theory, with linear R2 0.99994; twelve decades of precision bought 1.2 to 28.7 time units. The gain per decade over the last three decades was 1.02 times that over the first three. Figure 3 shows the line.

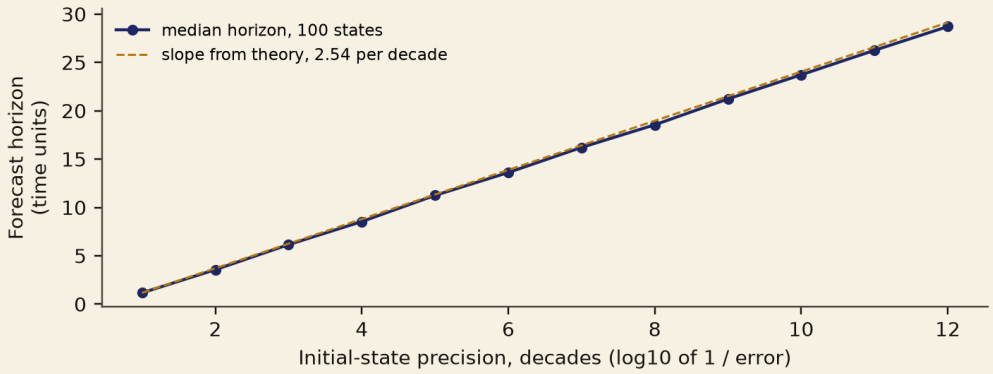


Figure 3. Lorenz-63 forecast horizon against starting precision, median over 100 attractor states, with the slope from the Lyapunov exponent. Drawn from the result file.

In the crowd toy with steep reaction (β = 20) and 1,000 members, an unpublished forecast with every member measured scored R2 1.00, the same forecast published naively scored -1.55, and a fixed-point forecast that solves for the reaction scored 0.9999. Measuring 100 of the 1,000 members cut the fixed-point forecast to 0.40 while the unpublished one scored 0.98. With identical members there is no fixed point, and the best published forecast scored -0.02.

In the overfitting check, plain least squares with 150 variables on 150 forecasters fitted the training data at R2 0.96 and scored -135,817 held out. Ridge on shuffled targets stayed at or below -0.005 at every count, and ridge on real targets reached 0.067 with all variables. P5 failed on its own wording, because it asked for a training R2 of at least 0.99 at 150 variables and got 0.96; every other rule in it held.

8 Discussion

The direction of the hope holds: skill grows with measured variables in every target and always beats the shuffled control. The strong form fails. In four of the five targets with a valid ceiling, the model stops between 0.18 and 0.65 of the way there, and the remainder is mostly reliable signal, not noise.

The exception teaches the most. The similarity variables were learned from the same triplet corpus, so they were built for the target, and they reach the ceiling. Our 0.660 sits close to the ceiling near 0.67 that Hebart and colleagues reported for an earlier version of the task [3], which we read at the level of abstracts. Variables chosen because they were available, not because they were fitted to the target, stall.

The gaps closed most when a different kind of variable arrived. For forecasters this agrees with Himmelstein, Atanasov and Budescu [5]: past accuracy beats the trait battery. For trees, prior growth beats size, site and stand. A target's own history carries signal its present description misses.

Salganik and colleagues [1] found that many methods on rich data barely beat a simple benchmark. Our ceilings

add a reading of that kind of result: a stall well below the noise ceiling says the missing information is in the world and not in the measurement error, so a different measurement can still help.

The toys bound the hope from the other side. With an exact model, chaos turns each tenfold gain in precision into a fixed slice of forecast time. A published forecast that people react to can be right only when the forecaster can solve for the reaction, and that solution degrades fast when most of the crowd is unmeasured.

The tier B* places this as a strong signal worth expanding. A rerun with non-linear learners and a preregistered tree ceiling on fresh splits would raise it.

9 Limitations

- Ridge and logistic models may leave signal that a non-linear model would find. A stall may belong to the model class.
- The datasets are a convenience sample of what was at hand, and only variables available in them were tried.
- The mood ceiling is a lower bound, the music rater count is assumed, and the tree ceiling estimator failed. A post hoc estimate in our ledger, outside the verdict and not in the result file, puts the tree ceiling at 0.74 to 0.82.
- The second run was not blind. Its one change, clipping, was fixed in a file hashed after the first run had been read.
- Several plateau intervals reached the fit bound, so they carry little information, and the P2 rule could not pass for those targets.
- The similarity task's shuffled control sits above chance by construction.
- The toys model their stated mechanisms only and say nothing direct about weather or markets.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 every skill, ceiling, share, strong-variable value, variance split, plateau fit, control bar, case count, toy value, overfitting value and verdict in this paper was re-read from the second run's result file, and the first run's forecaster R2 of -1.686 from the first run's result file. These match the findings ledger to rounding; the tree ceiling reads -3.9645 in the file and -3.97 in the ledger. Both preregistration files' SHA-256 values match their hash files and the hashes recorded in the two result files. The post hoc tree ceiling of 0.74 to 0.82 was read from the ledger only. All three figures were drawn by a script that reads only the second run's result file. The study was not re-run for this paper. The literature was found by searches run on 9 November 2026 and read at the level of abstracts, repository records and publisher pages; the full texts were not read.

11 Corrections

- 29 October 2026: the first run's forecaster R2 of -1.686 came from unclipped scaling of nearly constant variables. A disclosed addendum clipped scaled values at 5 in either direction and nothing else; the second run is primary and not blind.
- 29 October 2026: the addendum recorded three defects left in place on purpose: the failed tree ceiling estimator, the P2 rule that cannot pass for an unflattened curve, and the similarity control's above-chance start.
- 9 November 2026: first published version. The identification test and the target built on a private image collection are left out.

12 Reproduction

- Data: Forest Inventory and Analysis for Washington and Oregon [8]; openESM dataset 0062 [9]; DEAM static ratings [6]; Good Judgment Project data [7]; THINGS odd-one-out triplets and the 66 object dimensions [4].
- Hold out by plot, by each person's last 30 percent of beeps, by artist, by forecaster and by between-subject repeated triplets. Add variables in 8 random orders and correlation order at about 14 counts; refit and score held out at each count. Clip scaled values at 5 in either direction.
- Compute each ceiling from the data's own repeats and the share with eq. 2. Fit eq. 3 to the curves. Refit each target on shuffled training labels 20 times.
- For the toys, integrate Lorenz-63 from perturbed states until the error norm exceeds 1, and simulate the crowd with the attendance rule in the methods.
- Preregistration SHA-256: ee38c3bd5b2a255fe65bbe3c94c895bde618261da9a2c638e16fce87f2f7b602. Addendum SHA-256: e722f68052046c926fe562f66bfb368fef7b6c3120dc28de781b861e727e4d54. Plan SHA-256: 7ae9de2b411eaa112e2644a19075028358534a613be6439c97d6fd040b9b33d9. Runs 20261029T172001Z and 20261029T175624Z (primary).
- Cost: the second run took about 16 minutes on one desktop computer, most of it in the forecaster and similarity targets.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the first run. A verdict passes only when every rule holds. Values are read from the second run's result file. P6, the identification test on a private image collection, is omitted.

Table 2. Preregistered verdicts and the rules that decided each one.

Verdict	Deciding rules	Value	Holds
P1	at least 4 of 6 targets at 0.8 of ceiling or more	1 of 6	no
P2	at least 3 of 7 plateaus moved by the strong variable	1 of 7	no
P3	linear R2 at least 0.98; slope within 0.75 to 1.25 of theory; late to early decade gain within 0.75 to 1.33	0.99994; 0.988; 1.015	yes
P4	unpublished R2 at least 0.99; naive published R2 below 0.5; fixed point R2 at least 0.95; 100 of 1,000 measured costs at least 0.2; identical members below 0.5	1.00; -1.55; 0.9999; 0.58; -0.02	yes
P5	every target above its shuffled 95th percentile; least squares at 150 variables below 0 held out and at least 0.99 in training; shuffled ridge at most 0.02; real ridge above 0	all above; -135,817 and 0.96; -0.005; 0.067	no

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **noise ceiling**. The best held-out score any model could reach on a target, given how far repeated measurements of that same target disagree; it is estimated from the data's own repeats. Field: measurement theory and computational neuroscience. Excludes the best score reached so far by any model, and a bound set by missing variables.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field:

research method (our usage). Excludes an interpretation written after reading the numbers.

- **Brier score**. The mean squared difference between a probability forecast and the outcome coded 0 or 1; lower is better. Field: forecast verification (meteorology). Excludes measures that score only whether the most likely outcome happened.
- **held-out skill**. A model's score on cases kept apart from fitting: R2 for numbers, the area under the ROC curve for yes-or-no outcomes, and accuracy for triplet choices. Field: statistics and machine learning. Excludes the score on the data used to fit the model.
- **AUC**. The area under the ROC curve: the chance that a randomly chosen positive case is ranked above a randomly chosen negative one, with 0.5 for guessing. Field: statistics (signal detection). Excludes accuracy at any one threshold.
- **ridge regression**. Least squares with a penalty on the size of the coefficients, chosen by leave-one-out error, so that many variables can be fitted without chasing noise. Field: statistics. Excludes non-linear models such as boosted trees or neural networks.
- **parallel forms**. Two items meant to measure the same thing, answered at the same time; their correlation estimates reliability, here happy and cheerful at the same beep. Field: psychometrics. Excludes the same item repeated later, which also picks up real change.
- **split-half reliability**. The correlation between scores computed on two halves of the same items, here a forecaster's accuracy on the odd and the even questions. Field: psychometrics. Excludes agreement between different people or different occasions.
- **share of ceiling**. Held-out skill above chance divided by the noise ceiling above chance; 1 means the model scores as well as the measurements allow. Field: our usage. Excludes the raw skill, which is not comparable across targets.
- **saturating curve**. The curve s(k) = a k / (k + h) fitted to skill above chance against the number of variables k, where a is the level it approaches. Field: curve fitting. Excludes a forecast of skill with variables not in the dataset.
- **strong variable**. A variable of a different kind from the base set, named for each target before the run: past behaviour of the same unit, or a human rating of a related quantity. Field: our usage in this study. Excludes one more variable of the kind already in the base set.
- **forecast horizon**. The time until a forecast from a slightly wrong starting state differs from the truth by more than a set tolerance; in a chaotic system it grows with the logarithm of the starting precision. Field: dynamical systems (chaos). Excludes errors from a wrong model, which the toy does not have.
- **reflexive forecast**. A published forecast that the forecast population reads and reacts to, so that publishing it changes the outcome it predicts. Field: economics and social science. Excludes a forecast kept private from the people it describes.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **honest null**. A working test that found no effect beyond a stated bound, a forecast that failed, or an error of our own found and fixed, reported with the same weight as a positive result. Field: research method (our

usage). Excludes a void run, and a result left unpublished because it was negative.

- **tier**. How ready a finding is to publish: A candidate, B signal worth expanding (B* strong), C internal, D honest null worth reporting, E not reliable as it stands. Field: our publication register. Excludes a measure of how large or how true an effect is.

15 References

[1] Salganik, M. J., Lundberg, I., Kindel, A. T. and others (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences 117(15), 8398-8403. doi:10.1073/pnas.1915006117

[2] Lage-Castellanos, A., Valente, G., Formisano, E. and De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. PLOS Computational Biology 15, e1006397. doi:10.1371/journal.pcbi.1006397

[3] Hebart, M. N., Zheng, C. Y., Pereira, F. and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature Human Behaviour 4(11), 1173-1185. doi:10.1038/s41562-020-00951-3

[4] Hebart, M. N., Contier, O., Teichmann, L. and others (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. eLife 12, e82580. doi:10.7554/eLife.82580

[5] Himmelstein, M., Atanasov, P. and Budescu, D. V. (2021). Forecasting forecaster accuracy: contributions of past performance and individual differences. Judgment and Decision Making 16(2), 323-362. doi:10.1017/S1930297500008597

[6] Aljanaki, A., Yang, Y.-H. and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. PLoS ONE 12(3), e0173392. doi:10.1371/journal.pone.0173392

[7] Good Judgment Project. GJP data. Harvard Dataverse, CC0 1.0. doi:10.7910/DVN/BPCDH5

[8] USDA Forest Service. Forest Inventory and Analysis database, Washington and Oregon. FIA DataMart.

[9] openESM. Experience sampling dataset 0062 (Neubauer). CC BY 4.0.

[10] Measurement-bounded predictive inference in international large-scale assessment: how the plausible-value ceiling governs attainable accuracy and the stability of predictor rankings in PISA 2022. Research Square preprint rs-10732065 (authors and date not read; seen as a search summary).