A BETTER METRIC WILL NOT SAVE YOU: WHAT YOU COMPARE BEATS HOW YOU COMPARE IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Representational similarity measures (RSA, CKA, Procrustes) are chosen for the transformations they treat as irrelevant, and all of them take the set of conditions as given. That set is fixed by two decisions, which inputs are admitted and how they are grouped, both taken before the matrix a measure operates on exists: no invariance property reaches them. We measure what they are worth where the arrangement the conditions ought to have is known independently of the systems compared, the acoustic organisation of vowels, across fifteen languages and twenty-seven representations. Changing which categories are compared, at a fixed number of them, moves the gap between two representations by more than twice the effect being measured, while changing how many moves it not at all. In this setting rankings of models are identified under these choices, and rankings across languages are not.

1 INTRODUCTION

Claims of the form *model A represents X more like B than C does* are now routine: between successive layers of a network (Raghu et al., 2017; Kornblith et al., 2019), between networks differing only in random initialisation (Ding et al., 2021), and between artificial models and biological recordings (Kriegeskorte et al., 2008; Khaligh-Razavi & Kriegeskorte, 2014), among others. They are used to decide which architectures are worth pursuing and which accounts of a cognitive function remain viable.

Comparing representations is harder than it looks, because it is not obvious what *similar representation* could mean. Two systems given the same inputs produce representations of different widths and in different bases, with no correspondence between individual dimensions: a dimension of a transformer's hidden state has no counterpart in a second network, still less in a cortical recording. No shared coordinate system exists, so there is nothing to correlate, and were a number nonetheless produced, nothing would say whether it is a large one. Nor is either representation interpretable: with no account of what any dimension encodes, we cannot even specify what the two systems ought to agree about. Similarity must be defined without knowing what is being compared.

The measures that answer this, such as Representational Similarity Analysis (Kriegeskorte et al., 2008), Centred Kernel Alignment (Kornblith et al., 2019) or Procrustes-based shape metrics (Williams et al., 2021), were introduced in cognitive neuroscience, and machine learning adopted them largely as they stood. They compare the *organisation* of each representational space rather than the representations themselves: one fixes a set of N *conditions* and computes, within each system separately, the $N \times N$ matrix of pairwise distances between its representations; the two matrices are then compared. Fixing that set is two decisions, and we keep them apart throughout: *which inputs are admitted* to the comparison at all, and *how the admitted inputs are grouped* into the units among which distances are taken. A condition is thereby characterised by its relations to the other conditions and never by the dimensions encoding it, which is what dissolves the correspondence problem.

There is more than one such measure, however, and applied to the same data they disagree (Ding et al., 2021). A substantial literature has accordingly grown up around choosing among them, proceeding axiomatically and empirically (§2). Both routes are *intrinsic* by nature: nothing external says what a good similarity score would be, so a measure can only be held against other measures, against its own stability, or against perturbations whose importance is argued rather than observed. That is why the choice among these measures is still open after a decade.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Answering these questions calls for a change of method rather than a refinement of it: an *extrinsic* evaluation, one that holds a measure against an account of how the conditions ought to be arranged that owes nothing to any system being compared. Such accounts are rare, which is why the literature has done without one. Showing that a choice *changes* a conclusion requires no more than varying it; showing that a conclusion is *wrong* requires knowing what it should have been, and in most domains the structure the conditions ought to exhibit is exactly what is being estimated.

Speech supplies a domain where it is not. The acoustic arrangement of vowel categories is a genuine low-dimensional geometry, recoverable from the waveform by signal processing that shares nothing with the models under evaluation, documented for over seventy years (Peterson & Barney, 1952), and different in every language. **This paper uses it to measure how much of a comparative conclusion is fixed by the condition set rather than by the data**, in a setting where a choice that changes the answer can be shown to make it wrong about the claim as stated, not merely different. §3 describes the instrument and its scale, on which a similarity score of 0.87 can be called large or small rather than merely positive.

Every one of the evaluations above varies the measure and holds the set of conditions fixed, which is less an oversight than a consequence of the framing: the invariances of a measure are properties of what it does *once the $N \times N$ matrix has been formed*, so no invariance argument reaches a decision taken before that matrix exists. Both decisions that fix the condition set are taken before it. Each is mandatory, unconventioned, almost never reported, and, as far as we are aware, unmeasured. That analytic choices of this kind go undeclared, and change conclusions when they do, is documented at scale in neighbouring fields: almost every one of 241 surveyed fMRI studies used a unique analysis pipeline (Carp, 2012), and seventy teams given one dataset and the same nine hypotheses chose seventy different workflows and reported different results (Botvinik-Nezer et al., 2020). A filter can even attenuate the effect studied while creating a spurious one of opposite sign (Tanner et al., 2015).

Our own question makes both decisions visible. Asking whether a self-supervised speech model recovers that geometry better than a non-learned spectral representation does requires first deciding which stretches of audio count as instances of a vowel, and then which vowel categories those instances are grouped into. Neither is a matter of presentation: how many degrees of vowel height to recognise, or whether a diphthong is one condition or two, are contested theoretical questions on which specialists disagree (Appendix I). Each choice is therefore constrained but not determined, which is the dangerous case: one recognised as arbitrary would be reported, whereas one that is constrained but underdetermined is mistaken for a fact about the data.

It turns out to dominate. Changing which vowel categories are compared, at a fixed number of them, moves the gap between two representations by 0.146 on average: more than two and a half times the effect being measured, 0.0545. Changing which stretches of audio count as instances moves the gap in one language of fifteen from -0.047 to -0.002. Changing *how many* categories are compared, by contrast, has no measurable effect at all: from five to twenty-eight the gap moves by 0.005, of the order of its own standard error. What matters is neither how many conditions there are nor how widely they are spread through the reference space, but which ones they are. §4 gives the measurements.

Contributions. We treat the condition set as an object of measurement rather than as a given. Against a reference owing nothing to the systems compared, we quantify what each decision determines and which aspect of the set is responsible, show the effect to be structured rather than noisy, establish that a damaging selection cannot be recognised without recomputation, and report which claims remain identified once both decisions vary. Each answer is about correctness and not only stability. We also report what our pre-registered protocol caught (Appendix L).

2 Background: measuring representational similarity

The measures that answer the problem of §1 compare the internal structure of each system rather than the two sets of vectors: how the things being represented are arranged relative to one another, within one system and then within the other. They form one family among several (Klabunde et al., 2025), and the one on which the methodological debate has concentrated. This section fixes notation for the three we use, and locates within them the decision this paper measures. Approaches that instead fit a map between the two systems, or compare them by their behaviour, are discussed in Appendix F.

The rows of that arrangement are the *conditions* of §1, and what a row is varies by tradition. In much of the machine-learning literature a row is a single input (one image, one utterance) and the only decision is which inputs enter (Kornblith et al., 2019; Raghu et al., 2017). In the neuroscience tradition from which these measures come, a row is an experimental condition summarised over repeated trials (Kriegeskorte et al., 2008; Nili et al., 2014), and in analyses at the level of categories (object classes, word types, phonetic categories) a row pools every input carrying the same label. Where rows aggregate, both decisions apply: which inputs are admitted, and how the admitted inputs are grouped. Where a row is a single input, only the first does.

Nor is the grouping always the same on both sides. In model-to-brain comparisons the neural side is averaged over trials while the model side is deterministic, so the two sides are pooled by different rules as a matter of course, so that the estimator on one side is not the estimator on the other.

Write $\mathbf{X} \in \mathbb{R}^{N \times p}$ and $\mathbf{Y} \in \mathbb{R}^{N \times q}$ for the two systems at the level of conditions: N conditions, one per row, in p and q dimensions respectively. The comparison is now between two objects of size $N \times N$, one per system, and N is the size of the condition set rather than the number of inputs.

RSA (Kriegeskorte et al., 2008) builds, for each system, the $N \times N$ representational dissimilarity matrix (RDM) of distances between its conditions, and correlates the two, typically with Spearman's ρ_s, so that no linear relation between the two distance scales need be assumed:

$$(\mathrm{RDM}_{\mathbf{X}})_{ij} = d(\mathbf{x}_i, \mathbf{x}_j), \qquad \mathrm{RSA}(\mathbf{X}, \mathbf{Y}) = \rho_s\big(\mathrm{vec}^{\triangle}\mathrm{RDM}_{\mathbf{X}},\ \mathrm{vec}^{\triangle}\mathrm{RDM}_{\mathbf{Y}}\big)\,, \tag{1}$$

where $\mathbf{x}_i$ is row i of $\mathbf{X}$ and $\mathrm{vec}^{\triangle}$ vectorises the strict upper triangle of a symmetric matrix, $m = N(N-1)/2$ entries, one per pair of conditions; m bounds what the correlation can resolve.

Linear CKA (Kornblith et al., 2019) builds a different $N \times N$ object from the same rows, the Gram matrix of inner products between conditions, and normalises the alignment of the two through HSIC, which measures that alignment once the mean has been removed from each. Orthogonal Procrustes (Williams et al., 2021) takes a third form, minimising $\|\mathbf{XQ} - \mathbf{Y}\|_F$ over rotations $\mathbf{Q}$; since $p \neq q$ in general, the two must first be brought to a common dimension, at a projection rank the definition does not supply (Appendix A). The three also differ in what they leave to the analyst: RSA fixes neither the distance d nor the statistic comparing the triangles, which CKA and Procrustes fix by definition.

Every object in these definitions is indexed by conditions and not by dimensions, which is what makes each invariant to permutation of dimensions by construction, and to orthogonal rotation whenever the underlying proximity is Euclidean. That invariance is bought at a price, and the price is the subject of this paper: the coordinate space has been discarded and only the arrangement among conditions retained, so everything downstream depends on *how that arrangement was built*, and the condition set is its first ingredient. Which inputs enter, and how they are grouped, is not a detail of the procedure: it is the object the procedure operates on.

The methodological debate around these measures is, with no exception we are aware of, a debate about invariance, and for the reason §1 gives: with no external criterion for what a good similarity score would be, a measure can only be held against controls internal to the comparison. RSA was introduced to compare measurement modalities with no common units at all (Kriegeskorte et al., 2008) and supplied with a reliability analysis shortly after (Nili et al., 2014). CCA and its variants sought invariance to any invertible linear map (Raghu et al., 2017; Morcos et al., 2018); Kornblith et al. (2019) showed this to be too permissive, since discarding the distinction between high- and low-variance directions makes trained and untrained networks resemble one another, and proposed linear CKA with the invariance weakened to the orthogonal and isotropic cases. Williams et al. (2021) responded to both, on the grounds that neither RSA nor CKA is a metric and so neither licenses clustering models or averaging over them. Two later contributions replaced axioms by experiment while keeping the controls internal: Ding et al. (2021) perturbed representations in ways known on independent grounds to change what a network computes and found that the measures differ in whether they register the change at all, so that the answer to a comparative question can depend on which measure was used, and Davari et al. (2023) showed that CKA can be driven to arbitrary values without materially changing what a network computes.

Every one of those proposals concerns what a measure does to the $N \times N$ matrix. What none of them leaves open is the condition set, which each receives already formed.

3 AN INSTRUMENT WITH A KNOWN ANSWER

To measure a choice no invariance argument can reach, we need a case in which the arrangement the conditions ought to have is known before any model is consulted. Such cases are uncommon: in most domains where these measures are applied, the structure the conditions should exhibit is what is being estimated. Vowels offer one. Vowel categories occupy distinct and stable regions of a two-dimensional acoustic plane, the vowel triangle, documented since the 1950s (Peterson & Barney, 1952) and used in phonetics as a measurement tool rather than a theoretical idealisation. Its coordinates are the first two *formants*, F_1 and F_2, the two lowest resonances of the vocal tract, which track tongue height and backness and are recovered from the waveform by standard signal processing.

What this yields is a *geometry* and not a set of labels. Positions in the (F_1, F_2) plane admit distances, so for any set of vowel categories a pairwise structure can be written down against which a model's own may be compared, which class labels or a proxy task could not supply. It comes from the same recordings the models receive, by a route sharing no component with any of them: the alignment locating the tokens is produced by a trained model (McAuliffe et al., 2017), but one not among the systems evaluated, so its errors are common to all rather than favouring any.

We call this an *external reference* and not a ground truth because it is itself a measurement, with its own assumptions and errors; what makes it usable is not that it is beyond doubt but that it is fixed before any model is consulted. It does not state what a speech representation ought to encode, only one measurable property of the signal against which a comparative claim can be evaluated.

Because that arrangement is known, a choice that changes the answer can be shown to make it wrong about the claim as stated, and not merely different; and because inventories differ, the question of §1 can be asked once per language.

Data and protocol. We use fifteen of the 102 languages of FLEURS (Conneau et al., 2023), a read-speech corpus with content parallel across languages, chosen to span inventories of five to twenty-eight categories and to vary in tone and in family (Appendix G); the corpus carries no speaker labels, which constrains what can be controlled for.

Forced alignment (McAuliffe et al., 2017) gives a phone label and a time interval for every vowel token, formants are estimated with Praat (Boersma & Weenink, 2021), and two levels of aggregation reduce each system to one vector per category: frames are mean-pooled within a token, then tokens sharing a label are summarised by their mean on the representation side and their coordinate-wise median on the reference side. The two sides therefore use different estimators, each with a local justification and the pair with none, which §4.4 measures. Comparing a representation with the reference is then RSA as defined in equation 1, its two open choices fixed to cosine distance on the representation side, Euclidean on the reference side, and Spearman's ρ between the vectorised triangles: the default throughout, from which each subsection of §4 varies one step (Appendix B).

Representations compared. Twenty-seven representations are evaluated, chosen so that the scale defined below has known points along its whole length rather than only at the top. Three families are the objects of the comparison: six self-supervised models, one of them pre-trained on music, two encoders trained or fine-tuned on transcriptions, and four neural audio codecs. Two fix the scale from below: four non-learned representations, and nine randomly initialised controls matching the trained architectures layer for layer, which isolate what the architecture contributes before any training. The last is a floor: one model trained on images, run at two frame rates, which shares the architecture family but not the modality. Appendix H gives each checkpoint with its revision hash, the architecture and initialisation of every control, and how frames are obtained from the vision model.

An anchored scale. A similarity score has no natural units, which is tolerable when the only question is which of two models scores higher. Here it is not: we must say whether the effect of a pipeline choice is large *relative to the claimed effect*, and without a common yardstick the sentence *the choice of categories moves the result more than pre-training does* cannot even be evaluated.

The reference provides that yardstick. Table 1 reports, per family of representations, the median over the fifteen languages of the RSA score in the default configuration just described. Three rows are controls rather than models, and they are what makes the others readable: the reference passed through the representation side of the pipeline and compared with itself, which scores 0.99997 and

Table 1: The anchored scale: median ρ over the fifteen languages, one row per representation family.

ρ	anchor	content
0.99997	reference ceiling	identity control: the geometry against itself
0.9003	pre-trained speech and audio models	8 models: 6 self-supervised, 2 supervised
0.8535	non-learned spectral features	3 baselines
0.8421	learned compression	4 neural codecs
0.7571	raw formant measurements	F_1–F_4 as a feature vector
0.5024	architecture alone	9 randomly initialised controls
0.3298	out of modality	1 vision checkpoint at 2 frame rates

which no representation can better; the trained architectures with their weights left at initialisation; and the model trained on images, which shares the architecture family but has never heard speech. Together they bound the axis from above, from the middle and from below.

We express every effect as a percentage of the interval between the first control and the second: the distance between an untrained architecture and exact recovery of the geometry. Both ends are controls and not models, deliberately. A unit defined as the distance between two named models would be the very quantity the scale is used to judge, and every effect would be reported as a fraction of itself. That interval is the unit of §4 throughout, the instrument that makes one pipeline choice comparable with another and with the effect being claimed.

Anchoring that scale on RSA is itself a choice, made for resolution: each measure has an interval of its own, RSA's is the widest of the three at 0.49755 against 0.3331 and 0.1742 (Table 2), and the widest interval is the one on which two effects of similar size are least likely to be confused. The choice is not neutral, and §4.1 measures what it is worth: read each on its own interval, the three return the pre-training gap with the same sign and order of magnitude.

How much of the axis is noise. The reference is estimated, so its reproducibility bounds what any score against it can mean. A split-half reliability of the reference matrix, by utterance with the Spearman–Brown correction (Nili et al., 2014), runs from 0.968 to 1.000 across languages with a median of 0.997, and no model–language cell exceeds the ceiling of its language: what §4 reports are differences and not noise. One qualification limits what we recommend. Reliability is not validity: a systematic error, an octave jump made consistently by the formant tracker, reproduces in both halves and leaves the ceiling untouched, so a high ceiling certifies that the reference is stable, not that it is right, and the control of Appendix N inherits that blind spot.

Bifurcations. Every step just described involves a choice the procedure requires and nothing determines: which tokens are admitted, which labels kept apart, which estimator summarises a category, which distances are taken. We call such a choice a *bifurcation* when it is mandatory, unconventioned, rarely reported and, decisively, *asymmetric*: when it does not affect the compared systems alike.

That last property carries the argument. A choice shifting every score equally is a nuisance parameter, changing values and leaving conclusions alone; one that affects the compared systems differently changes rankings. Only the second threatens identification, so we report for every bifurcation a *common effect*, the mean shift across systems, and a *differential effect*, the variance of that shift between them. Only the second bears on what may be concluded.

4 RESULTS

We begin with the choice the literature has debated, then turn to the ones it has not. The unit of replication differs by bifurcation: choices applying to every language are bootstrapped over the fifteen, whereas composition is evaluated over repeated draws within one language, so the two intervals are not comparable estimates of the same uncertainty and we do not read them as such. Figure 1 is the shape of what follows: those choices displace the gap between pre-trained models and non-learned spectral features, the largest by more than the gap itself.

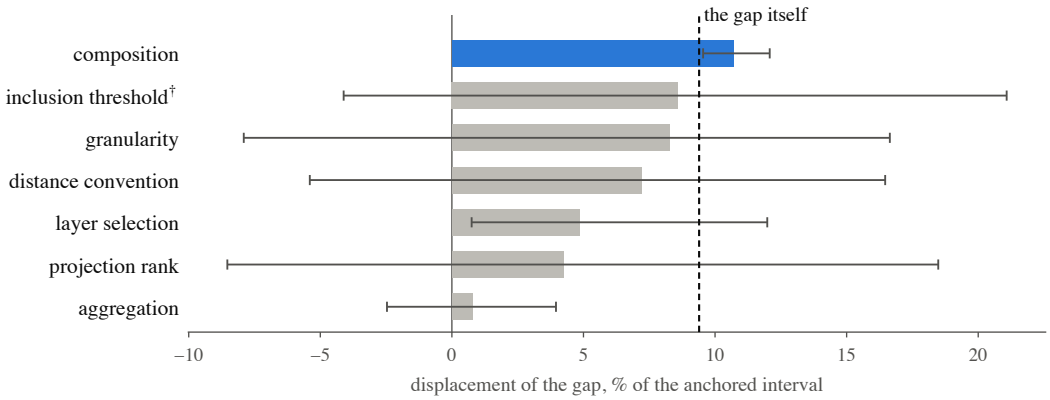

Figure 1: Each undeclared choice, and the effect it is meant to help measure. Bars give the displacement of the pre-training gap in percent of the anchored interval (§3); the vertical line is the gap itself. The dagger marks the sweep run over eight representations, not twenty-seven.

Table 2: The three measures on one question. Top: rank correlation between two measures of the ranking of representations. Bottom: the gap between pre-trained speech and audio models and non-learned spectral features under each measure, on that measure's own anchored interval.

pair of measures	27 representations	16 trained	16 trained, English only
RSA / CKA	+0.969 [+0.913, +0.985]	+0.868 [+0.617, +0.959]	+0.856
RSA / Procrustes	+0.967 [+0.894, +0.988]	+0.909 [+0.638, +0.976]	+0.635
CKA / Procrustes	+0.944 [+0.893, +0.974]	+0.779 [+0.609, +0.918]	+0.785
measure	*floor – ceiling*	*gap*	*% of its own interval*
RSA	0.5024–0.99997	0.0467	9.4% [4.8, 21.4]
linear CKA	0.6669–1.0000	+0.0289	8.7% [5.6, 18.2]
Procrustes, rank 2	0.8258–1.0000	+0.0117	6.7% [4.1, 16.8]

4.1 What the measures decide, and what they leave open

The literature's question is which of these measures to use, so we ask first what that choice is worth. RSA, linear CKA and Procrustes at matched rank are applied to the same representations, conditions and reference, and compared on the orderings they produce.

Less than the literature would lead one to expect. That these measures can answer the same question differently is well established (Kornblith et al., 2019; Ding et al., 2021; Davari et al., 2023); here it does not reach the ordering. The three rank the twenty-seven representations alike, and the gap between pre-trained models and non-learned spectral features keeps its sign and its order of magnitude under all three, each read on its own anchored interval (Table 2). They part company where comparisons are actually made: among the sixteen trained and spectral representations, without the controls, agreement falls markedly. Much of it comes from the distance between trained models and controls rather than from the ordering among trained models, and CKA moves those most, promoting a codec and a music-pretrained model and demoting a supervised encoder. None of this makes the three interchangeable: they agree on the coarse ordering here and disagree substantially once the controls are removed.

That reassurance is conditional, and on the kind of choice this paper is about. A measure is not a procedure: RSA fixes neither the distance nor the statistic comparing the triangles, and Procrustes at matched rank requires a projection rank (§2). Varying those, four of five rank-by-domain configurations change verdict between the two distance conventions, in both directions, and the two factors interact (Table 8, Appendix E). The reversal is not an artefact of our own pipeline: under the same change, a quantity that the convention cannot reach does not move at all (Appendix E).

The sensitivity this paper is about is therefore already visible in the question the literature has been asking. The measures agree, until one of the choices they leave open is made differently.

One limit no measure escapes. A rank correlation over k conditions moves on a grid of spacing $12/\big(m(m^2-1)\big)$ with $m = k(k-1)/2$, so below some k the difference reported is smaller than the step between attainable values: at five categories every observed value falls on a grid point and four models share one. The criterion is operational, not a rejection of small inventories, and §4.2 gives the case that fails it.

4.2 How the inputs are grouped

An analyst comparing two systems over k vowel categories must choose which k; nothing in the procedure fixes the choice and it is rarely reported. Drawing five hundred subsets at each of six cardinalities separates the effect of the size of the set from that of its composition (Table 11, Appendix E). Size has no measurable effect: from five categories to twenty-eight the gap moves by less than its own standard error. Composition has a large one: two analysts drawing five categories each report gaps separated by 0.146 on average, against an effect of 0.0545 that both are trying to measure. We report the two quantities separately rather than their ratio, whose denominator is not distinguishable from zero and whose value depends on two undeclared choices (Appendix E).

Random subsets are not what an analyst would pick, so we restrict the draws to the dispersion of the fifteen attested inventories: that removes a fifth of it at five categories and leaves the rest twice the effect, so the result is not an artefact of implausible sets. Nor would a summary serve in its place: four coverage statistics account for 0.101 of the pooled variance and cardinality alone for none of it (Appendix C), so the set has to be reported category by category.

What matters for published work is where an inventory somebody built falls among those they could have built instead. The twelve-category English merge table used here was built for this project (Appendix C); it falls at percentile 71.8 among five hundred random subsets of its size and at 75.2 among merges matched for the profile of category sizes, while four languages whose inventory changed without a hand-built table fall between 38 and 47, the lower bound of the English interval lying above all four. A hand-built inventory is therefore not one arrangement among many: it lies where the distribution is thin, which is what makes it unrepresentative of the choices a reader might have made.

Which categories are compared is half of the grouping decision; how finely they are cut is the other, and merging is the commonest manipulation of an inventory, always on the ground that finer distinctions are unreliable. Merging that raised every score equally would be a nuisance parameter. Merging the twenty-eight English categories into twelve raises every score, by +0.0768 on average, and does not raise them equally. The shift is monotone in the quality of the representation ($\rho = -0.773$, $p = 3.7e{-}06$): the six weakest representations gain +0.1225, the six strongest +0.0284, and the spectral baselines lose ground. A coarse inventory flatters weak models. The consequence is that the gap between a strong self-supervised model and its randomly initialised counterparts falls from 0.317 to 0.224: the sign of the comparison is preserved, and the margin supporting it depends on a convention that is rarely declared.

4.3 Which inputs are admitted

Tokens shorter than some duration are routinely discarded, on the ground that a formant estimate over too few periods is unreliable. The threshold is never the same twice and never reported. We sweep it from 0 to 100 milliseconds, extracting once at zero and filtering afterwards so that the sets are nested.

The rule changes what is concluded. The ranking of languages survives only partly, and in Polish, the only language in which the spectral baseline is ahead, the threshold moves the gap from -0.047 to -0.002. About half of what the threshold does is to move the compared systems relative to one another rather than together.

A filter's bias is measurable without recomputing anything: one compares the distribution of categories among the tokens kept and those discarded. We pre-registered the prediction that this quantity would predict the threshold's differential effect, and that both would be largest in the language included in the corpus for its extreme vowel reduction.

The prediction fails twice. Russian is indeed the most heavily cut language and first on both measures of bias, yet its differential effect ranks fourth of fifteen, and across the fifteen languages the correlation between measured bias and differential effect is +0.118 ($p = 0.676$). The Russian case keeps its force as a matter of construct validity, since the filter removes precisely the phenomenon the language was included for. But the one language in which the damage was visible in advance is not the one in which it was worst, and the cheap diagnostic does not substitute for the sweep.

4.4 What the analyst does downstream

Once the condition set is fixed, two choices remain: how a condition's tokens are summarised and which layer of a deep model is reported. Neither belongs to the set and neither is usually declared. The two estimators of §3, a mean on the representation side and a coordinate-wise median on the reference side, are a bifurcation of the kind this paper is about, found in our own work; we recompute the representation side with the reference's estimator (Table 12, Appendix E).

Matching the two estimators displaces the comparison by 2.33% of the anchored interval and lowers every anchor. We had predicted that the gap would narrow, since the spectral baselines carry the heavier-tailed within-category distributions and should gain most from a robust centroid. The premise is correct and the gap nonetheless widens: the reference is itself a coordinate-wise median over formant measurements, and the tokens it discards are not those lying far from a centroid in a spectral space, so matching the name of an estimator does not match its effect. Neither configuration is principled and we report both. As with the inclusion threshold, a quantity available without recomputation orders the systems correctly and fails to predict the consequence.

Where a model has many layers and a baseline one, reporting each at its best layer compares a maximum over many draws with a maximum over one, an asymmetric multiple-comparison advantage that grows with the number of layers and so cancels in no comparison. Matched relative depth is one control among several; we use it to size that advantage, not to claim it removed.

Of the 32 paired comparisons, 0 survive the change of protocol. The inflation is 5.2% of the anchored interval against a median gap of 0.0564, so selection accounts for 46% of the reported margin. This is the one bifurcation we can test outside speech, and there the reference is exact rather than measured: on dSprites, whose images are generated from known factors, a baseline contests five of twelve comparisons between three trained vision models and four unlearned ones at matched relative depth, where raw pixels are ahead of all three (Appendix J).

4.5 What survives

Each effect above was reported in the quantity natural to its analysis, and those quantities are not commensurable, so we ask one question of each: how far does it move the gap between pre-trained models and non-learned spectral features. Figure 1 gave the shape, Table 13 the values (Appendix E).

One choice separates and five are comparable. Only composition has an interval excluding the others; the inclusion threshold, granularity, the distance convention, layer selection and projection rank follow, ordered by estimate, with overlapping intervals all containing zero but the layer. Aggregation is negligible. They support reading *composition first, alone in separating; then a group of five of comparable magnitude, the choice of measure among them; then aggregation*, not a strict ordering of seven lines.

Two qualifications belong with that comparison, beyond the unit of replication noted above, which also leaves composition without a multilingual interval since only one language has twenty-eight categories. Two lines change rank once brought to this unit, aggregation falling and granularity rising, because a spread between models and a displacement of the gap are two quantities (Appendix D). The dominance is not an artefact of the protocol: standardising the reference once over the full inventory rather than on each subset gives +12.7% instead of +10.7%, with the curve unchanged in shape.

That comparison measures displacement, whereas a published comparison usually reports an ordering, which may be more robust. We compute the rank correlation of the twenty-seven representations between the extremes of each choice and between pairs of draws.

Orderings survive more than values do, but not uniformly. The ranking of models correlates above +0.922 across both undeclared choices and at +0.926 between two draws of fourteen categories.

Table 3: What survives each choice: rank correlation of a ranking between the two extremes of the choice, with a bootstrap interval and the number of items it is computed on.

choice	models	trained models	languages, by level	languages, by gap
inclusion threshold, 0 vs 100 ms	+0.922 [+0.59, +1.00], $n = 8$	—	+0.786 [+0.41, +0.95], $n = 15$	+0.507 [-0.07, +0.88], $n = 15$
distance convention	+0.929 [+0.80, +0.97], $n = 27$	—	+0.432 [-0.14, +0.81], $n = 15$	+0.518 [+0.04, +0.86], $n = 15$
English only, like for like				
composition, $k = 20$	+0.969 [+0.965, +0.972], $n = 27$	+0.921 [+0.915, +0.924], $n = 16$	—	—
composition, $k = 14$	+0.926 [+0.921, +0.930], $n = 27$	+0.838 [+0.832, +0.847], $n = 16$	—	—
measure, RSA vs Procrustes	+0.872, $n = 27$	+0.635, $n = 16$	—	—
composition, $k = 5$	+0.539 [+0.512, +0.565], $n = 27$	+0.313 [+0.282, +0.344], $n = 16$	—	—

Restricted to the sixteen trained and spectral representations the last falls to +0.838, which is still higher than two measures agree with each other on the same language. At five categories the ordering collapses: +0.313 between draws over the trained representations, with a quarter of draw pairs below +0.06. That regime is not hypothetical: Spanish, Mandarin and Polish have inventories of that size.

The thesis is therefore best stated by regime. On inventories of typical size, the composition of the condition set moves the values more than the choice of measure, and disturbs the ordering less. Only on small inventories does it disturb the ordering more than the choice of measure.

In this setting comparisons between models are identified and comparisons across languages are not, since two unrelated undeclared choices each suffice to scramble that ranking.

5 Conclusion

A decade of work on representational similarity has gone into choosing among measures; that is not where the leverage is. The three we compared rank twenty-seven representations alike and agree on the pre-training gap more closely than the choices we measured differ, while the composition of the condition set moves that gap by more than the gap itself. What is compared beats how it is compared.

The argument is not against the invariance literature, which answered its own question, but about ordering. A measure's invariances are properties of what it does once the $N \times N$ matrix exists, so no result there reaches a decision taken before that matrix is formed, as both decisions fixing the condition set are. A measure can be invariant to everything it was designed to be invariant to yet return whatever the analyst's inventory determines.

Three things follow for practice. A comparative claim should report its condition set category by category, since no summary of the set stands in for it; report the rule that admitted its inputs, and sweep it, since no cheap diagnostic predicts what that rule will do to a comparison; and compare systems at matched depth, or say that it did not. Appendix N states these alongside our own record, offered as a base rate and not as evidence of rigour: pre-registration, a bug journal and a test suite still left the defects listed in Appendix L.

What this does not establish. The two decisions differ in reach: grouping bears on work in which a row aggregates several inputs, admission wherever these measures are used at all. Two of our results rest on English, the only language of our sample in which the space of condition sets can be explored widely and whose hand-built inventory is itself unusual among those we sample (§4.2); only the layer-selection artefact was tested in a second modality, and at least two further bifurcations meet our criterion unexamined. The magnitudes are therefore established for vowel geometry in multilingual speech; the structural claim above is not so confined (Appendix K).

What made any of this measurable was a criterion external to the comparison. Intrinsic validation can establish that two measures disagree, not which is wrong, nor see a choice taken before any measure is applied. Domains where the arrangement of the conditions is known independently of the systems compared are rare, but not confined to speech: the layer-selection artefact reproduces, and is larger, where generative factors supply the reference exactly, so that the objection to our own reference cannot arise (Appendix J). We would rather the field looked for more such domains than a better measure.

Reproducibility Statement

Every measured value in this paper, in the text, tables and captions alike, is generated programmatically from the stored result files rather than transcribed by hand, so the prose cannot drift from the results it reports; the few exceptions are fixed protocol constants and a handful of figures from exploratory runs whose scripts write no result file. Each value is bound to an artefact and its checksum, and the generation chain refuses to build when two names are bound to the same value of the same artefact, when a table declares a quantity another table declares from a different source, or when a percentage cannot be recomputed from its operands. The pipeline was frozen after the last defect recorded in Appendix L and every value regenerated from it, so no number reported here predates a correction; the largest of those corrections moved the anchored interval in its fifth decimal and changed no conclusion. The extraction, analysis and figure code, together with the scripts that regenerate every reported value, are included in the anonymous supplementary material.

Appendix M states which of our pre-registered predictions can be checked from outside the project and which cannot, and Appendix L gives the defect record in full. The protocol is in Appendix B, the language sample in Appendix G, and the checkpoints with their revision hashes in Appendix H. One sweep is narrower than the others for a reason that is recoverable: the stored representations predate the change that records segment durations, so extending the duration threshold from eight representations to twenty-seven requires re-extracting nineteen models over fifteen languages, of the order of three GPU-days. That is a decision about time, not a technical obstacle, and we say so.

Ethics Statement

This work analyses public corpora and public model checkpoints and collects no new data from human participants. FLEURS is used as published, with no annotation added; it carries no speaker identifiers, which bounds what could be inferred about any individual from our artefacts.

One of our results bears on a class of claim with equity implications. Comparisons of how well a representation serves one language against another are precisely the comparisons §4.5 finds not to be identified under choices that go undeclared. Statements that a multilingual model represents some languages as well as others, or serves a low-resource language adequately, are therefore more fragile than their published form suggests, and we would not want the instrument of §3 read as licensing them. The remedy is the one this paper recommends throughout: report the condition set, and the rule that admitted the inputs.

AI Use Statement

The authors used a large language model as a writing aid, for drafting and revising prose and for checking internal consistency across sections, and as a coding assistant in the preparation of the analysis code. It was not used to generate the research ideas or the hypotheses tested here, and no number in the paper is produced by a language model: every value is computed by the pipeline described in Appendix B from stored artefacts, and bound to them as the Reproducibility Statement describes. The authors take full responsibility for the content, including every claim and citation.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Paul Boersma and David Weenink. Praat: Doing phonetics by computer, 2021. Computer program, version 6.1.38.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 2020. doi: 10.1038/s41586-020-2314-9.

Joshua Carp. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289–300, 2012. doi: 10.1016/j.neuroimage.2012.07.004.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023. doi: 10.1109/SLT54892.2023.10023141.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13890–13902. Curran Associates, Inc., 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, 10(11):e1003915, 2014. doi: 10.1371/journal.pcbi.1003915.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Peter Ladefoged and Ian Maddieson. *The Sounds of the World's Languages*. Blackwell, Oxford, 1996.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5727–5736, 2018.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10 (4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Jennifer Nycz and Lauren Hall-Lew. Best practices in measuring vowel merger. In *Proceedings of Meetings on Acoustics*, pp. 060008, 2014. doi: 10.1121/1.4894063.

Gordon E. Peterson and Harold L. Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952. doi: 10.1121/1.1906875.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

Eli Stafford, Aimée Lahaussois, and Guillaume Wisniewski. What do neural speech models know about phonology? Evidence from structured phoneme confusions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25860–25873, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1288.

Darren Tanner, Kara Morgan-Short, and Steven J. Luck. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, 52(8):997–1009, 2015. doi: 10.1111/psyp.12437.

J. C. Wells. *Accents of English*. Cambridge University Press, Cambridge, 1982. Three volumes.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

A THE TWO MEASURES STATED IN FULL

Equation 1 of §2 defines RSA, which is the default throughout. The other two are stated here. Linear CKA (Kornblith et al., 2019) compares the Gram matrices of inner products between conditions,

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top, \quad \mathbf{L} = \mathbf{Y}\mathbf{Y}^\top, \qquad \mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathrm{HSIC}(\mathbf{K}, \mathbf{K})\,\mathrm{HSIC}(\mathbf{L}, \mathbf{L})}}, \tag{2}$$

with $\mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \mathrm{tr}(\mathbf{KHLH})/(N-1)^2$ and $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ the centring matrix, so that HSIC measures the alignment of the two structures once the mean has been removed from each. Orthogonal Procrustes (Williams et al., 2021) minimises $\|\mathbf{XQ} - \mathbf{Y}\|_F$ over rotations $\mathbf{Q}$, which requires the two systems to be brought to a common dimension first; we project both to a matched rank d, and d cannot exceed $k - 1$ for k centred conditions, which is why §4.1 treats the rank as a bifurcation in its own right rather than as a preprocessing detail.

B EXPERIMENTAL PROTOCOL

Aggregation. On the representation side, a token's interval selects a span of frame-level representations, mean-pooled into one vector per token; tokens carrying the same phone label are then summarised by their mean. On the reference side, categories are summarised by the coordinate-wise median of their tokens' (F_1, F_2) pairs. Formant estimation carries occasional gross errors, octave jumps and F_1–F_2 confusions, to which a mean is not robust, whereas the representations show no comparably structured outliers; measuring vowel differences by Euclidean distance between category centroids is in any case an established approach in phonetics (Nycz & Hall-Lew, 2014). Taking the median per coordinate commutes with any monotone per-coordinate transformation, so the order of the median, the Bark warping and the standardisation is immaterial: a candidate degree of freedom that turns out to be inert, which we note because so few of the others are. It is not equivariant to rotation of the (F_1, F_2) plane, which is acceptable because those axes are interpretable rather than arbitrary.

The space of analyses explored. Not every grouping of vowel categories would be defensible, and we do not claim otherwise: pooling two categories at opposite corners of the plane would produce a centroid corresponding to no attested vowel. Our principal manipulation is selection rather than arbitrary merging. We draw subsets of the categories a language actually distinguishes, so that every condition in every draw is a genuine category of that language; where merging is at issue (§4.2) the draws are matched to the profile of category sizes of the published inventory, and restricted to the dispersion spanned by attested inventories. The space we explore is the space of defensible analyses, not the space of all partitions.

Table 4: Variance of the gap explained by four coverage statistics of the condition set, and by its cardinality alone.

scope	R^2	raw σ	residual σ	n
k=5	0.096	0.144	0.138	500
k=7	0.125	0.082	0.077	500
k=12	0.260	0.047	0.041	500
k=15	0.238	0.037	0.033	500
k=20	0.338	0.023	0.019	500
pooled: coverage only	0.101	0.080	0.076	2500
pooled: cardinality only	0.000	0.080	0.080	2500
pooled: coverage and cardinality	0.115	0.080	0.075	2500

Why vowels admit a reference at all. The vowel triangle is a measurement tool rather than a theoretical idealisation: it is reproducible in any laboratory from a microphone and an hour of recordings, and recovered directly from the signal, with no model and no annotation anywhere in the procedure. This is what makes it independent of the systems compared, rather than one more estimate from a related pipeline.

Two remarks on the noise ceiling. The Spanish ceiling is exactly 1.000000: with five categories there are only ten distances to order and both halves ordered them identically, which is a property of a coarse grid and not a measurement, and we use it as a denominator nowhere. The limitation that matters more, that a split-half reliability measures reliability and not validity, is stated in §3 because the recommendation of Appendix N depends on it.

Reading the tables. Unless a table says otherwise, its cells are computed in the default configuration of §3, with the layer fixed at the RSA argmax; a score for a representation is the median of its per-language scores over the fifteen languages, and an interval is a percentile bootstrap over languages with two thousand resamples. On the anchored scale, a measure's floor is the median of the nine random-weight controls under that measure and its ceiling is the reference compared against itself, so each measure has an interval of its own. Where a projection rank is varied, cells that cannot exist are left empty rather than set to zero: a rank of d requires $k \geq d + 1$ conditions, so rank and cardinality cannot be crossed freely and the space of pipelines is not a hypercube. We are not aware of this being noted elsewhere. Where a row corresponds to a single draw, such as the full inventory, it has no dispersion to report. Where a table correlates two rankings of the languages, *by level* ranks them by the median score over representations and *by gap* by the gap between a self-supervised model and a spectral baseline, each taken at both extremes of the choice. A dash marks a cell that was not computed, which is not the same as one that cannot exist. Departures from any of this are stated in the subsection that uses the table.

Units. Frequency in hertz is not perceptually uniform: the same interval spans a large perceptual distance low in the spectrum and a small one high up, so Euclidean distances computed in hertz would exaggerate differences along F_1 relative to F_2 and misrepresent the geometry the reference is meant to encode. Warping to a perceptual scale before measuring distances is standard practice in phonetics, and Bark is one of the scales in common use. Each coordinate is then standardised within language.

C THE CONDITION SET: COVERAGE, PLAUSIBLE INVENTORIES, PERCENTILES

The three tables below support §4.2. The first regresses the gap on four coverage statistics of the condition set, pooled and by cardinality. The second restricts the draws to the range of dispersion the fifteen attested inventories span. The third places the hand-built English inventory among the subsets and the size-matched merges it could have been. Its contrast row is that English percentile minus the mean percentile of the four languages whose inventory changed without one, each among random subsets of its former size, with a bootstrap interval over both; the range printed for those four is their spread, not the mean the contrast subtracts.

Table 5: Dispersion restricted to the range of coverage spanned by the attested inventories.

k	in band (%)	σ all	σ in band	reduction (%)	mean gap
5	67.6	0.144	0.113	22	0.036
7	71.8	0.082	0.069	16	0.043
12	78.6	0.047	0.041	12	0.051
15	85.4	0.037	0.034	9	0.053
20	95.6	0.023	0.022	3	0.057

Table 6: Where an inventory sits in the distribution of inventories that could have been chosen instead.

inventory	percentile	95% CI
hand-built table, subsets	71.8	[67.6, 75.8]
hand-built table, merges	75.2	[71.4, 79.0]
inventories changed without one (4 languages)	38–47	—
contrast	+27.7	[+23.2, +32.1]

D DISPERSION IS NOT DISPLACEMENT

Each effect in §4 was first reported as the quantity natural to its analysis. Brought to one unit, the displacement of the SSL–spectral gap in percent of the anchored interval, two of them change rank. Table 7 gives both for each choice.

Table 7: The quantity each choice was first reported by, against the displacement of the SSL–spectral gap it produces. The value column mixes units by construction: that is what the table documents.

choice	quantity first reported	value	displacement (% interval)
composition	between-draw σ of the `hubert_large` − `mfcc_40` gap, $k = 5$	0.1444	+10.7% (IQR, $k = 14$)
inclusion threshold	median differential effect on the same two-model gap	0.0309	+8.6%
granularity	spread across systems of the per-system shift	0.0564	+8.3%
distance convention	spread across models of the per-model shift	0.0361	+7.2%
layer selection	simulated inflation of an SSL–baseline margin	5.2%	-4.9%
projection rank	slopes of a two-model gap against the number of pairs	—	-4.2%
aggregation	spread across models of the per-model shift	2.33%	+0.8%

E THE RATIO, AND THE CONVENTIONS OF OUR OWN PIPELINE

Table 9 gives the ratio of between-draw dispersion to the effect of cardinality under the two choices that were not declared when it was first computed, the distance convention of the draws and the summary of their dispersion. Its denominator is not distinguishable from zero, which is why §4.2 reports the two quantities separately. Table 10 lists what a change of distance convention displaces in our own pipeline, from a numeric shift to a change of verdict.

The reversals of verdict in Table 8 do not come from our pipeline. Under the change of distance convention, the slope of RSA against the size of the condition set, at rank 2 on the full domain, changes by a factor of 2.5, while that of CKA is identical under both conventions, at +0.00722: the representation-side distance is free for RSA and fixed by the linear kernel for CKA.

Table 11 separates what the size of the condition set does from what its composition does (§4.2), and Table 12 gives the displacement of each anchor when the two aggregation estimators are matched (§4.4).

Table 8: RSA minus CKA slope, paired, by projection rank, domain and distance convention.

rank	domain	distance	difference	95% CI	excludes 0
2	full	euclidean	-0.0004	[-0.0027, +0.0020]	no
2	$k \geq 11$	euclidean	+0.0013	[-0.0014, +0.0039]	no
4	full	euclidean	+0.0013	[-0.0010, +0.0035]	no
4	$k \geq 11$	euclidean	+0.0083	[+0.0066, +0.0102]	yes
10	$k \geq 11$	euclidean	-0.0049	[-0.0068, -0.0030]	yes
2	full	cosine	+0.0097	[+0.0062, +0.0130]	yes
2	$k \geq 11$	cosine	+0.0142	[+0.0106, +0.0181]	yes
4	full	cosine	+0.0010	[-0.0021, +0.0038]	no
4	$k \geq 11$	cosine	-0.0009	[-0.0033, +0.0013]	no
10	$k \geq 11$	cosine	-0.0009	[-0.0029, +0.0010]	no

Table 9: Ratio of between-draw dispersion to the effect of cardinality, under both undeclared choices.

distance	summary	σ	range	ratio
Euclidean	mean	0.080822	0.004624	17.5
cosine	mean	0.066793	0.004768	14.0
Euclidean	median	0.061235	0.004624	13.2
cosine	median	0.047017	0.004768	9.9

Table 13 is the closing comparison of §4.5, read by Figure 1 in the body.

F Alternatives to structural comparison

Two common alternatives show why a reference-based framing differs from what is usually done; neither compares structure. Fitting a map, as in linear probing (Alain & Bengio, 2016) or the encoding models of neuroscience, makes the score depend on the capacity of the class the map is drawn from, so that with a rich enough class everything becomes predictable from everything; what is measured is decodability, which is why probes are now paired with control tasks that bound what a probe can learn unaided (Hewitt & Liang, 2019). Comparing systems by their behaviour is silent about representation: error consistency stays low between systems whose accuracies match closely (Geirhos et al., 2020), and systems with very different internal geometries can produce similar error profiles (Stafford et al., 2026).

The control tasks of Hewitt & Liang (2019) are the nearest relative in that literature of what §3 does, in that both bound a score by something the comparison does not supply. They differ in what the bound is about: a control task bounds what a probe can learn from a representation, whereas the reference geometry here states what the conditions ought to look like, so a choice that changes the answer can be shown to make it wrong about that claim rather than merely unstable.

G Languages

Fifteen of the 102 languages of FLEURS were used. They were chosen to vary along the three dimensions we expected to bear on the geometry, the size of the vowel inventory, the presence of lexical tone or pitch accent, and language family, rather than by resource availability, since the size of the condition set is itself one of the quantities under study. The resulting sample spans eleven families, includes three languages with complex tone systems and two with simple ones, and ranges from five vowel categories to twenty-eight. FLEURS itself covers 102 languages with roughly twelve hours of speech per language, recorded by three native speakers each from a common set of sentences; its content is parallel across languages, a property few multilingual corpora have and one that matters here, since the language is the unit of replication.

Table 10: Quantities displaced by a change of distance convention, measured in our own pipeline.

quantity displaced	magnitude	nature
raw formants vs identity control	1.61 units	numeric
SSL–spectral gap	$\times 1.77$	numeric
percentile of the published inventory	18.6 points	numeric
pooled coverage R^2	$\times 4.8$	numeric
RSA-vs-CKA verdict	**4 of 5 cells**	**qualitative**

Table 11: Mean gap between two representations, and its dispersion across draws of which categories to include, by condition-set size.

k	mean gap	σ across draws	SE of mean
5	0.055	0.144	0.0065
7	0.054	0.082	0.0037
12	0.059	0.047	0.0021
15	0.058	0.037	0.0017
20	0.058	0.023	0.0010
28	0.059	—	—

H REPRESENTATIONS

A dagger on a revision marks a checkpoint for which the extraction cache holds two revisions and the pipeline did not record which was used (Appendix L); a control inherits the dagger of the trained model whose configuration it reads.

Every representation is evaluated at the layer selected as described in §3; §4.4 reports what that selection is worth.

Random-weight controls. The nine controls are three architectures, those of `hubert_large`, `wav2vec2_base` and `xlsr53`, each instantiated under seeds 0, 1 and 2. A control reads the configuration of its trained checkpoint, sets the seed with `torch.manual_seed` and builds the model from that configuration alone, through `AutoModel.from_config` of `transformers` 4.57.6 (PyTorch 2.6.0); no trained weight is loaded, the convolutional feature encoder included. The weights therefore follow the library's default initialisation for that class: linear layers drawn from a centred normal of standard deviation 0.02 (the checkpoint's `initializer_range`) with zero biases, normalisation layers set to the identity, and convolutions drawn from a He normal; the wav2vec 2.0 class additionally gives its positional convolution and its feature projection laws scaled by their fan-in. Layer count and width are those of the trained model (Table 15), so each control differs from its trained counterpart by training alone. A control is scored like any other representation: for each language, at its best layer, then the median over the fifteen languages. The floor of the anchored scale is the median of these nine scores. It pools architectures and seeds: it says where an untrained network of this family falls, not how much one architecture varies from seed to seed.

Vision checkpoint. The out-of-modality floor is one checkpoint, `google/vit-base-patch16-224-in21k`, at the revision given in Table 15: a ViT-Base with patches of 16 pixels on 224-pixel images, pre-trained on ImageNet-21k, as its identifier states, and never exposed to audio. Audio reaches it as images. The signal, resampled to 16,000 Hz, gives one log-mel spectrogram (128 bands, an FFT of 2,048 samples, a hop of 32 ms, decibels relative to the utterance maximum); the spectrogram is cut into windows of 480 ms, each window is rescaled to $[0, 1]$ by its own minimum and maximum, resized bilinearly to 224×224 and copied to the three colour channels, without the ImageNet normalisation. One window gives one frame: the class token of each of the 13 hidden states (the patch embedding and each transformer block), of width 768. The two rows of Table 15 are two strides between windows, named 50 Hz and 10 Hz in the configuration. Because windows start on whole mel frames, the first advances by one mel frame (32 ms, 31.25 frames per second) and the second by 96 ms (10.4 frames per second); the names are kept as labels. A vowel is represented, as for every other model, by the mean of the frames that fall in

Table 12: Displacement of each anchor of Table 1 when the representation side is pooled with a coordinate-wise median instead of a mean.

family	anchor	displacement	% interval
pre-trained speech and audio models	0.9003	-0.0074	-1.49
non-learned spectral features	0.8535	-0.0114	-2.29
learned compression	0.8421	-0.0019	-0.37
raw formant measurements	0.7571	-0.0160	-3.22
architecture alone	0.5024	-0.0009	-0.18
out of modality	0.3298	-0.0108	-2.16

Table 13: Every choice on one unit: the displacement of the gap between pre-trained speech and audio models and non-learned spectral features, in percent of the anchored interval.

choice	extremes	n	displacement	% interval	95% interval
composition (IQR, $k = 14$)	25th vs 75th percentile draw	1 language, 500 draws	+0.0533	+10.7%	[+9.5, +12.1]
composition (σ, $k = 14$)	between-draw dispersion	1 language, 500 draws	+0.0435	+8.7%	[+8.1, +9.3]
inclusion threshold	0 ms vs 100 ms	15 languages	+0.0428	+8.6%	[-4.1, +21.1]
granularity	28 categories vs 12 categories	1 language, 1000 token resamples	+0.0411	+8.3%	[-7.9, +16.6]
distance convention	Euclidean reference vs cosine reference	15 languages	+0.0358	+7.2%	[-5.4, +16.5]
layer selection	argmax vs relative depth 0.05	15 languages	-0.0242	-4.9%	[-12.0, -0.8]
projection rank	rank 2 vs rank 10	11 languages	-0.0210	-4.2%	[-18.5, +8.5]
aggregation	mean vs median	15 languages	+0.0039	+0.8%	[-2.5, +4.0]

its interval, and the layer is selected as for every other model. The two strides give two sets of frames, hence two scores and two rows. That there are two is a matter of record rather than of design: the coarser stride was introduced as a speed measure, raising the configured hop from 20 to 100 ms so as to cut the number of windows fivefold, and the two were then kept side by side deliberately, as an ablation on whether temporal resolution changes what is recovered of the vowel geometry.

Two properties of that construction are worth stating plainly, because they bound what the last column of Table 15 may be used for. A window is labelled by the mel frame it starts on rather than by its centre, and the frame rate is recovered after the fact, as the number of frames divided by the duration of the utterance; the correspondence between a time in the signal and a frame index therefore drifts along the utterance rather than holding exactly. That column is accordingly descriptive. It names a configuration and enters no denominator anywhere in this paper – in particular not the anchored interval of §3, which is computed from scores and never from frame rates – and no quantity reported in the body is derived from it.

I WHY NO INVENTORY IS CANONICAL

§1 claims that how many degrees of vowel height to recognise, or whether a diphthong is one condition or two, are contested questions rather than matters of presentation. This appendix makes that claim good, states what our corpus supplied and what we did with it, and gives the inventory used for every language.

The count is an analytic decision, not an observation. A vowel inventory is not read off a signal. It is the output of decisions about which contrasts are phonemic, and four of those decisions recur across the descriptive literature (Ladefoged & Maddieson, 1996). Whether length is a property of the vowel or of the syllable determines whether a long and a short vowel are one category or two. Whether a diphthong is a single moving target or a sequence of two determines whether it occupies one row of the matrix or two. Whether centralised or reduced vowels are categories in their own right or allophones of full vowels determines how many rows unstressed syllables contribute. And how finely height is divided, into three levels or four or five, is a choice about which the descriptive tradition of a single language can change within a generation. Each decision is defensible either way, each changes N, and none of them is a decision about the recording.

Two defensible inventories for the same language, in this paper. We do not have to look outside our own material for an example. The English inventory of our corpus carries twenty-eight categories as transcribed, and the merge table of §4.2 reduces it to twelve. Both describe the same recordings.

Table 14: The fifteen languages, ordered by the size of their inventory.

language	FLEURS code	family	tone (WALS 13A)	k	vowel tokens	split-half ceiling	grid step
Spanish	es_419	Romance	none	5	2,500	1.000	1.21e-02
Mandarin	cmn_hans_cn	Sinitic	complex	7	3,500	0.998	1.30e-03
Polish	pl_pl	Slavic	none	8	4,000	0.997	5.47e-04
Swahili	sw_ke	Bantu	none	10	4,063	0.996	1.32e-04
Japanese	ja_jp	Japonic	simple	12	5,765	0.996	4.17e-05
Hausa	ha_ng	Chadic	simple	13	6,153	0.997	2.53e-05
Korean	ko_kr	Koreanic	none	13	6,479	0.997	2.53e-05
Portuguese	pt_br	Romance	not coded	14	7,000	0.999	1.59e-05
Russian	ru_ru	Slavic	none	14	6,684	0.995	1.59e-05
French	fr_fr	Romance	none	15	7,444	0.998	1.04e-05
German	de_de	Germanic	none	16	8,000	0.999	6.94e-06
Vietnamese	vi_vn	Austroasiatic	complex	17	8,025	0.998	4.77e-06
Thai	th_th	Tai-Kadai	complex	18	8,636	0.999	3.35e-06
Turkish	tr_tr	Turkic	none	20	7,491	0.989	1.75e-06
English	en_us	Germanic	none	28	11,718	0.968	2.22e-07

Table 15: The twenty-seven representations, grouped by the families of Table 1.

representation	checkpoint	revision	layers	width	layer used	
pre-trained speech and audio models (8)						
hubert_large	facebook/hubert-large-ll60k	ff022d0†	25	1024	5 (0–24)	
mert_330m	m-a-p/MERT-v1-330M	5240c27†	25	1024	8 (1–22)	
mms_300m	facebook/mms-300m	4ee317c†	25	1024	2 (0–18)	
wav2vec2_base	facebook/wav2vec2-base	0b5b8e8†	13	768	1 (0–12)	
wav2vec2_english	facebook/wav2vec2-base-960h	22aad52	13	768	0 (0–9)	
wavlm_large	microsoft/wavlm-large	c1423ed†	25	1024	3 (0–24)	
whisper_medium	openai/whisper-medium	abdf7c3	25	1024	2 (0–5)	
xlsr53	facebook/wav2vec2-large-xlsr-53	c3f9d88†	25	1024	2 (0–17)	
learned compression (4)						
dac_encoder_24khz	DAC 24 kHz (library weights)	—	1	1024	0	
encodec_encoder_24khz	EnCodec 24 kHz, 6 kbps (library weights)	—	1	128	0	
mimi_encoder_24khz	kyutai/mimi	89091b3	1	512	0	
speechtokenizer_encoder	fnlp/SpeechTokenizer	4d54939	1	1024	0	
non-learned spectral features (3)						
logmel_80	no training	—	1	80	0	
mfcc13_delta	no training	—	1	39	0	
mfcc_40	no training	—	1	40	0	
raw formant measurements (1)						
formants_f1f4	F1–F4, formant tracking (no training)	—	1	4	0	
architecture alone (9)						
hubert_large_random_seed0	facebook/hubert-large-ll60k	ff022d0†	25	1024	0 (0–18)	seed 0
hubert_large_random_seed1	facebook/hubert-large-ll60k	ff022d0†	25	1024	5 (0–24)	seed 1
hubert_large_random_seed2	facebook/hubert-large-ll60k	ff022d0†	25	1024	1 (0–21)	seed 2
wav2vec2_base_random_seed0	facebook/wav2vec2-base	0b5b8e8†	13	768	0 (0–12)	seed 0
wav2vec2_base_random_seed1	facebook/wav2vec2-base	0b5b8e8†	13	768	0 (0–12)	seed 1
wav2vec2_base_random_seed2	facebook/wav2vec2-base	0b5b8e8†	13	768	0 (0–9)	seed 2
xlsr53_random_seed0	facebook/wav2vec2-large-xlsr-53	c3f9d88†	25	1024	3 (0–24)	seed 0
xlsr53_random_seed1	facebook/wav2vec2-large-xlsr-53	c3f9d88†	25	1024	2 (0–18)	seed 1
xlsr53_random_seed2	facebook/wav2vec2-large-xlsr-53	c3f9d88†	25	1024	1 (0–14)	seed 2
out of modality (2)						
vit_imagenet_10hz	google/vit-base-patch16-224-in21k	b456956	13	768	6 (1–10)	10 Hz
vit_imagenet_50hz	google/vit-base-patch16-224-in21k	b456956	13	768	6 (1–12)	50 Hz

The first follows the phone set of the aligner (McAuliffe et al., 2017); the second follows the lexical sets of Wells (1982), which group vowels by the words that carry them rather than by a phonemic analysis, and which are the standard descriptive device for comparing English accents whose phoneme counts differ. Neither is a coarsening of a true inventory that the other reports correctly. §4.2 measures what passing between them does: every score rises, and it rises most for the weakest representations.

What we used, and what we built. Inventories are used as observed. For each language the categories are the vowel phones the aligner's phone set distinguishes, with no merging, which is why the sample ranges from five categories to twenty-eight without our having chosen that range. Two exceptions are declared. The twelve-category English merge table is ours: it was built for this project rather than borrowed, organised by the lexical sets above, and archived with the configuration.

Table 16: The inventory used for each language, as the aligner transcribes it, with no merging.

language	k	categories
Spanish	5	a e o i u
Mandarin	7	i o a u e y ə
Polish	8	a ɛ ɔ ɨ i u ɔ̃ ɛ̃
Swahili	10	ɑ i u ɛ ɔ ɑː uː iː ɛː ɔː
Japanese	12	a o i e oː ɯ eː ɨ ɨː iː aː ɯː
Hausa	13	aː a ɪ ə eː oː iː ʊ ɛ uː e ɔ o
Korean	13	ɐ i ʌ o e u ɛː ɨ ʌː iː oː eː uː
Portuguese	14	ɐ i a u ɨ o e ɐ̃ ẽ ɛ ɔ õ ĩ ũ
Russian	14	a ɐ o e ə i ɨ ʊ ɪ u ɛ æ ɵ ʉ
French	15	a e i ɛ ɑ̃ ɔ ɔ̃ y ɛ̃ o u ø œ ɑ ə
German	16	ə a ɪ ɐ ɛ iː eː aː ʊ ɔ oː uː yː ʏ øː œ
Vietnamese	17	aː iː əː ɔː a ɨ oː ə eː ɨː ɔ i ɛː u uː o e
Thai	18	aː a iː ɔː o ɛː i uː eː oː e ɛ ɤː ɯː u ɯ ɔ ɤ
Turkish	20	a e i ɯ o u y ø ɛ ɨ ʊ aː ɯː uː eː oː øː iː ʏ yː
English	28	a i ə ɛ ɪ ɒ e iː ɐ ɑ ɒː ɔ æ eː ɚ ʉː uː aː ɑː ɜː oː ʉ o ɝ ɛː u ʊ ɜ

And the draws of §4.2 are subsets of the observed categories, never arbitrary partitions, so that every condition in every draw is a category the language actually distinguishes (Appendix B).

The same underdetermination outside speech. Nothing in the argument is particular to vowels. A comparison over ImageNet classes must decide which classes and at what level of the hierarchy, and the answer is not given by the images; a comparison against neural recordings must decide which experimental conditions count as distinct, and that is the question the experiment was designed to ask rather than one it answers. The reference geometry is what our domain supplies and most do not (§3); the underdetermination of the condition set is what our domain shares with them.

J THE LAYER-SELECTION ARTEFACT IN VISION

The layer-selection artefact of §4.4 reproduces in a second modality. On dSprites, whose images are generated from known factors, three trained vision models (DINOv2, CLIP, ResNet-50) are compared with four unlearned baselines under the three protocols of Table 17. Under the best layer per condition, a baseline contests one of the twelve pairs; at one relative depth for every model, the rule of §4.4, it contests five, and raw pixels are ahead of all three trained models. The artefact is therefore more visible here than in speech, where no baseline contests a model under either protocol. The setting also has a property ours lacks: its reference geometry is exact rather than measured, which would make the *bias* of the estimator measurable there, and not only its variance. The composition of the condition set was not tested in vision, and the result of §4.5 remains a result about speech.

K WHAT THIS PAPER DOES NOT ESTABLISH

The paragraph of §5 states these in brief. Each is given here with what would be needed to lift it.

The two decisions do not have the same reach. §4.2 and the estimator of §4.4 bear on how inputs are grouped, and therefore on work in which a row aggregates several inputs: the neuroscience tradition these measures come from, analyses at the level of categories, and comparisons between a model and trial-averaged recordings. §4.3 bears on which inputs are admitted, which applies wherever these measures are used. A reader whose rows are individual examples should take the second and not the first.

One domain, and one language for two results. Vowel geometry is the instrument. One result travels: the layer-selection artefact of §4.4 reproduces, more sharply, on a dataset whose generative factors supply the reference (Appendix J). The others were not tested in a second modality, and whether their magnitudes transfer is not established here; what transfers in any case is the argument

Table 17: The layer-selection artefact in a second modality (dSprites; three factors × three values of k = nine conditions).

	argmax	matched depth	leave-one-out
pairs a baseline contests, of twelve	one	five	three
mean ρ over the nine conditions			
DINOv2-base	0.629	0.448	0.530
CLIP ViT-B/16	0.649	0.466	0.561
ResNet-50	0.731	0.605	0.605
raw pixels	0.693	0.693	0.693
PCA-64 on pixels	0.568	0.568	0.568
Gabor bank	-0.057	-0.057	-0.057
HOG	0.564	0.564	0.564

that a choice upstream of the matrix is beyond the reach of any invariance. Within the domain, the composition and granularity results rest on English alone, the only language of our sample with twenty-eight categories.

The space is not enumerated. Two choices that meet our criterion are not examined: the units the reference is expressed in, and how representations are normalised. We do not claim to have enumerated the space.

Two bifurcations, two sets of models. The duration threshold was swept over eight representations and the distance convention over twenty-seven, so the wider interval on the first reports the count rather than a less stable ranking. Extending the sweep is a matter of compute and not of method, and the Reproducibility Statement says what it would take.

What we did not measure. The instrument would also permit asking which measure best recovers a known geometry at a fixed condition set. We point at this rather than pursue it, because §4.1 implies the answer would itself be conditional on the choice this paper is about.

L THE RECORD: PREDICTIONS AND DEFECTS

A paper about choices that go unreported owes an account of its own. We registered each prediction before computing it and kept a journal of the defects found in our own pipeline. This appendix gives both, grouped by what would have been needed to notice them; Appendix M shows which predictions can be checked from outside the project and which cannot.

Two things in the record bear on how the results of §4 should be read. Half the predictions we registered were refuted outright, among them two whose premise was correct and whose conclusion was not (§4.3, §4.4); and the defects that cost most were those where the artefact was internally consistent, since an error that produces a plausible number and a file that is merely stale both survive every check that asks whether the pipeline ran.

Predictions registered, and refuted. Of the ten predictions we registered, five were refuted outright, four were confirmed only in part and one was never tested. Two refutations are worth more than a line, because what they overturned was a way of reasoning rather than a number.

The premise held and the conclusion did not. We predicted that matching the two aggregation estimators would narrow the gap, on the ground that the spectral baselines carry the heavier-tailed within-category distributions. The premise is correct and the gap widens (§4.4). A mechanism can be right about what it describes and wrong about what follows from it.

The cheap diagnostic, twice. Twice we predicted that a quantity available without recomputation would predict a bifurcation's differential effect, and twice the diagnostic ordered the systems correctly

on the property it measures and failed to predict the consequence (§4.3, §4.4). The pattern is the finding, and it is what makes the recommendation to sweep rather than diagnose worth stating.

A third case never reached a result. An early version of the unaggregated analysis produced a correlation near -0.9 with a p-value indistinguishable from zero, between the number of categories present in a draw and the dispersion of its level. The quantity says nothing: dispersion is constant within a level, so the correlation only restates that both terms move with the number of rows, and it would have come out large whatever the mechanism. It was caught because the protocol addendum, written three days earlier, had specified the test at fixed n, which is the one the prediction engaged and which returns nothing. This is the failure of §4.2's ratio in another guise: a statistic whose value is imposed by its construction rather than by the data.

One choice we declined to change. Late in the work we found that the family we had been calling self-supervised contained two supervised encoders. Redefining it would have been defensible and we did not, because the panel was named in the pre-registration of the closing analyses and changing it after seeing their results is the move this paper documents in others. We report the sensitivity instead (§4.5).

Four classes of defect. The defects we found fall into four kinds, distinguished by what would have been needed to notice them.

Class A: an error that produced a plausible result. An asymmetry in how the two sides were pooled produced an effect four times its true size in one language, which stood for six weeks as the motivating result of an earlier draft. It was plausible, it was in the right direction, and nothing about the output marked it. The entry below is of the same class, with a mechanism the others do not share.

A storage type decided the order of the smallest distances. The out-of-modality anchor of Table 1 was printed as 0.3280 in one table and 0.3298 in another. The second was the value computed in double precision and the first the value computed in single precision, and it took three hypotheses, two of them wrong, to establish it. Until its last regeneration, the pipeline computed cosine distances between category centroids held in single precision. For most representations this is harmless. It is not for those whose centroids are nearly collinear: randomly initialised networks, the vision model and the last layers of some trained encoders, where the smallest cosine distances are of the order of 10^{-6} and the subtraction that defines them loses most of their digits to cancellation. There the order of the small distances, which is all a rank correlation sees, was set by the last bits of the representation. Of the 5,985 model–language–layer values, 698 moved by more than 10^{-5} once the distances were computed in double precision, and the largest moved from 0.794 to 0.891, or 19.5% of the anchored interval. At the layers this paper reports, the displacement is at most 0.0066, and the default layer changed for one model–language pair. No conclusion changed, and the anchored interval moved in its fifth decimal.

Two features make the case worth recording beyond its correction. The loss strikes where there is no structure: the representations it affects are those that bound the scale from below, against which every other value is read. This is the lattice of §4.1 seen from the other end. There, the estimator cannot resolve below the step of its grid; here, it cannot resolve below the precision of its storage, and near either limit the order of the distances is decided by something other than the representation. And the natural remedy would have hidden the problem. Rounding both computations, or recomputing in single precision to match, produces agreement, and agreement is more plausible than disagreement: the obvious diagnostic confirms instead of revealing. The build now declares every artefact that carries this quantity and stops if any two disagree beyond the rounding of storage.

Class B: the correction depended on what was neither in the code nor in the output. A switch that controlled the pooling above was read from a configuration file that nothing wrote to, so that setting it changed nothing and the code that consumed it was correct. A second instance is below; a third is that the artefacts on disk predate a change to the extraction that would have made a later analysis possible, and nothing about them says so: they load, they carry the expected keys, and the only symptom is a downstream script refusing a file that looks intact.

What identified a computation was not written down when it was made. Twice, a piece of information that was available at the moment of computing was not stored with the result, and could

not be recovered afterwards. The first is the identity of the token. Our store of representations kept each vector and its phone label and dropped what linked it to the token it came from, so that no row could be put back beside the duration or the formants measured on the same segment. Nothing was falsified; two analyses simply became impossible. The duration sweep of §4.3 could not go beyond 8 of 27 representations, and the unaggregated arm could not be run until the representations were extracted again with their provenance. The second is the revision of each checkpoint. Of the eleven checkpoints taken from a public hub, six have two revisions in the cache of the machine that extracted them, and nothing the pipeline kept says which one was used (Appendix H). This prevents nothing that we report; it weakens the claim that anyone can reproduce it. The two costs differ, one closing an analysis for months and the other opening a doubt that cannot be closed, but the mechanism is the same, and so is the remedy (Appendix N). A variant of it met us in the vision arm: two files of the same name, at the same path on two machines, one a partial run and the other the complete one, and nothing in either said which was which. Like a table that has stopped being generated but is still included, each looked right and only one was.

Class C: the code was correct and gave two answers. The percentile of §4.2 moved by seven points between two runs of the same script on the same data. Neither run was wrong. The draws were generated from a shared random stream, and an earlier part of the analysis consumed a different number of values in each run, so the subsets drawn afterwards differed. The defect is not in any line of the code but in the coupling between two parts of it that appear independent.

Class D: the check did not check. Three controls ran, reported nothing, and did not look where we believed they looked: the pooling switch above, a test suite that collected no test file and therefore passed, and the entry below. This is the class we would expect to be rarest and it is the best attested.

A check that passed because it did not look. The build checks that no table is wider than the text, since an over-wide table does not fail: it runs into the margin, and what passes the edge of the page is not printed. The check read the compilation log and attributed each overfull box to the file open when it was reported. Parentheses in the log that are not file names closed the table's file early, so that its overflows were attributed to nothing and the check passed. Meanwhile a column of the closing table ran past the page, in a build that reported no error. The check existed, it ran, and it did not look where we believed it looked; it now fails on any overflow it cannot attribute.

M PREREGISTRATIONS

Table 18 gives, for each prediction of Appendix L, the first commit that carries it and what that commit can prove. Not every prediction can be checked from outside: two were written in a plan that is not versioned, and the journal that records the verdicts entered version control only after the first six had been computed. We report the discipline as it was, imperfect, rather than as it should have been, because it still caught what Appendix L describes.

N A CHECKLIST

The items below follow from what §4 showed to be consequential and from what Appendix L shows to be easy to miss. Several of them we wrote after the defect they would have caught. The first three are the findings of this paper turned into practice and are stated in §5; the rest come from our own record.

- **Report the condition set, category by category.** No summary of the set stands in for the set: four coverage statistics capture a tenth of what is at issue (§4.2).
- **Report the inclusion rule, and sweep it.** Which inputs are admitted take a comparison from -0.047 to -0.002 in one language of fifteen, and no diagnostic available without recomputation predicts what it will do (§4.3).
- **Compare at matched depth, or say that you did not.** Reporting each system at its best layer compares a maximum over many draws with a maximum over one, and accounts for 46% of the margin (§4.4).

Table 18: The predictions of Appendix L, with the commit that carries each before its computation, and what that commit proves.

	prediction	first commit	verifiable by	verdict
P1	The noise ceiling falls mechanically with the number of pairs	`5f2d11b`, 2026-09-08	the script that states it, before it ran	refuted, in the opposite direction
P2	RSA, CKA and Procrustes columns nearly identical on the first bifurcations	—	no external trace	refuted
P3	After normalisation, the SSL - MFCC gap separates tonal from non-tonal languages	—	no external trace	stated; not tested
P4	The Procrustes slope in k survives a rank control	`3a3221c`, 2026-09-08	the script that states it, before it ran	refuted, sign reversed at rank 2
P5	The threshold's differential effect is largest in Russian	`3452e65`, 2026-09-08	the script that states it, before it ran	refuted on both parts
P6	The paired estimator narrows the SSL - spectral gap	`0c8db60`, 2026-09-15	the script that states it, before it ran	refuted, the gap widens
P7	Without aggregation the between-draw dispersion becomes negligible	`6354f5d`, 2026-09-18	its preregistration file	principal confirmed, subsidiary refuted
P8	Composition ranks above the distance convention on one unit	`e47871c`, 2026-09-23	its preregistration file	confirmed; subsidiary refuted under the IQR
P9	The three measures rank the representations alike	`e47871c`, 2026-09-23	its preregistration file	confirmed over 27, partial over 16
P10	The ranking of representations survives composition	`13855b9`, 2026-09-23	its preregistration file	confirmed at k = 14, collapses at k = 5

- **Check that the estimator can resolve the comparison.** A rank correlation moves on a grid whose step grows as the condition set shrinks, and below some size the difference reported is smaller than the step (§4.1).
- **Report a noise ceiling, and state what it does not cover.** A split-half reliability bounds what the reference can support; a systematic measurement error reproduces itself in both halves and leaves it untouched, so the control is blind to exactly the failure that would matter most (§3).
- **Compute distances in the precision in which they will be compared.** A rank correlation is decided by the order of the smallest distances, and the representations that bound a scale from below are those whose smallest distances fall under the resolution of single precision (Appendix L).
- **Record what identifies what was computed, at the moment it is computed.** The token behind each representation and the revision of each checkpoint cost nothing to store then and cannot be recovered later (Appendix L).
- **Before reporting a correlation, check that both its terms vary within the unit it is computed over.** A statistic whose value is imposed by its construction comes out large whatever the mechanism, and a ratio whose denominator is not distinguishable from zero has no value to report (§4.2).
- **Make a control fail when it cannot check.** Three of the defects we found passed a control that ran and did not look (Appendix L).

A checklist is not a guarantee, and Appendix L is the evidence: we were following one, and it did not contain these items until each of them was needed.