Softmax Masking Is a Choice Model: The Evidence for Linear versus Gaussian Renormalization of Probability Vectors

Peter Cotton

Provisional draft, 17 August 2026

Abstract

When a probability vector over mutually exclusive outcomes must be transported to a smaller support, because an outcome is withdrawn, scratched, delisted, disallowed or masked out of a softmax, proportional renormalization is the reflex and in many models the architecture. It is the Gumbel point of the independent additive random-utility class[2]. Re-running a Gaussian contest among the survivors is Case V. Psychology has tested renormalization under restriction since 1957, as the constant-ratio rule, and we score the two out of sample on thirty-nine population comparisons, twenty-seven observing the smaller menu and twelve inducing subset choice from rankings. Gaussian renormalization has the lower held-out log loss in thirty. The rows are dependent and unequally clustered across studies, so that count describes this corpus rather than estimating a rate. The losses concentrate on one-dimensional and strongly confusable stimulus sets, which motivates a boundary rule rather than establishing one.

1 A two horse race between zero parameter choice models

A probability vector is given over a set of mutually exclusive outcomes, and the set changes. A horse is scratched, a product is delisted, a candidate fails to qualify for a ballot, a response option is disallowed, a token is masked out of a softmax, a distribution is truncated to a sub-support. In each case a probability vector has to be transported from the original support to a smaller one. What are the probabilities over what remains?

We take the case where the original vector is a population's choice shares over a menu, since that is where data exists to score the answers, but the operation is the general one and the two answers below apply wherever it is performed.

Nothing has been observed about the smaller menu, so the answer must come from an assumption about how the original shares arose. The standard assumption is stated by Luce's axiom [5], which assigns each item i on menu S a worth $u(i)$ with

$$p_i = \frac{u(i)}{\sum_{j \in S} u(j)}, \tag{1}$$

so the shares determine the worths up to scale. The prediction for a subset $T \subseteq S$ follows by changing the denominator alone, which leaves every ratio p_i/p_j with $i, j \in T$ as it was in S.

[2]The Gumbel construction is due to Holman and Marley [35, 33]; the uniqueness is Yellott's [39].

For a two-item menu the predicted share preferring i is $p_i/(p_i + p_j)$. This is proportional, or *linear*, renormalization. The invariance it imposes is Independence of Irrelevant Alternatives, and applied to a reduced set of response alternatives it is what Clarke [12] named the *constant-ratio rule*.

There is another approach that is equally natural, if not more so, and also demands no estimation or free parameters. Read the shares as the winning probabilities of a contest among independent Gaussian performances, which is Thurstone's Case V [37] generalized from pairs to a multiway field. Item i carries a location a_i, draws $X_i \sim N(a_i, 1)$ on each occasion, and the highest draw is chosen, so

$$p_i = \int \varphi(x - a_i) \prod_{k \neq i} \Phi(x - a_k)\, dx. \tag{2}$$

Given the shares, this is inverted for the locations, which are identified up to a common translation and fixed here by $\sum_i a_i = 0$. The prediction for a subset T is the same contest run among T alone. For a pair it is one normal evaluation,

$$P(X_i > X_j) = \Phi\left(\frac{a_i - a_j}{\sqrt{2}}\right). \tag{3}$$

The two are termed *linear renormalization* and *Gaussian renormalization* throughout. Both take the same shares and return shares over the survivors. The first holds the ratios of the shares fixed, the second holds the latent locations fixed.

Each map travels under several names, and we treat the names of each as synonyms. For the first: proportional renormalization, linear renormalization, the ratio rule, Clarke's constant-ratio rule, and masking a softmax and rescaling the survivors. All denote Equation (1) applied to a subset, and we write linear renormalization except where the historical name is what a cited paper used. For the second: Gaussian renormalization, Thurstone's Case V re-run among the survivors, and the independent unit-variance Gaussian contest on the restricted menu. The names differ by discipline and by decade rather than by content, and part of why the comparison below had not been made is that the two literatures holding the pieces did not share a vocabulary.

The two horses disagree, and in a fixed direction. If $p_i > p_j$ then Equation (3) lies between $\frac{1}{2}$ and the renormalized $p_i/(p_i + p_j)$: winning a large field requires an unusually high draw and beating a single rival does not, so withdrawing the other alternatives counts the favourite's advantage for less. Section 4.1 shows that these two are the two conventional members of the independent additive random-utility class, and Proposition 1 makes the direction of their disagreement exact.

2 Scoreboard

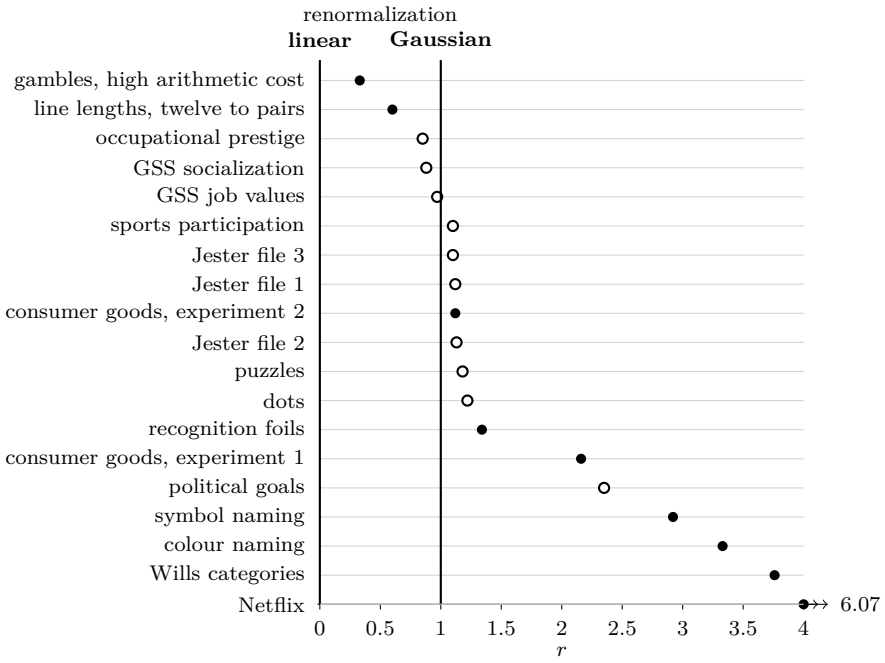

Figure 1: Where each population sits between the two maps. The coordinate is $r = \lambda_{\text{observed}}/\lambda_{\text{Case V}}$, the observed contraction slope divided by the slope Case V implies for that population's own shares, so linear renormalization is at $r = 0$ by construction and Gaussian renormalization at $r = 1$. Filled circles are restricted menus the observer actually faced, open circles are ranking-induced. Most populations lie beyond $r = 1$: choice contracts by more than Case V predicts, so the Gaussian point is not far enough along the family, which is the same signed error the after-stimulus arms and Table 9 report. The two nearest linear renormalization are the gambles, whose shares are concentrated, and the line lengths, the one perceptual continuum here. Nineteen of the thirty-nine comparisons appear; the rest restrict to sets of three or more, so no pairwise q_{ij} exists and no slope can be fitted, which excludes the tone matrices, the Getty conditions and Townsend and Landon. Precision varies by an order of magnitude across rows: the recognition foils rest on 360 pairs, the gambles on five and Netflix on three, so the extremes of this axis are its least certain points. Positions here do not reproduce Table 1: λ projects a vector of pair-specific predictions onto a single slope and is not a proper score, so the twelve-line population sits nearer Case V on this axis while linear renormalization has the lower held-out log loss on it. This axis places populations, the table scores forecasts.

Figure 1 and Table 1 state the result before the argument for it. The figure asks where each population falls between the two maps, on a scale where linear renormalization is one point and Gaussian renormalization another. Of the nineteen it can place, eighteen lie nearer the Gaussian point, the median sits at 1.13, and fourteen lie beyond it. The single population nearer linear renormalization is the gamble row, which rests on five pairs. That includes the perceptual continuum: line lengths sit at 0.60, so even where linear renormalization has the lower held-out log loss it is the further of the two maps from the observed contraction. The table asks the different question of which map forecasts better. Thirty-nine comparisons have full-menu shares

that calibrate both maps and restricted-menu choices that score them by held-out log loss. Rows are comparisons, not independent populations: several reuse the same subjects or source study. Section 5 gives the protocol, the datasets and the null.

Gaussian renormalization has the lower held-out log loss in thirty of the thirty-nine rows. That count is a description of this table and not an estimate of any rate, because the rows are not independent, the families contribute unequal numbers of them, and the collections were assembled by searching for restriction data rather than sampled from a population of applications. The effect sizes and their families are what the table is for.

Eight of the nine rows where linear renormalization predicts better are one phenomenon. Four are tone-identification conditions, three are line-identification restrictions and one is the Getty condition whose four survivors are mutual confusions, all of which place the alternatives on a single perceptual dimension; Section 5.8 treats them together. The ninth is Wikipedia, whose magnitude is -0.0001.

The tail column separates two things that the earlier draft of this table conflated. A positive gain says Gaussian renormalization predicted better with estimated shares, and part of that can be shrinkage. A small tail says the population departs from the constant-ratio rule in the direction of contraction. The consumer experiments appear twice each because the two answers must come from the same analysis population: the forced-choice subgroup admits no deferral coding, the all-subject figure uses the whole sample, and pairing a gain from one with a tail from the other would compare different people. Three further populations appear in Section 5.8 rather than here, because published aggregates fix the direction of their failure without supporting this protocol. Wikipedia has a small tail and a negative gain, so it departs from the rule in the right direction while still predicting worse. Sports participation has a positive gain and a tail of 0.51, so its gain is shrinkage and it carries no evidence about the rule. Section 5.7 treats the second question and Section 5 the first.

3 Prior tests of the constant-ratio rule

Clarke tested the rule on spoken consonants in noise, predicting confusion matrices over reduced response sets from a larger one, and Clarke and Anderson [11] followed with a second test. Pollack and Decker [13] predicted four-way consonant matrices from an eight-way matrix and reported average deviations near four per cent. Hodge [20] ran nested ensembles in three modalities, Rich [21] carried the rule into short-term memory, and Getty, Swets, Swets and Green [19] held eight stimuli fixed while allowing only four responses.

Morgan [22] built a likelihood-ratio test and rejected the rule on the earlier data. Townsend and Landon [23] ran one master set of five letters with three nested response sets, blocked and counterbalanced, and published the full matrices per subject. Rouder [24] states the modern summary: conditioning on a reduced choice set is better than a guessing correction predicts and worse than the ratio rule predicts. Wills, Reimers, Stewart, Suret and McLaren [25] reject the rule in categorization and argue for a Thurstonian replacement, and they go furthest toward the contest studied here: alongside their fitted four-parameter model they build a “simple winner-take-all” system that picks the largest noisy magnitude, note that “like the ratio rule, the simple-WTA system has no free parameters”, and report that Gaussian noise “produces comparable results”.

[4]The news tail rests on sixty replicates, so .016 is the floor 1/61 rather than a measured frequency.

Population	Restriction	Gain	Tail	Family
News slates	observed impressions	+.0496	.016[4]	clicks
Sounds, eight to four, $\{1,3,5,7\}$	observed menus	+.0490	$\le .005$	Getty
Sounds, eight to four, $\{1,2,5,6\}$	observed menus	+.0453	$\le .005$	Getty
Consumer goods 1, forced choice	observed menus	+.0442	.005	Costa-Gomes
Occupational prestige	ranking-induced	+.0390	$\le .005$	prestige
Wills categories, three to two	observed menus	+.0299	.002	categorization
Symbol naming, six to two	observed menus	+.0278	$\le .005$	Yeon–Rahnev
Colour naming, four to two	observed menus	+.0147	$\le .005$	Yeon–Rahnev
Consumer goods 1, all subjects	observed menus	+.0140	.010	Costa-Gomes
Consumer goods 2, forced choice	observed menus	+.0140	.129	Costa-Gomes
Sushi	ranking-induced	+.0111	$\le .005$	preference
Gambles, high arithmetic cost	observed menus	+.0101	.035	Aguiar
Letters, five to three, $\{F,H,X\}$	observed menus	+.0093	.005	Townsend–Landon
Political goals	ranking-induced	+.0069	$\le .005$	preference
Gambles, medium arithmetic cost	observed menus	+.0068	.124	Aguiar
Consumer goods 2, all subjects	observed menus	+.0059	.199	Costa-Gomes
GSS socialization	ranking-induced	+.0055	$\le .005$	GSS
Sports participation	ranking-induced	+.0050	.51	preference
Letters, five to three, $\{A,E,X\}$	observed menus	+.0040	.010	Townsend–Landon
Recognition foils, four to two	observed menus	+.0037	$\le .005$	Utochkin
Jester file 3	ranking-induced	+.0024	$\le .005$	Jester
Jester file 1	ranking-induced	+.0020	$\le .005$	Jester
Jester file 2	ranking-induced	+.0018	$\le .005$	Jester
GSS job values	ranking-induced	+.0017	$\le .005$	GSS
Dots	ranking-induced	+.0015	$\le .005$	perceptual
Netflix	ranking-induced	+.0009	$\le .005$	film
Puzzles	ranking-induced	+.0007	$\le .005$	perceptual
Letters, five to four, $\{A,E,F,H\}$	observed menus	+.0006	.129	Townsend–Landon
US state ballots 2024	observed ballots	+.0002	.005	ballots
Gambles, low arithmetic cost	observed menus	+.0002	.667	Aguiar
Wikipedia links	observed removals	−.0001	.005	clicks
Tones, wide, ten to eight	observed menus	−.0041	n/a	Stewart
Tones, narrow, ten to eight	observed menus	−.0051	n/a	Stewart
Lines, twelve to eight	observed menus	−.0057	1.000	Rouder
Lines, twelve to four	observed menus	−.0090	1.000	Rouder
Sounds, eight to four, $\{3,4,5,6\}$	observed menus	−.0127	1.000	Getty
Tones, narrow, ten to six	observed menus	−.0133	n/a	Stewart
Tones, wide, ten to six	observed menus	−.0173	n/a	Stewart
Lines, twelve to two	observed menus	−.0322	1.000	Rouder

Table 1: Held-out log loss of linear renormalization minus that of Gaussian renormalization, in nats per prediction, so a positive gain means Gaussian renormalization predicted better. The tail is the Monte Carlo frequency with which a population obeying the constant-ratio rule exactly reaches that gain on the same menus, and it therefore tests departure from the rule rather than predictive superiority. Replicates are two hundred except for the news row, sixty, and the Wills row, four hundred; the tone matrices are participant-averaged, so no tail is computed.

The principle is therefore theirs. The map is not, and the difference is in what goes in. Their magnitudes are supplied by a model of the stimulus, a function of how many category-appropriate

symbols a display contains, and their conclusion is conditional on it: the ratio rule is incorrect "given the assumption" that magnitudes take that form. The problem here has no stimulus in it. The input is a probability vector, not a probability vector together with a picture of whatever produced it, so an analyst holding shares alone has nothing to race. Their Experiment 2 nonetheless withdraws an alternative, and their data has been deposited throughout, so it enters the comparison below on the same terms as everything else.

Three findings recur across those studies. The rule is a fair first approximation. It fails by concentrating the withdrawn mass onto whichever survivor most resembles the removed alternative rather than spreading it proportionally, which Townsend and Landon attribute to Debreu and Rouder reaches from the opposite direction. And every claim of conformity that was later retested with a test that had power was overturned. Holloway reversed his own 1968 conclusion in 1971 [16, 17], and Morgan's likelihood-ratio test rejected the rule on both of the datasets he applied it to, Clarke's and Egan's [15, 22].

The reason the question stayed open is stated in the literature itself. Clarke returned to it in 1959 with four kinds of signal ensemble and three competing models, and reported that "the simplified version of the constant-ratio rule and the simplified version of the theory of signal detectability were both compatible with the data obtained in the speech experiments" [14]. The theory of signal detectability is the Gaussian account. So the two candidates compared in this paper were already run against each other in 1959, and the verdict was that percent correct could not separate them. The confusion matrices that would have separated them were printed but never used for that purpose.

Clarke's 1959 abstract also records the boundary condition of Section 5.8, in the same sentence pattern: "No model tested was sufficiently complex to account for data when the sinusoidal signals varied only in amplitude or only in frequency". Both accounts failed on unidimensional continua at the outset, which is where the tone and line collections below place them.

What the literature does not contain is a comparison against a second assumption carrying no free parameters either. Lee [34] proposed one analytically, observing that detection theory and the constant-ratio rule diverge most for unidimensional stimuli and that the gap could serve in diagnosis, but did not carry it out on data. Nor was any of this scored out of sample; the tests are goodness-of-fit statements about the matrix in hand. Appendix A is a census of sixty-seven sources, recording which of them printed or deposited enough to score either assumption.

4 Theory

4.1 Why these two maps

Renormalization is not a consequence of probability theory. It is the posterior for mutually exclusive events only when the information that one is impossible says nothing about which of the rest obtained, and Monty Hall is the familiar counterexample. Withdrawal is a third case, neither flat-likelihood conditioning nor informative conditioning, because no information about a fixed experiment has arrived. An alternative is gone and the problem is a different one. Conditioning requires a model of how the probabilities were generated, and linear renormalization supplies one.

Which model it supplies is known. Within the class of independent additive random-utility models, Gumbel errors generate Equation (1) [35, 33], and Yellott [39] showed the noise law is identified once menus of three or more are included, though not from pairwise comparisons alone. So the constant-ratio rule is one point of a family.

Nothing requires standing on exactly one point of it. Anyone with restricted-menu observations to fit can estimate an interpolating model, which is no worse in sample and may or may not be better out of it. The question here is narrower. Linear normalization is applied almost unconsciously, in survey analysis, in pricing and inside any model that masks a softmax, and the question is whether it is typically better or worse than Gaussian renormalization when a default has to be chosen and there is nothing to fit.

Equation (2) is a one-dimensional integral for any number of alternatives, since conditioning on $X_i = x$ makes the other events independent, and it is inverted numerically by the lattice method of [2], which recovers the whole location vector in one pass over a shared grid.

Any noise law fixed in advance up to location gives a map with nothing left to choose, so the class supplies many zero-parameter candidates and not two. We compare two of them because they are the conventional ones. Gumbel gives linear renormalization, which is the reflex and the architecture of a masked softmax. Unit-variance Gaussian gives Case V, which is the classical model of a latent contest and the one Thurstone proposed.

The asymmetry between the two is worth naming. Case V is a model somebody proposed and defended. Linear renormalization is the Gumbel-noise member of the same class, and it is not adopted because anyone argues that choice errors are double exponential. It is adopted because dividing by a smaller total is the obvious thing to do. Nothing in a masked softmax expresses a belief about extreme-value noise. The incumbent therefore holds its position by default rather than by argument, and the question below is not whether a challenger can overturn a defended position but whether the default was ever the right place to stand.

Interpolating between them, or choosing the standardized shape of some third law, requires a number that the shares themselves do not supply. Overall scale does not: it is a gauge, absorbed exactly by the fitted locations. Both are calibrated to the same $|S| - 1$ free shares and neither sees any observation from a smaller menu. Both are *representative-agent* maps, treating the population as one chooser with one set of worths or locations and transporting its shares from S to T. A fitted Plackett–Luce likelihood [36] uses the whole ranking and a fitted Bradley–Terry model [10] uses the pairwise outcomes directly; neither is estimated here.

The direction of the disagreement is a theorem, and it holds for noise laws other than the Gaussian. Let f and F be the density and distribution of the common noise, independent and identically distributed across alternatives up to location, with $f > 0$ and $0 < F(x) < 1$ at every finite x, so that the reverse hazard $r = f/F$ is defined everywhere and $\log r$ is differentiable. The highest draw wins, as above.

Proposition 1 (Removing alternatives contracts the odds)**.** *Suppose* $\log r$ *is concave. Let* $T \subset S$ *and* $i, j \in T$ *with* $a_i > a_j$*. Then*

$$\frac{P(i \mid S)}{P(j \mid S)} \;\geq\; \frac{P(i \mid T)}{P(j \mid T)} \;\geq\; 1,$$

with the first inequality strict when $\log r$ *is strictly concave and* H *is non-constant on a set of positive measure, which holds whenever* $S \setminus T$ *is non-empty.*

Proof. Write $A(x) = f(x - a_i)F(x - a_j)$ and $B(x) = f(x - a_j)F(x - a_i)$ for the two items in question, $C(x) = \prod_{k \in T \setminus \{i,j\}} F(x - a_k)$ for the other survivors, and $H(x) = \prod_{k \in S \setminus T} F(x - a_k)$ for the items removed. Then $P(i \mid T) \propto \int AC$ and $P(i \mid S) \propto \int ACH$, and likewise for j with B in place of A. Now

$$\frac{A(x)}{B(x)} = \frac{r(x - a_i)}{r(x - a_j)},$$

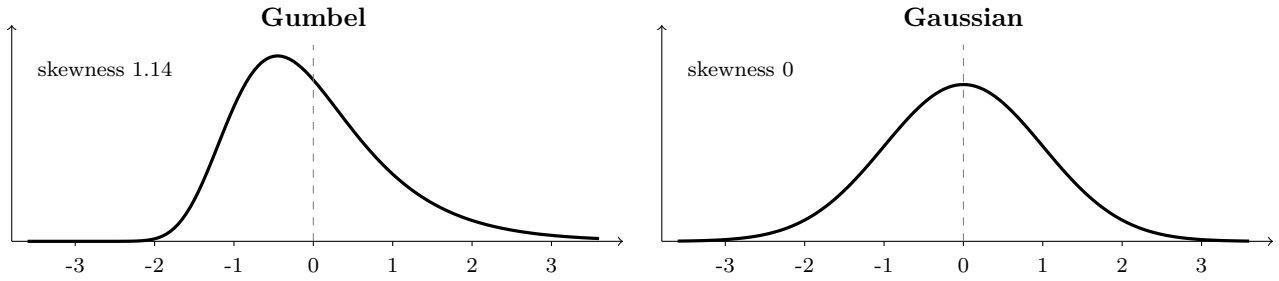

Figure 2: The two noise laws, each standardized to mean zero and variance one, so that only shape distinguishes them. Give every alternative a location, add a draw from the left-hand density to each, and take the highest: the winning probabilities are Equation (1), and restricting the field renormalizes the shares proportionally. Do the same with the right-hand density and the shares contract instead. Renormalization is therefore a race too, run with a noise law that is asymmetric, bounded below in effect and long-tailed above. We withhold our opinion as to which is the more *a priori* reasonable.

which is increasing in x when $\log r$ is concave, since $a_i > a_j$ places $x - a_i$ below $x - a_j$ and $(\log r)'$ is decreasing. Under the probability measure proportional to $B(x)C(x)\,dx$ the functions A/B and H are both increasing, hence positively correlated, so

$$\frac{\int ACH}{\int BCH} \geq \frac{\int AC}{\int BC},$$

which is the first inequality, strict under strict concavity when $S \setminus T$ is non-empty. For the second, raising a_i stochastically increases X_i and so raises $P(i \mid T)$ while lowering $P(j \mid T)$, and the two are equal at $a_i = a_j$ by exchangeability; hence $a_i > a_j$ gives $P(i \mid T) \geq P(j \mid T)$. □

The proposition says that removing alternatives moves the ratio toward one.

Two laws sit at the ends of that condition. The Gaussian satisfies it strictly, since $\frac{d^2}{dx^2} \log r(x) = -\operatorname{Var}(Z \mid Z < x) < 0$. The Gumbel has $r(x) = s^{-1}e^{-x/s}$, so $\log r$ is affine, the inequality becomes an equality, and renormalization is recovered exactly, which is the Holman–Marley construction [35, 33]. The axiom is a boundary of the class, and it is where machine learning sits: a softmax over scores z_i is Equation (1) with $u(i) = \exp(z_i)$, so masking a softmax and rescaling the survivors is proportional renormalization.

An objection is worth answering here, because it is the first one a reader from machine learning will raise. A deep network is not committed to the axiom. Its logits are the output of an arbitrarily flexible function, and it can represent any distribution over its full vocabulary, including distributions that no independent additive random-utility model generates. All of that is true, and none of it touches the case at hand. The constraint applies to a *fixed* logit vector at inference time. Depth governs the map from context to logits; it says nothing about how the distribution should change when the support is cut and the logits are not recomputed. Masking and rescaling a fixed vector leaves every surviving ratio exactly as it was, which is Equation (1) on the restricted set, however that vector was produced and however deep the network that produced it.

Escaping the axiom therefore requires recomputing the scores with the permitted set as part of the input. A model trained that way can express any dependence on the menu, and its masked

distribution is then not a transport of its unmasked one, so the question this paper asks does not arise for it. The question arises exactly where a score vector is computed once and then restricted, which is what constrained decoding, vocabulary masking, restricted action spaces and post-hoc class removal all do.

The proposition gives the direction of the effect and not its size. Size depends on the shares as well as on the noise law, and recalibrating several laws to one share vector and then to another can reverse which of them contracts more, so tail thickness alone does not order the laws by how much they contract.

4.2 Gaussian taste heterogeneity as a mechanism

Nothing so far explains why the Gaussian member of the class should be the one that fits. A mechanism is available, and it is elementary.

Let every individual obey the axiom, and let their tastes differ. Individual θ assigns systematic utility $V_i(\theta)$ to alternative i and chooses by Luce at scale $\beta > 0$,

$$P(i \mid S, \theta) = \frac{\exp\{V_i(\theta)/\beta\}}{\sum_{j \in S} \exp\{V_j(\theta)/\beta\}}, \tag{4}$$

which by Yellott's theorem is the same as choosing the largest $V_i(\theta) + \beta G_i(\theta)$ for independent standard Gumbel $G_i(\theta)$.

Proposition 2 (Gaussian taste heterogeneity)**.** *Suppose $V_i(\theta) = \mu_i + \sigma Z_i(\theta)$ with $Z(\theta) \sim N(0, I_K)$ independently across individuals. Then the aggregate choice probabilities are exactly the winning probabilities of a contest with utilities*

$$\mu_i + \sigma Z_i + \beta G_i, \tag{5}$$

independent across alternatives. If $\mu_i = \sigma a_i$ and $\sigma/\beta \to \infty$, they converge to those of Case V with locations a.

Proof. Aggregating over individuals integrates over the taste draw Z, leaving the per-alternative utility (5), independent across i because both components are. Dividing every utility by σ preserves the arg max and gives $a_i + Z_i + (\beta/\sigma)G_i$, whose last term vanishes in probability. Ties have probability zero under the Gaussian limit, so the winning probabilities converge. □

This is a mixed logit: Gaussian heterogeneity between individuals, Gumbel noise within them. Such models are standard, and McFadden and Train [7] establish how general they are. Nothing here claims otherwise. What the proposition supplies is the connection between the high-heterogeneity limit of that familiar object and the particular restriction map the paper tests.

The Gaussian assumption on taste is where a limit theorem belongs.

Corollary 1 (Many small taste components)**.** *Suppose $V_i(\theta) = \sqrt{m}\, a_i + \sum_{\ell=1}^{m} \xi_{\ell i}(\theta)$, where the vectors $\xi_\ell(\theta)$ are independent and identically distributed across ℓ with mean zero, covariance I_K and finite second moments, and the individual chooses by (4) at fixed β. Then the aggregate choice probabilities converge as $m \to \infty$ to those of Case V with locations a.*

Proof. Dividing the utilities by $\sqrt{m}$ gives $a_i + m^{-1/2} \sum_\ell \xi_{\ell i} + (\beta/\sqrt{m})G_i$. The multivariate central limit theorem sends the middle term to $N(0, I_K)$ and the last to zero, and the arg max probabilities are continuous at a limit with no ties. □

A correction to this limit must be a correction to shape, not to scale. For any location-scale family $X_i = a_i + s\varepsilon_i$ the winning probabilities satisfy $P_i^S(a; s) = P_i^S(a/s; 1)$, so if $b(p)$ is the unit-scale location vector calibrated to the full-menu shares p, then $s\,b(p)$ is the scale-s calibration and $P_i^T(s\,b(p); s) = P_i^T(b(p); 1)$ for every restricted menu T. A common noise scale is a gauge. It is absorbed exactly by the fitted locations, not approximately, and restricted-menu observations cannot identify it either unless the locations are anchored by cardinal information from outside the shares.

The same holds for the composite noise $W_\varepsilon = Z + \varepsilon G$ at finite σ/β, with $\varepsilon = \beta/\sigma$. Subtract the common Gumbel mean, which cancels in the arg max, and divide by $s_\varepsilon = (1 + \varepsilon^2\pi^2/6)^{1/2}$, which is a gauge by the previous paragraph. The standardized law then has skewness of order ε^3 and excess kurtosis of order ε^4, with no ε^2 term surviving. Under smooth dependence of the winning probabilities on the noise law, and a locally nonsingular Jacobian of the full-menu share map in the $K - 1$ location contrasts, the implicit function theorem gives $a_\varepsilon(p) = a_0(p) + O(\varepsilon^3)$, so the first generic departure of the recalibrated restriction map from Case V is cubic in ε. Symmetric configurations can cancel the cubic term and leave a higher order. Any parameter that interpolates between the two maps has to be a shape or relative-mixture parameter. It cannot be a common noise scale.

The covariance assumption is doing real work and should not be waved through. If the taste components carry covariance Σ instead of I_K, the same argument delivers a Gaussian race with covariance Σ, which is correlated multinomial probit rather than Case V. Case V additionally requires the covariance to be isotropic on the $(K-1)$-dimensional utility-contrast subspace. In the usual reference-difference coordinates $D_i = \varepsilon_i - \varepsilon_K$ that is not a diagonal matrix: independent equal-variance errors give $\mathrm{Cov}(D) = \sigma^2(I_{K-1} + \mathbf{1}\mathbf{1}^\top)$, and it is in an orthonormal contrast basis Q that $Q^\top \mathrm{Cov}(\varepsilon)Q = \sigma^2 I_{K-1}$. So the limit theorem motivates Gaussian random utility in general, and independence across alternatives is a further restriction, adopted here because it is the parameter-free member of the class.

σ	Top share	Observed λ	Case V λ	Ratio
0.5	0.255	0.052	0.225	0.23
1	0.300	0.128	0.227	0.57
2	0.343	0.197	0.229	0.86
3	0.359	0.216	0.229	0.94
4	0.365	0.228	0.230	0.99
6	0.372	0.225	0.230	0.98
8	0.373	0.230	0.231	1.00
16	0.375	0.233	0.231	1.01

Table 2: A population of Luce choosers whose log-worths carry Gaussian taste variation of scale σ against a Gumbel choice-noise scale of 1, with locations scaled as σa so that the shares do not degenerate. The contraction slope λ is defined on pairs by $\delta_{ij} = \mathrm{logit}\, q_{ij} - \log(p_i/p_j)$ for i the higher-share alternative, with q_{ij} the observed share preferring i from $\{i, j\}$, fitted through the origin as $\delta = -\lambda \log(p_i/p_j)$; $\lambda = 0$ is exact renormalization. Observed λ is the population's contraction; Case V λ is what the independent unit-variance Gaussian race implies for that population's own shares. The top share is reported to show the limit is not the near-indifference corner. Parameters: $K = 5$, $A = (0.55, 0.25, 0, -0.25, -0.55)$ with $\mu = \sigma A$, Gumbel scale 1, and 400,000 direct Monte Carlo draws per row from `numpy.random.default_rng(0)`; reproduced by `python heterogeneity_limit.py`.

No individual need be non-Lucean for the Gaussian transport to be the right one in aggregate. A population of Luce choosers whose tastes vary is a contest with composite noise, and as taste variation comes to outweigh within-person choice noise the composite approaches Gaussian and the population contraction approaches the Case V prediction. That mechanism also says where the fit should be worst: in populations whose members are nearly alike.

Independent evidence for the same phenomenon exists in memory research. Robinson, DeStefano, Vul and Brady [8] put softmax against Gaussian signal detection across m-alternative recognition tasks, and report that "the d' parameter of the Gaussian signal detection model was more stable across m-afc conditions than the β parameter of the softmax model". Their primary analysis holds each model's parameter fixed across all m conditions and compares log likelihood, and the Gaussian wins in both experiments, $t(29) = 4.26$ and $t(29) = 4.42$, $p < .001$, with thirty participants each. That is the contraction studied here, seen through the parameter that has to move to absorb it.

Their design is nearer to this paper's than anything else we have found, and the distance that remains is this. No response set is restricted over shared alternatives: m varies by adding fresh items, not by withdrawing named ones. Linear renormalization is never the competitor, since the comparison is softmax against Gaussian with both fitted. And the parameters are estimated on the data that scores them, so the cross-condition constraint does the work that holding out a menu does here. What their result establishes is that one parameter has to move with m and the other does not, which is the signature of contraction rather than a measurement of it.

4.3 What the shares cannot identify

Correlated multinomial probit can represent near-substitution, by giving two similar alternatives a positive covariance. It does not follow that it represents any near-substitute behaviour exactly: menu-dependent attention and the dimensional retuning of Section 5.8 are not captured by a static covariance. And it is not available as a default. One full-menu share vector supplies $K - 1$ free numbers, and those are already exhausted by the $K - 1$ location contrasts. After choosing a reference alternative and normalizing scale, the covariance matrix of the $K-1$ utility differences carries a further $K(K - 1)/2 - 1$ parameters. A single $(K - 1)$-dimensional share vector cannot jointly identify $K - 1$ location contrasts and an unrestricted covariance shape, so for every $K \geq 3$ the covariance is unidentified from full-menu shares alone. A correlated model is available to anyone with restricted-menu data to fit it on. That is a different exercise from the one conducted here.

5 Empirical

5.1 Protocol

Both maps are calibrated on full-menu shares alone, then scored by log loss on restricted menus neither has seen. The input is a probability vector and nothing else, which is what makes the two maps comparable and what the exercise assumes.

Both maps hold the latent quantity fixed and change only the menu, and Section 5.8 collects the cases where that fails. Two further assumptions are less visible. The noise is independent and identically shaped across alternatives, which near-substitutes violate. And the population is treated as one chooser with one latent vector, though Section 4.2 shows heterogeneous Luce choosers reproducing aggregate contraction with no individual departing from the axiom. Three

things make that possible now. Trial-level archives are open, so restricted menus can be scored on the observations that produced them rather than on printed marginals. Inverting Equation (2) from shares alone is a single pass over a shared grid, so the Gaussian map costs nothing to apply. And a generative null is available. Contraction of a noisily estimated share vector lowers log loss whether or not the axiom fails, so every raw gain below is accompanied by the gain a population obeying the axiom exactly would produce on the same menus, its centred excess, and a Monte Carlo tail. The excess is a diagnostic centred on that particular fitted null, not an estimate of a model-free structural effect: under a non-Luce population there is no reason for the null's bias to match the alternative's. Section 5.7 details the construction.

5.2 Restrictions that were observed

Ten collections record choices from a smaller menu directly, and they are reported first because nothing about the restriction is inferred. The first three are controlled: the same subjects, or the same stimuli, before and after a menu change. The remainder are ecological, and Section 5.2 says where that weakens them.

5.2.1 Consumer goods, all subsets of five

In the ranking-and-menu benchmark of Tables 3 and 6, twelve collections are complete rankings from which subset choice is read off, and one observes menu choice directly. That one comes first.

Costa-Gomes, Cueva, Gerasimou and Tejiščák [1] ran two experiments in which subjects chose separately from every non-singleton subset of five consumer goods, twenty-six menus; the first also included the five singletons. The archive names the goods by two-letter codes and we describe them no more precisely. Restriction is observed rather than imputed, which makes this the cleanest test in our collection, and the two experiments are separate studies with disjoint subjects, so we report them separately rather than pooled.

Subjects are split into five folds. On the training subjects, shares come only from the full five-item menu; both maps are calibrated to those shares and nothing else; each then predicts every menu of size two, three and four; and we score the choices actually made by held-out subjects. Intervals resample subjects and repeat the whole procedure, calibration included. Subjects who may defer are reported alongside the forced-choice subgroup, for whom no deferral coding enters. Of the 373 subjects, 364 record at least one usable choice from a non-singleton menu and enter the analysis.

The race is ahead in both experiments and in both treatments, and by more where deferral cannot interfere. The two experiments do not agree on strength. Experiment 1 clears its null comfortably, with Monte Carlo tail frequencies of 0.010 and 0.005; experiment 2, at fifty-eight per cent of the sample, does not, at 0.199 and 0.129. The direction replicates and the magnitude is not established by these data alone.

5.2.2 Colours and symbols, with the menu revealed after the stimulus

The consumer experiment observes restriction but announces it: the subject sees the menu before deciding. Two perceptual experiments remove that confound, and one of them removes it by design.

Yeon and Rahnev [26] show a display whose dominant colour, among four, has to be named. In one condition the full menu is available. In a second the observer sees the identical display and

	n	Choices	Renorm.	Race	Gain	Excess
Experiment 1, all	230	5,134	0.9438	0.9298	+0.0140	+0.0158
Experiment 1, forced	76	1,900	0.9639	0.9197	+0.0442	+0.0466
Experiment 2, all	134	3,105	0.9162	0.9103	+0.0059	+0.0056
Experiment 2, forced	61	1,525	0.8939	0.8799	+0.0140	+0.0131
Pooled	364	8,239	0.9263	0.9164	+0.0100	+0.0123

Table 3: Held-out log loss per choice on menus of size two to four, both maps calibrated only to training-fold full-menu shares under a common add-α convention, $\alpha = 1/2$. Excess is the gain net of the median gain under a population that renormalizes exactly, as in Section 5.7. Subject-bootstrap intervals for the gains are $[+0.0072, +0.0335]$, $[+0.0171, +0.0867]$, $[+0.0009, +0.0306]$, $[+0.0034, +0.0560]$ and $[+0.0048, +0.0246]$ respectively.

is told only *after* it has gone which two named alternatives the answer lies between. Conditions were blocked and trained, so the observer knew a two-alternative response was coming; what could not be anticipated while the percept formed is the identity of the surviving pair. The design therefore prevents pair-specific advance attention rather than making the restriction itself a surprise. Their Experiment 2 repeats this with six symbols. Calibration uses that observer's full-menu row for that dominant item alone, and the two-alternative cells on the same observer and item are then predicted and scored.

Arm	Cells	Renorm.	Race	Gain	Excess
Four colours, menu revealed after	384	0.4942	0.4795	+0.0147	+0.0157
Six symbols, menu revealed after	300	0.6276	0.5999	+0.0278	+0.0314
Same pairs, announced in advance	384	0.3967	0.3928	+0.0038	+0.0048

Table 4: Mean nats per cell, a cell being one observer, one dominant item and one surviving alternative. Subject-bootstrap intervals for the gains are $[+0.0118, +0.0179]$, $[+0.0234, +0.0324]$ and $[+0.0006, +0.0072]$. The null here resamples the calibration row as well as the pair counts, so it charges the race for calibration noise as well as for sampling noise in what it predicts; all three tails are at the 0.005 floor.

The six-symbol gain is the largest of the two arms here. The same pairs were also run *announced in advance*, so the observer can aim attention at them. That arm was not analysed by the original authors and is recovered here from the trial files. There the advantage nearly disappears. Accuracy rises from 0.780 to 0.850, linear renormalization becomes almost exactly right, and the race begins to over-contract. So the race's advantage is not a property of these pairs, these colours or these observers. It appears when the surviving pair is selected only after the percept has formed, so that attention cannot be tailored to that pair, and it does not appear when the pair is named in advance.

In both genuine restriction arms *both* maps over-predict the favourite: linear renormalization by 0.069 and the race by 0.055 at four alternatives, and by 0.096 and 0.064 at six. Contraction has the right sign and removes a fifth to a third of the error. Neither of the two maps is calibrated here, and a fitted shape or mixture parameter has room to improve on both. That over-prediction is Yeon and Rahnev's own finding, reached with a fitted model, so their result is independent evidence that the correction runs toward contraction while bounding what either

default can claim.

Utochkin, Azarov and Grigorev [27] supply the same structure in memory. An observer sees a studied photograph among three foils, or the same target against one named foil, with the same images in both arms. The race wins by +0.0037 per cell over 360 nested cells, interval [+0.0029, +0.0047], the tightest in the paper. The contraction is the same size against a foil from the target's own category as against a foil from another, which matters in Section 5.8: category membership is not what breaks independence.

5.2.3 Gambles, ballots, news slates and Wikipedia links

Restriction is observed in four further populations, which differ in what does the removing.

A fifth candidate was tried and discarded. Grocery stockout data records hourly sales by product, and a shopper may buy two items from one category, so the alternatives are not mutually exclusive and the object is a basket rather than a choice. Recovering choices would need transaction-level baskets, which the release does not contain.

Aguiar, Boccardi, Kashaev and Kim [9] observe choices from all thirty-one non-empty subsets of five lotteries with a default always available, so menus run from two to six alternatives, in a cross-section giving at most two disjoint menus per person. It is adversarial on the authors' own analysis, since they report that random utility cannot explain the population behaviour while a logit attention model survives.

The experiment was run three times on disjoint subject pools, differing only in how much arithmetic is needed to value an alternative: one operation, three, or five. The excess over the null rises with that requirement, from −0.0012 to +0.0045 to +0.0081, and the full-menu shares flatten alongside it, the largest falling from 0.357 to 0.257 against a six-alternative maximum entropy of 1.792.

We read this as the thesis working rather than as instability. When the arithmetic is one step the task has a computable answer and is closer to an examination question than to a preference judgement; subjects converge, the shares concentrate, and there is little preference left for any restriction map to describe. When it takes five steps nobody can compute the answer, so people fall back on what they actually prefer and the choice becomes psychological. A model of preference under noise should win in the second case and not the first. Winning in both, including where the answer is objectively determined, would be the result to distrust.

US presidential ballots supply restriction by statute. The alternatives are the same named candidates in every state and which of them appears varies with petition thresholds and filing deadlines, so the states are nested menus over identical alternatives. States are not interchangeable populations, so a state may only predict another whose two-party log odds are within 0.5.

News slates supply the largest sample and the only menus that are recorded rather than reconstructed. The Microsoft News dataset logs, for each impression, the exact set of articles displayed and which one was clicked [38]. Slates recur across users and one slate is frequently a subset of another, so click shares on the larger slate calibrate both maps and the choices made on the nested smaller one score them. Of 153,727 impressions, 846 slates recur ten or more times and yield 6,135 nested pairs, of which the first 6,000 are scored. Because those pairs are drawn from a few hundred slates and overlap heavily, the interval below resamples slates rather than pairs.

Wikipedia links supply editorial removal, in the weakest form of observation used here. The clickstream release is not a record of impressions: it is a monthly count of (referrer, resource)

transitions extracted from request logs, so what it supplies is the aggregate distribution of destinations conditional on a transition having been made from a given article, and not a table with one row per reader carrying a menu and one exclusive response. Destinations are the same articles from month to month and editors change the set, so adjacent dumps give the earlier month's conditional shares and the later month's. The dump censors pairs below ten clicks, so a vanished destination may have been removed or may have fallen below threshold; it records no link position, so layout is uncontrolled; and readership and page content move between months along with the link set. This row is a transport comparison on conditional transition shares rather than an observed withdrawal.

Population	Units	Gain	Null median	Excess	MC tail
Lotteries, high cost	3,928	+0.0101	+0.0020	+0.0081	.035
Lotteries, medium cost	3,928	+0.0068	+0.0023	+0.0045	.124
Lotteries, low cost	3,928	+0.0002	+0.0014	−0.0012	.667
News slates	6,000 pairs	+0.0496	−0.0088	+0.0584	.016
US state ballots 2024	263 pairs	+0.0002	−0.0000	+0.0002	.005
Wikipedia links	3,544 pages	−0.0001	−0.0002	+0.0001	.005

Table 5: Held-out loss differences on populations where restriction is observed rather than read off a ranking. Excess is the gain net of what a renormalizing population produces on the same menus. The lottery frames are disjoint subject pools; ballot units are matched state pairs; Wikipedia units are pages whose destination set shrank; One row, Wikipedia, has a negative raw gain and a positive excess, which means Gaussian renormalization loses in absolute terms while still departing from the axiom in the direction of contraction.

The news result is the largest here by an order of magnitude, on the only menus in the paper that were recorded at the moment of choice. Its weakness is assignment: the slates are chosen by a recommender, so a slate's composition is correlated with what the system expects readers to want, and a smaller slate may reach a different readership than the larger one. Both maps inherit that identically and the null runs on the same slates. That does not make the comparison clean. The null holds the worths fixed and so does not simulate a shift in readership, position or context, and two maps facing the same confound need not respond to it equally well. This row and the two below it are transport exercises on data that was not generated by withdrawing an alternative from a fixed population, and they are weaker evidence about withdrawal than the controlled restrictions of Section 5.2 for that reason.

The lottery gradient runs with arithmetic cost, and the mechanism of Section 4.2 predicts its direction, since computation error is per-alternative and roughly Gaussian, so loading more of it onto each alternative raises σ against β. Tightening the filter that identifies a genuine removal from a five per cent share to thirty-five per cent raises the excess from +0.0001 to +0.0005 across 3,544, 817 and 121 pages. That is consistent with attenuation by measurement, and also with the threshold selecting a different population of pages, which these data cannot separate. Wikipedia and the two General Social Survey collections put the truth between the two maps, contracting when an alternative goes but by less than Case V says. They are the exception. Fitting the contraction slope λ of Table 2 to each collection's observed pairwise choices puts the observed slope *above* the race's in eight of eleven, so in most collections choice contracts by more than Case V predicts and Gaussian renormalization errs by over-predicting the favourite. That is the same direction as the after-stimulus arms of Section 5.2.2, where both maps over-predict,

and it says the Gaussian point of the family is not far enough along it for these populations.

5.3 Restrictions induced from rankings

Twelve archival collections have human respondents and rankings, or ratings fine enough to yield them. Two rank-order card tasks from the General Social Survey ask what a child should learn and what matters in a job. Three independent samples of Jester users rate the same ten jokes. Sushi rankings come from a Japanese web survey, film rankings are induced from Netflix ratings, and dots and puzzles are perceptual discrimination tasks. The rest are rankings of seven sports by participation preference, ten occupations by perceived prestige, and four political goals from a five-nation survey. The thirteen collections comprise 147,719 responses, of which the menu experiment contributes 373 subjects; the twelve ranking collections contribute 147,346. These are not thirteen independent experiments, since three are Jester files, two are General Social Survey items, and dots and puzzles are companion tasks.

The protocol matches Section 5.2.1 with respondents in place of subjects, and the outcome for a subset is the respondent's highest-ranked member of it.

Dataset	n	K	Subsets	Renorm.	Race	Gain
Occupational prestige	143	10	1,013	1.0173	0.9783	$+0.0390$ $[+0.0227, +0.0555]$
Sushi	5,000	10	1,013	1.3724	1.3613	$+0.0111$ $[+0.0094, +0.0128]$
Political goals	2,262	4	11	0.8835	0.8766	$+0.0069$ $[+0.0059, +0.0078]$
GSS socialization	5,000	5	26	0.7984	0.7929	$+0.0055$ $[+0.0041, +0.0070]$
Jester file 3	5,000	10	1,013	1.5189	1.5165	$+0.0024$ $[+0.0015, +0.0032]$
Jester file 1	5,000	10	1,013	1.5287	1.5267	$+0.0020$ $[+0.0012, +0.0027]$
Jester file 2	5,000	10	1,013	1.5247	1.5230	$+0.0018$ $[+0.0010, +0.0025]$
GSS job values	5,000	5	26	0.8593	0.8576	$+0.0017$ $[+0.0005, +0.0030]$
Dots	3,183	4	11	0.8285	0.8270	$+0.0015$ $[+0.0005, +0.0025]$
Netflix	4,379	3	4	0.7932	0.7922	$+0.0009$ $[+0.0007, +0.0012]$
Puzzles	3,180	4	11	0.8180	0.8173	$+0.0007$ $[-0.0003, +0.0017]$
Sports participation	130	7	120	1.2579	1.2528	$+0.0050$ $[+0.0017, +0.0087]$

Table 6: Held-out log loss per prediction over every subset of size two or more. Datasets are capped at five thousand respondents for runtime. The race is lower in twelve of twelve and the interval excludes zero in eleven. Occupational prestige is scorable here only because the shared smoothing removes a zero training cell. Sports participation sits below the rule: its gain is real but Section 5.7 shows it carries no evidence about the restriction map. Intervals resample respondent losses with the fitted training models held fixed, so unlike Table 3 they omit calibration uncertainty and are too narrow, most seriously in the two smallest rows.

5.4 The constant-ratio-rule datasets rescored

Four subjects of Townsend and Landon [23] identified briefly displayed block capitals, in blocks with the master response set $\{A, E, F, H, X\}$ and blocks with $\{A, E, F, H\}$, $\{A, E, X\}$ and $\{F, H, X\}$, counterbalanced, 240 trials per row. Responses were spoken letter names, so the labels are identical across menus by construction. They published the full confusion matrices per subject, which is why the dataset is still usable forty-four years later.

It is the cleanest calibration in the paper for a reason none of the others share. Calibration row and target row come from the *same subject* in different blocks, so there is no population-heterogeneity confound of the kind every ranking collection carries, and the aggregation argument of Section 4.2 is not available as an explanation of any gain found here.

Two rows are dropped because the published table prints cells that cannot be reconciled with their own totals. Over the remaining 38 restricted rows and 9,120 held-out trials, linear renormalization scores 0.9454 and the race 0.9412, a gain of +0.0042 with a row bootstrap of $[+0.0022, +0.0064]$, against a fitted-Luce null median of −0.0011, so the excess is +0.0053 at $p = 0.005$. The subsets give +0.0006 $[-0.0002, +0.0015]$ for $\{A, E, F, H\}$, +0.0040 $[-0.0003, +0.0079]$ for $\{A, E, X\}$ and +0.0093 $[+0.0051, +0.0141]$ for $\{F, H, X\}$. That ordering is mostly the removal count: the first subset withdraws one letter and the other two withdraw two, and Table 7 shows the gain growing with how much is removed and vanishing when nothing is. The unconfounded comparison is between the two subsets that each remove a similar pair and retain the other, +0.0040 against +0.0093. Those two designs are structurally symmetric, each removing a similar pair and retaining the other, so the difference between them is not attributable to similarity.

What the row-level output does show is the limit of a parameter-free correction. For subject one, stimulus A, subset $\{A, E, X\}$, the observed, renormalized and raced shares are 0.621/0.672/0.635 on A, 0.217/0.129/0.149 on E and 0.163/0.199/0.217 on X. The race corrects the favourite downward and the near-substitute upward, both in the right direction and the second far too little, and moves the dissimilar letter the wrong way. It wins because the dominant terms improve. Townsend and Landon diagnosed their residual as mass concentrating onto the surviving near-substitute instead of spreading, and Case V has no covariance to represent that, so the residual survives. A parameter-free Gaussian recovers part of the discrepancy and does not remove it. Lee [34] proposed exactly this comparison analytically in 1968, observing that detection theory and the constant-ratio rule diverge most for unidimensional stimuli and that the gap could serve in diagnosis; what is added here is the out-of-sample follow-through.

Wills et al. (2000)

Wills et al. deposited their Experiment 2, in which twelve participants chose among three categories and twelve others had one category disallowed, over the same stimuli. Over 39 cells and 1,560 held-out two-choice trials, linear renormalization scores 0.6839 and the race 0.6540, a gain of +0.0299 against a null median of −0.0015, so the excess is +0.0314 at $p = 0.002$. Graded and dummy stimuli agree, +0.0287 against +0.0326.

The restriction is between groups, with four participants behind each disallowed category, and the gain sits in two of the three: −0.0001, +0.0413 and +0.0485. The participant bootstrap is wide and right-skewed, $[+0.0175, +0.0717]$.

The withdrawn response occurs zero times in the two-choice condition and 510 times in the three-choice condition, so it is a real competitor on the master menu, and cells are balanced at forty trials.

5.5 How the difference is distributed

5.5.1 By size of the surviving menu

Splitting the held-out predictions of Section 5.3 by the size of the surviving menu shows the gain concentrated where most of the menu has been withdrawn.

Dataset	$\lvert T\rvert = 2$	3	4	5	6	7	8	9	10
Sushi	+.0412	+.0263	+.0150	+.0081	+.0040	+.0017	+.0005	+.0000	−.0000
Jester file 3	+.0064	+.0050	+.0033	+.0020	+.0012	+.0006	+.0003	+.0001	−.0000
Jester file 1	+.0053	+.0041	+.0026	+.0016	+.0010	+.0005	+.0003	+.0001	+.0000
GSS socialization	+.0130	+.0011	+.0004	−.0000					
Political goals	+.0079	+.0071	+.0000						
Dots	+.0025	+.0003	−.0000						
Netflix	+.0013	+.0000							

Table 7: Held-out gain by the size of the surviving menu, for a representative seven of the twelve ranking collections. In most collections the pairwise gain exceeds the all-subsets aggregate, by factors from 1.2 to 3.7, though occupational prestige reverses it at 0.73. The last entry in each row is zero to the precision of the pipeline, because nothing has been removed and the two maps coincide exactly in theory. The aggregate quoted elsewhere is therefore lower than the pairwise figure. Occupational prestige and sports participation are non-monotone in the size of the surviving menu and are omitted here.

5.5.2 By rank of the item chosen

Averaged over outcomes the difference between the two maps is small. Scoring each held-out prediction by the rank the chosen item held in the full-menu shares, pooled over the four ten-item collections, shows what the average is made of.

Rank of chosen item	1	2	3	4	5	6	7	8	9	10
Share of outcomes	.17	.13	.13	.10	.10	.10	.08	.07	.07	.04
Mean gain	−.060	−.045	−.033	−.003	+.010	+.020	+.054	+.067	+.116	+.242

Table 8: Held-out gain by the full-menu rank of the item actually chosen, pooled over Sushi and the three Jester files. Monotone in rank, crossing zero between fourth and fifth. The contributions sum to +0.0043.

Gaussian renormalization pays 0.060 nats whenever the favourite wins, which happens 17 per cent of the time, and collects 0.242 when the tenth-ranked item wins, which happens 4 per cent of the time. That is what a contraction does: it moves probability off the favourite and onto the tail, buying insurance against upsets and paying a premium on every favourite win. Three earlier observations follow. The aggregate gain is small because large rare wins net against small frequent losses. The ballot result is negligible because third-party candidates are the tail and there is almost no tail mass to collect. And the effect is largest on news slates, where readers spread widely over the displayed set.

5.6 An aside on sequential heuristics

The next comparison is not the restriction question, and is separated for that reason. Removing the observed winner and re-running is conditioning on the outcome of a fixed contest, not withdrawing an alternative from it, and learning who won is informative about the latent draw. Neither rule below is the exact ordering law of the model it is named for. What follows is therefore a comparison of two practical heuristics used in place of those laws, and it is evidence about the heuristics rather than about either transport map.

Collection	Observed	Renormalization	Race
Political goals	0.123	0.318	0.309
GSS job values	0.186	0.303	0.294
GSS socialization	0.191	0.302	0.294
Sushi	0.206	0.250	0.245
Jester file 1	0.138	0.160	0.158

Table 9: Probability that the full-field favourite finishes second. Both columns are sequential rules: remove the winner, then apply the map to the survivors. That is Harville's construction in the first column and its re-run-the-contest analogue in the second, and neither is the exact ordering law of its model, so this compares two practical rules rather than two models. Renormalization overstates the probability in every collection, by 0.022 to 0.195, and Gaussian renormalization overstates it by a little less. The mirror holds for the last-ranked item, whose chance of finishing second both rules understate everywhere.

This is the pattern reported for the Harville formula in racing place markets, favourites over-priced and outsiders under-priced, appearing in survey rankings of sushi, jokes, child-rearing values and political goals. It is the most useful result here for a practitioner, because it is a large, replicable, signed error in the default rather than a small average edge. Gaussian renormalization removes 4.9 to 12.4 per cent of it, so both rules share a bias much larger than the distance between them. What produces the remainder is not identified; the obvious candidate is that a single respondent's ranking is internally consistent, so adjacent places are positively correlated in a way no independent contest reproduces.

5.7 The shrinkage null

A positive gain has two possible sources and log loss cannot tell them apart. The race may be closer to how the population chooses. Or the race may be wrong about that and win anyway, because contracting a noisily estimated share vector is shrinkage, and shrinkage lowers log loss when the input is over-confident. The second mechanism needs no departure from the axiom, so properness of the score licenses nothing here: a proper score says the true probability vector is optimal, not that a plug-in estimate of it beats a biased transformation of the same estimate.

The remedy is to build a population where the axiom holds by construction and run the identical procedure on it. Taking the observed full-menu shares as worths, choices or complete rankings are generated from that exact process, and everything is repeated. Rankings are drawn by sorting $\log u_i + G_i$ for independent standard Gumbel G_i, which is the Plackett–Luce ranking law; for $K \geq 4$ this is a particular joint law consistent with those subset marginals rather than the only one, so the null is specifically a Plackett–Luce ranking null.

The shrinkage mechanism is real and it accounts for the entire gain in the smallest dataset. Sports participation, with 130 respondents, sits exactly on its null and is excluded from every claim here. Everywhere else the mechanism runs against the race, and the excess over the null exceeds the raw gain in every remaining collection. It also rescues puzzles, whose raw interval in Table 6 covers zero while its excess over the null does not. The menu experiment clears the null on both readings. Among the collections entering this table it is the one that observes menu choice directly; the twelve ranking collections induce subset choices from complete rankings.

Dataset	Observed	Null median	Excess	MC tail
Menu experiment, forced choice	+0.0265	−0.0041	+0.0306	0.010
Menu experiment, all subjects	+0.0100	−0.0023	+0.0123	0.015
Occupational prestige	+0.0390	−0.0227	+0.0617	≤ 0.005
Sushi	+0.0111	−0.0063	+0.0175	≤ 0.005
GSS socialization	+0.0055	−0.0066	+0.0121	≤ 0.005
Political goals	+0.0069	−0.0012	+0.0081	≤ 0.005
GSS job values	+0.0017	−0.0047	+0.0065	≤ 0.005
Jester file 3	+0.0024	−0.0015	+0.0039	≤ 0.005
Jester file 1	+0.0020	−0.0014	+0.0034	≤ 0.005
Jester file 2	+0.0018	−0.0014	+0.0032	≤ 0.005
Dots	+0.0015	−0.0014	+0.0029	≤ 0.005
Puzzles	+0.0007	−0.0011	+0.0019	≤ 0.005
Netflix	+0.0009	−0.0001	+0.0010	≤ 0.005
Sports participation	+0.0050	+0.0051	−0.0000	0.51

Table 10: The same gain when the axiom holds by construction. Two hundred replicates per row. The diagnostic is the excess of the observed gain over the null median; the tail probabilities are indicative rather than exact tests, and for the menu rows they are optimistic, because the null redraws each subject's thirty-one choices independently given the aggregate worths and so omits the within-subject dependence that the bootstrap intervals of Table 3 do preserve. Wherever shares rest on more than about two thousand observations the null is negative, because contraction is then a bias rather than useful insurance, so comparing an observed gain with zero understates it.

5.8 Boundary conditions

The four losses in Table 1 are one task, and three further populations fail both maps in ways that published aggregates settle without this protocol. Taken together they describe a boundary that can be stated as a rule rather than a list.

A perceptual continuum. Stewart, Brown and Chater [28] had listeners name one of ten pure tones, and ran the same tones with the middle eight and the middle six labels, so the label sets nest over physically identical stimuli. Calibrating on the ten-label row and predicting the restricted rows, linear linear renormalization wins all four conditions, by 0.0133 and 0.0051 nats at narrow spacing and 0.0173 and 0.0041 at wide. One row of the wide ten-to-eight condition is excluded: it contains an exact zero, which has to be floored, and two independently written calibrators disagree on the resulting location by enough to move that condition's gain. The other fifteen rows agree between them to 10^{-4}. The mechanism is visible in the matrices, which are strongly banded: a tone is confused almost entirely with its immediate neighbours, so the alternatives are near-substitutes by construction and removing the outer labels returns their mass to neighbours rather than proportionally. The loss is larger at six labels, where more of the continuum has been cut away.

A survivor's near-substitute. Scottish juries return one of three verdicts, and Ormston and colleagues [31] ran 64 mock juries on films that were identical except for the final section, in which the judge names the verdicts available. Half had three verdicts and half had two. Renormalizing

the three-verdict shares onto the survivors predicts Guilty at 54.9 per cent against an observed 38, and after deliberation 61.1 against 31. Both defaults fail, because the ordering *reverses*: Guilty leads Not guilty with three verdicts and trails it with two. Contraction moves odds toward one and never crosses over, so the race fails here too, and only slightly less. Not proven is a near substitute for Not guilty, so deleting it returns its mass to that neighbour rather than proportionally, which is Debreu's objection in a courtroom, on a fixed stimulus. Townsend and Landon's diagnosis [23] is the same one: their rule fails on the subset containing similar letters and holds on the subset that does not.

The same boundary inside one experiment. Getty, Swets, Swets and Green [19] had three observers identify eight complex sounds with all eight responses available, and then with only four, a different four in each of three conditions, the labels being the stimulus numbers. The stimulus set never changes; only the menu does. Over 72 cells the race wins by +0.0272 nats, and by +0.0111 on the rows whose stimulus survives in the menu, cell bootstrap [+0.0021, +0.0225]. But it loses one condition, and which one is not arbitrary.

Condition	Survivors	Within-set confusion	Gain
1	$\{1, 2, 5, 6\}$	0.103	+0.0453
3	$\{1, 3, 5, 7\}$	0.335	+0.0490
2	$\{3, 4, 5, 6\}$	0.790	−0.0127

Table 11: Within-set confusion is the fraction of a surviving stimulus's errors on the *full* eight-way menu that land on another survivor of that condition. It uses the master matrix only, so it is available before any restricted-menu observation exists. The one condition the race loses is the one whose survivors are each other's confusions.

The statistic turns the near-substitute condition from a description applied after a failure into a quantity computable in advance. Townsend and Landon [23] were using the same diagnostic in 1982, when they reported their rule holding on one subset and failing on the subset containing similar letters. With three observers the bootstrap resamples cells rather than observers and understates uncertainty, so only the signal-row interval excludes zero. The authors also report that observers retuned the weights they placed on perceptual dimensions to maximise discriminability of whichever subset they had to identify, with feedback given only on that subset, so both boundary conditions are present in this experiment at once.

Removal that changes the task. McMurray and colleagues [29] ran a two-label phoneme identification and a four-label version nesting it, on identical synthetic stimuli. The two extra labels absorbed under one per cent of responses, so linear renormalization predicts almost no change and contraction predicts movement toward even; the observed identification slope instead grew steeper, which is neither. Zoanetti and colleagues [30] report the same direction in assessment: deleting a distractor raised the share choosing the correct answer by more than proportionally. In both cases the survivors became easier to tell apart, so they are not the same alternatives before and after, and no map from old shares to new ones can be right.

A twelve-line continuum, within subject. Unpublished data from the Perception and Cognition Laboratory at the University of Missouri, public in `PerceptionCognitionLab/data0`,

has subjects learn a fixed number for each of twelve line lengths and identify them under feedback. The instructions state that "the number assignment for each line is constant throughout the experiment", and some blocks offer all twelve lines while others offer a named subset, so the same subject supplies both the full-menu distribution and the restricted one over identical labels. Conditions A and D restrict to $\{0,1,2,3\}$, $\{4,5,6,7\}$, $\{8,9,10,11\}$, $\{0\dots7\}$ and $\{4\dots11\}$; condition C restricts to twelve named pairs. Restricted blocks are bracketed by full-twelve blocks before and after.

Over 1,296 cells and 49 subjects, linear renormalization predicts better in every split. The gain is -0.0134 overall with a subject bootstrap of $[-0.0176, -0.0104]$, and by size of the surviving menu it is -0.0057 at eight, -0.0090 at four and -0.0322 at two, so the loss grows as more of the continuum is cut away. This is the largest disagreement in the paper running against Gaussian renormalization.

It is also the boundary rule's own prediction. Line length is the unidimensional continuum of the absolute-identification literature, and the rule was fixed in a circulated draft before this data was obtained. Together with the tone matrices and the Getty condition whose survivors are mutual confusions, it is the third collection of that kind and the third loss.

Induced values on a continuum. Duffy and Smith [18] pay subjects to select the longest of between two and six grey lines. Value is induced and one-dimensional by construction, which the authors make the point of the design: "objects are valued according to only a single attribute with a continuous measure and we can observe whether the choice was optimal". The boundary rule therefore predicts linear renormalization, and that is what their data gives.

Their lengths are redrawn on every trial, so there are no fixed alternatives carrying stable shares and the parameter-free protocol cannot be applied. The parametric version of the same question can: fit one scale parameter per model on menus of one size, hold it fixed, and score menus of another size by log likelihood per observation. Calibrating on the six-line menus and predicting the two-line menus, the restriction direction proper, linear renormalization is ahead by 0.0037 nats per observation with $t = 2.28$; calibrating on two and predicting three to six gives $+0.0059$, and on five and six predicting two and three, $+0.0051$. The fitted noise is $\sigma \approx 11$ pixels against a 16-pixel spacing between adjacent lengths.

One free parameter each makes this fair between the two models and unlike everything else in this paper, where neither map fits anything. It is reported because it is the same question on a dataset built to be one-dimensional, and because a referee who runs it will get this answer.

Withdrawn alternatives with negligible mass. A large top share is not by itself enough to make the two maps agree. Calibrating Case V to $p = (0.90, 0.09, 0.01)$ and removing the 0.01 alternative gives the favourite 0.9077 against a renormalized 0.9091, a difference of 0.0014. Removing the 0.90 favourite instead gives the surviving 0.09 alternative 0.8028 against a renormalized 0.9000, a difference of 0.0972. The condition that makes the two maps agree is that the withdrawn alternatives carry negligible mass and the dominant alternative survives, not that the shares are concentrated. The low-cost gamble frame, the ballot rows and the Wikipedia rows are of that kind, which is why they are draws or near-zero differences rather than evidence.

Figure 1 shows the same division on one axis. The collections nearest linear renormalization are the perceptual continuum and the concentrated shares; the collections furthest past Gaussian renormalization are categories, colours, symbols and consumer goods.

The rule that summarises this is narrower than the class of independent random utility models

and wider than any single dataset. *The Gaussian contest is the better default where the alternatives are distinct unordered items*, such as goods, news articles, memory foils, sushi, jokes, ballots, colours and symbols. It is not where they lie on a perceptual continuum, where a removed option is a near-substitute for a survivor, or where removal changes the discriminability of what remains.

Two consequences follow. First, an apparent conflict resolves. Meyer-Grant and colleagues [32] show that Gumbel-min uniquely predicts accuracy invariance as a recognition set grows uniformly, find that invariance, and report the Gumbel model outperforming the Gaussian throughout. That is Yellott's own condition, independently confirmed, in the same domain where Section 5.2.2 has the race winning on nested foils. Both hold, and their footnote supplies the reconciliation: invariance breaks down when systematic similarity is introduced, because latent strengths cease to be independent. Our foil split found contraction equal across same- and cross-category foils while the tones show similarity mattering greatly, which says it is the continuum that breaks independence and not category membership. Two photographs of different dogs are far less confusable than two tones six per cent apart in frequency.

6 Scope

The practical recommendation is narrow, and Section 5.8 makes it conditional. When a probability vector is transported to a smaller support and the alternatives are distinct unordered items, Gaussian renormalization should be tried alongside linear renormalization rather than passed over. Four questions decide whether that condition holds. None of them is settled by the share vector alone, and how well they can be answered in advance depends on what else is available. Do the alternatives lie along a single physical or perceptual dimension, so that neighbours are more confusable than distant pairs? Is any alternative a near-substitute for another, in the sense that a chooser who loses one would take the other rather than spread? Does removing an alternative change how easily the survivors can be told apart, as deleting a distractor or an implausible response option does? And do the withdrawn alternatives carry so little mass, with the leader surviving, that the two maps agree to within the noise?

A yes to the first two says linear renormalization or a fitted correlated model rather than Gaussian renormalization, though in the tone matrices and the Scottish verdicts linear normalization is merely the less wrong of two independence models, and near-substitution is an argument for a similarity-aware model rather than for proportional redistribution. A yes to the third says that neither transport map applies, because the surviving alternatives are not the alternatives that were there before, and the choice model has to be re-estimated under the new task. A yes to the fourth says the choice does not matter.

A confusion matrix answers the first two on inspection, which is the diagnostic Townsend and Landon [23] used, and the within-set confusion mass of Table 11 makes it numerical. That is much richer information than a share vector, and for goods, news articles, candidates, jokes and occupations no such matrix exists. There the questions have to be answered from the feature geometry or the semantics of the alternatives, and cannot be read off the shares.

One further caveat belongs here rather than in the theory. Distinct unordered alternatives do not imply $\Sigma = I$. Goods, articles, candidates and occupations can carry clustered and correlated tastes, and Section 4.2 shows the limit theorem delivering a Gaussian race with covariance Σ, of which Case V is the case that is isotropic on the contrast subspace.

The inversion from shares to locations is the only extra machinery, and the lattice algorithm

of [2] performs it in one pass over a common grid, returning the whole vector of winning probabilities at once. The restricted predictions are then one-dimensional quadrature on the same grid. At the field sizes here, ten alternatives at most, both steps are immediate; the method is designed for and tested at much larger fields [2]. So the recommendation needs no additional restricted-menu observations and no appreciable additional computation, although the boundary conditions below require structural information about similarity, ordering or task geometry that the shares do not contain. Within its condition it was better on every population examined here. Outside it, it was worse on all four tone rows, all three line-restriction rows and the Getty condition whose survivors are mutual confusions.

All twelve ranking collections carry an assumption the menu experiments do not. A respondent's ranking imposes one ordering on every subset by construction, so no ranking dataset can show a person choosing i over j from one menu and j over i from another, and none reveals whether that person would choose as their ranking implies when handed two options. Reading restriction off a ranking assumes they would. Section 5.2.1 does not.

Every quantity here is a population quantity. A population of people who each choose by Equation (1), with worths differing between people, generally does not satisfy the axiom in aggregate, which is why mixed logit is useful and why its probabilities are integrals over logit probabilities rather than logit probabilities [7]. Aggregate failure of the axiom is therefore not evidence that individuals violate it. Heterogeneous Luce choosers can reproduce aggregate contraction, and Section 4.2 exhibits a class in which they do. What is rejected is a representative-agent map.

The experiment compares two restriction maps, not two fitted likelihoods. A fitted Plackett–Luce model [36] uses the whole ranking and a fitted Bradley–Terry model [10] uses the pairwise outcomes directly. Neither was estimated here. Recommendations about ranking or paired-comparison likelihoods are therefore outside what these tables measure.

The score is a sum of marginal scores for top-item-within-subset outcomes. It is proper for those marginals and not for the ranking distribution. With $K = 4$ the scored marginals carry at most 17 degrees of freedom against 23 for a distribution over rankings, so distinct ranking laws can agree on everything scored here.

The independent-noise family nests Luce at the Gumbel law, so it excludes no Luce population. Choosing the Gaussian point of it is a separate decision, and Table 10 shows that decision losing when the axiom holds and the shares are well estimated.

Neither map dominates the other. When the shares are precise the better map is whichever matches the population, and when they are very noisy the dominant consideration is estimation risk rather than either structural rule. The recommendation above is for the case where an analyst holds aggregate shares and has no evidence about the axiom either way.

One limit is larger than the effect measured here. In the after-the-stimulus arms of Section 5.2.2 the observed restricted probability is available, so both maps can be checked against it rather than only against each other, and both err in the same direction by more than they differ from one another: Gaussian renormalization removes a fifth to a third of linear renormalization's error and no more. Table 9 is not a second such check. Observing the winner conditions on a realised contest instead of withdrawing an alternative from it, and neither rule compared there is the ordering law of the model it is named for, so it measures a heuristic error rather than the accuracy of either map. Choosing the better of these two maps is therefore a smaller change than fitting a shape parameter, wherever a restricted-menu observation exists to fit one. The case made here is for the situation where none does.

Pricing ordered outcomes by removing the winner and re-running is not the Gaussian contest's own prediction, and the two come apart on individual pairs and triples. We leave to future work the comparison of the exact Gaussian ordering law against Plackett–Luce ranking probabilities on complete-ranking data.

7 Conclusion

We have attempted to elevate the Thurstone Case V $N(0,1)$ model, which we have simply termed *Gaussian renormalization*, to the status of an equal partner to the ubiquitous practice of linear renormalization of probability. The question is not confined to menus of goods. It arises wherever a probability vector is transported to a smaller support, which includes masked softmax vocabularies, scratched fields, delisted catalogues and disallowed response sets. Four operations get run together under that heading and only two of them raise a question.

Conditioning a fixed categorical law on $X \in T$ gives $p_i / \sum_{j \in T} p_j$ by definition. That is what truncation ordinarily means in probability, linear renormalization is a theorem there rather than a choice, and nothing in this paper bears on it. Masking a fixed softmax and rescaling is the same arithmetic on a score vector: it does not merely permit the linear transport, it *is* the linear transport, so a masked softmax has IIA built into the operator. Neither is a case where two maps compete.

The question arises in the other two. When the feasible set is intervened upon, so that a horse is scratched or a response option is disallowed and the choice-generating experiment itself changes, the original share vector does not determine the new shares, and the two maps below are two conventional candidates among the many any regular noise shape supplies. And when a model recomputes its scores with the permitted set as an input, the result is not a transport of the old vector at all, so neither map describes it. This paper is about the third case, and the second is where its practical relevance lies, because that is where the linear transport is already deployed without anyone having chosen it.

Gaussian Case V is a credible, tuning-free alternative to linear renormalization. Across the collections examined it frequently produces useful odds contraction, with the most persuasive evidence coming from experiments in which the same observers perform tightly matched full-menu and restricted-menu tasks, the identity of the surviving alternatives being revealed only after the stimulus has gone. On those collections it is typically as viable as linear renormalization and usually better. It has the lower held-out log loss in thirty of the thirty-nine comparisons, and on the contraction axis of Figure 1 it is the nearer of the two maps in eighteen of the nineteen populations that axis can place, the exception resting on five pairs. Neither statement identifies a population rate, and the second is a projection rather than a proper score. What they jointly rule out is the position that the two maps trade blows.

The evidence does not establish universal superiority. Raw predictive gain must be distinguished from rejection of a Luce null, and Table 1 reports the two separately because the ballot, Wikipedia and low-cost gamble rows show them coming apart. Ecological comparisons must be separated from controlled restrictions, since a recommender, an electorate or a month of Wikipedia traffic can shift the population as well as the menu. And similarity, correlated tastes, task adaptation and the identity of the alternatives removed remain boundary conditions rather than details.

Four such conditions are worth carrying away, and none of them is settled by the share vector alone. The alternatives lie on a perceptual continuum, so that neighbours are more confusable

than distant pairs. A withdrawn option is a near-substitute for a survivor, so its mass returns to that neighbour rather than spreading. The withdrawal changes how easily the survivors can be told apart, in which case neither transport map applies and the choice model has to be re-estimated under the new task. Or the withdrawn alternatives carry negligible mass while the leader survives, in which case the two maps agree to within the noise. Figure 1 places the populations we can place along that axis. Where a confusion matrix exists, the within-set confusion mass of Table 11 measures the first two conditions directly. Where only shares exist, they have to be judged from the geometry or the semantics of the alternatives, and the judgement can be wrong.

To repeat, this note is about competing reflexes rather than competing careful modelling paradigms. Both maps are nested by the independent additive random-utility family, indexed by its standardized noise shape. Where both can be checked against an observed restricted probability, which is the after-stimulus perceptual arms of Section 5.2.2, both err in the same direction and by more than they differ from one another. The favourite's second-place exercise is not such a check, because observing the winner is conditioning on a realised contest rather than withdrawing an alternative from it, and neither of the rules compared there is the ordering law of the model it is named for. A model whose shape parameter nests the two can be fitted wherever a restricted-menu observation exists, and estimated under the same objective it is no worse in sample; whether it improves held-out prediction needs separate validation. A common noise scale cannot serve as that parameter, being a gauge.

We leave to future work the comparison of the exact Gaussian ordering law against Plackett–Luce ranking probabilities on complete-ranking data.

Data and code

Sushi rankings come from Kamishima [4]; Jester ratings from the Eigentaste project [3]; the General Social Survey cumulative file from NORC; the perceptual, leisure, prestige, film and political-goal rankings via PrefLib [6] and the CRAN packages PLMIX and ConsRank; the menu-choice data from the replication archive of [1]; the colour and symbol naming trials from the OSF archive of [26], with the advance-warning condition recounted from the raw session files because the authors' aggregate matrices omit it; the recognition trials from the OSF archive of [27]; the tone confusion matrices digitized from the vector figures of the author manuscript of [28]; the jury shares from the published tables of [31]; the eight-way and four-way sound-identification frequencies transcribed from Tables 6 and 8 of [19]; the letter confusion matrices transcribed from Tables 1 to 4 of [23]; the categorization trials from the CAM1 deposit of [25]; and the twelve-line identification trials from the public `PerceptionCognitionLab/data0` repository, which carries no licence file and no associated publication, so its authorship is recorded as unattributed. All are public.

Analysis code and every input needed to reproduce the tables are under `research/restriction` in `github.com/microprediction/winning`, with the inputs as column extracts rather than raw archives, one script per population, and, for the collections transcribed or downloaded for this paper, a `SOURCE.md` giving the fetched access route and the caveats. Each transcribed table carries its own arithmetic check: every row of the sound-identification and letter matrices reproduces its printed total, and the recounted Yeon–Rahnev condition reproduces the authors' saved matrix exactly.

Two storage conventions must be told apart when reproducing this. BayesMallows, Con-

sRank and PerMallows store ranks, so element j holds the rank of alternative j, while PLMIX stores orderings, so the first element names the favourite. Reading one as the other transposes the data silently and raises no error. We established which is which by cross-validation: the political goals data ships in two of these packages and they agree only when each is read under its own convention.

A Census of restriction experiments

The experiments were run in four literatures that do not cite one another: speech confusion, visual absolute identification, categorization and odor identification. Most of them published percent correct rather than cell-level shares, which scores neither map. That describes the majority of the sources we found.

We assembled a census of about a hundred sources, recording design, what numbers were printed or deposited, a fetched access route, and a usability verdict, negatives included. It is in the repository under `research/restriction/notes/crr`. Table 12 lists the nine sources that restrict a response set over shared alternatives, are not scored above, and could therefore still change the answer.

Source	Domain	Restriction	Status
Clarke [12]	speech in noise	$8 \to 4$, $6 \to 3$	library access
Pollack and Decker [13]	consonants	$8 \to 4$, three sets	not scoreable[3]
Hodge and Pollack	auditory, unidimensional	nested	library access
Lacouture et al.	line lengths	nested	library access
Kent and Lamberts	dot separation	$N = 2, 4, 6, 8, 10$	library access
Rouder lab, unpublished	line lengths	twelve, nested	raw trials, open
Negoias et al.	odor descriptors	$16 \to 8 \to 4$	never published
Jäger et al.	flavour terms	nested, $n = 735$	library access
Hörberg et al.	odor naming	free, then cued list	data request

Table 12: Restriction experiments not scored above. “Printed” means the matrices appear in the article and need transcription, as Townsend and Landon’s were for Section 5.4. The Rouder laboratory data is unpublished but public, with a master set, nested subsets and stimulus and response logged as global indices, which is the property that makes labels comparable across menus. Four large odor corpora are open and usable but carry a fixed four-descriptor menu, so they calibrate and cannot score.

None of the entries is a preference task, so the corpus scored above is not representative of this literature by domain. Two need only library access to a printed table rather than a data request, and both are from the 1950s.

References

[1] Miguel A. Costa-Gomes, Carlos Cueva, Georgios Gerasimou, and Matúš Tejiščák. Choice, deferral, and consistency. *Quantitative Economics*, 13(3):1297–1318, 2022.

[3] Pollack and Decker’s three overlapping four-way matrices were reduced to mean absolute deviations before publication, one scalar per set per signal-to-noise level, so the subset matrices do not exist in print. Their four eight-way masters were recovered and are scoreable only as master matrices.

[2] Peter Cotton. Inferring relative ability from winning probability in multientrant contests. *SIAM Journal on Financial Mathematics*, 12(1):295–317, 2021.

[3] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: a constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[4] Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 583–588, 2003.

[5] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

[6] Nicholas Mattei and Toby Walsh. PrefLib: a library for preferences. In *Algorithmic Decision Theory*, volume 8176 of Lecture Notes in Computer Science, pages 259–270, 2013.

[7] Daniel McFadden and Kenneth Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

[8] Maria M. Robinson, Isabella C. DeStefano, Edward Vul, and Timothy F. Brady. How do people build up visual memory representations from sensory evidence? Revisiting two classic models of choice. *Journal of Mathematical Psychology*, 117:102805, 2023.

[9] Victor H. Aguiar, Maria Jose Boccardi, Nail Kashaev, and Jeongbin Kim. Random utility and limited consideration. *Quantitative Economics*, 14(1):71–116, 2023.

[10] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[11] Frank R. Clarke and Clint D. Anderson. Further test of the constant-ratio rule in speech communication. *Journal of the Acoustical Society of America*, 29(12):1318–1320, 1957.

[12] Frank R. Clarke. Constant-ratio rule for confusion matrices in speech communication. *Journal of the Acoustical Society of America*, 29(6):715–720, 1957.

[13] Irwin Pollack and Louis Decker. Consonant confusions and the constant ratio rule. *Language and Speech*, 3(1):1–6, 1960.

[14] Frank R. Clarke. Constant-ratio rule for confusion matrices in speech communication. *Journal of the Acoustical Society of America*, 31(6):836, 1959. Abstract of a paper presented to the Acoustical Society.

[15] James P. Egan. Monitoring task in speech communication. *Journal of the Acoustical Society of America*, 29(4):482–489, 1957.

[16] Charles M. Holloway. Passing the strongly voiced components of noisy speech. *Quarterly Journal of Experimental Psychology*, 20:336–350, 1968.

[17] Charles M. Holloway. A test of the independence of linguistic dimensions. *Language and Speech*, 14(1):19–28, 1971.

[18] Sean Duffy and John Smith. An economist and a psychologist form a line: what can imperfect perception of length tell us about stochastic choice? *Theory and Decision*, 99(3):701–734, 2025.

[19] David J. Getty, John A. Swets, Joel B. Swets, and David M. Green. On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics*, 26(1):1–19, 1979.

[20] Milton H. Hodge. Some further tests of the constant-ratio rule. *Perception & Psychophysics*, 2(10):429–437, 1967.

[21] Robert Rich. Constant ratio rule for confusion matrices from short-term memory experiments. *British Journal of Psychology*, 62(1):27–35, 1971.

[22] B. J. T. Morgan. On Luce's choice axiom. *Journal of Mathematical Psychology*, 11(2):107–123, 1974.

[23] James T. Townsend and Douglas E. Landon. An experimental and theoretical investigation of the constant-ratio rule and other models of visual letter confusion. *Journal of Mathematical Psychology*, 25(2):119–162, 1982.

[24] Jeffrey N. Rouder. Modeling the effects of choice-set size on the processing of letters and words. *Psychological Review*, 111(1):80–93, 2004.

[25] A. J. Wills, Stian Reimers, Neil Stewart, Mark Suret, and I. P. L. McLaren. Tests of the ratio rule in categorization. *Quarterly Journal of Experimental Psychology*, 53A(4):983–1011, 2000.

[26] Jiwon Yeon and Dobromir Rahnev. The suboptimality of perceptual decision making with multiple alternatives. *Nature Communications*, 11:3857, 2020.

[27] Igor Utochkin, Daniil Azarov, and Daniil Grigorev. Invariant recognition memory spaces for real-world objects revealed with signal-detection analysis. *Psychological Science*, 36(11):831–845, 2025.

[28] Neil Stewart, Gordon D. A. Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological Review*, 112(4):881–911, 2005.

[29] Bob McMurray, Richard N. Aslin, Michael K. Tanenhaus, Michael J. Spivey, and Dana Subik. Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1609–1631, 2008.

[30] Nathan Zoanetti, Paul Beaves, Patrick Griffin, and Euan M. Wallace. Fixed or mixed: a comparison of three, four and mixed-option multiple-choice tests in a fetal surveillance education program. *BMC Medical Education*, 13:35, 2013.

[31] Rachel Ormston, James Chalmers, Fiona Leverick, Vanessa Munro, and Lorraine Murray. *Scottish Jury Research: Findings from a Large Scale Mock Jury Study*. Scottish Government, 2019. ISBN 978-1-83960-194-1.

[32] Constantin G. Meyer-Grant, David Kellen, Samuel M. Harding, and Henrik Singmann. Extreme-value signal detection theory for recognition memory: the parametric road not taken. *Psychological Review*, published online 4 June 2026. doi:10.1037/rev0000615.

[33] Eric W. Holman and A. A. J. Marley. Stimulus and response measurement. In E. C. Carterette and M. P. Friedman, editors, *Handbook of Perception, Volume II: Psychophysical Judgment and Measurement*, chapter 7, pages 173–213. Academic Press, 1974.

[34] Wayne Lee. Detection theory, micromatching, and the constant-ratio rule. *Perception & Psychophysics*, 4(4):217–219, 1968.

[35] R. Duncan Luce and Patrick Suppes. Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume III, pages 338–339. Wiley, 1965. Theorem 32 there carries the footnote that the proof is due to Eric Holman and A. A. J. Marley, with no title or date given; this is the source Yellott [39] cites at pp. 110 and 123, rather than [33].

[36] Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society, Series C*, 24(2):193–202, 1975.

[37] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[38] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: a large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, 2020.

[39] John I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144, 1977.