Think Before Rescaling Probabilities to Unity: Evidence from Thirty-Nine Tests of Linear versus Gaussian Renormalization

Peter Cotton

First version: August 12, 2026. This version: November 2, 2026.

Abstract

Rescaling a probability vector when an item is withdrawn or masked out is one of the great unconscious modelling acts. When the support shrinks because something was learned about a fixed experiment, rescaling is the conditional law. When somebody instead intervenes on what may be chosen, the old vector does not determine the new one, and rescaling amounts to assuming Gumbel noise in an additive random-utility model[1]. A Gaussian is an equally canonical, equally tuning-free choice. We race this *Gaussian renormalization* against linear renormalization on thirty-nine held-out comparisons, re-litigating the constant-ratio rule versus Thurstone's Case V. Gaussian renormalization has the lower log loss in thirty rows and leads in fourteen sources against one. Differences are small: a median of 0.004 nats.

1 A two-horse race between zero-parameter choice models

Suppose a probability vector is presented, as with a collection of de-vigged bookmaker odds for a horse race, or the output of a calibrated probabilistic machine learning model. Suppose next that the set changes. Perhaps a horse is scratched, a product is delisted, or a primary candidate drops out of a race. Alternatively a response option in a survey may be disallowed, or a token masked out of a softmax. The situations differ, but in each case a probability vector has to be transported from the original support to a smaller one. The natural question arises: what are the probabilities over what remains?

Occasionally there is a definitive answer. Suppose the probability vector describes a fixed experiment and we learn that the outcome fell in T. Then $p_i / \sum_{j \in T} p_j$ is the conditional law by definition, and there is nothing to choose. Masking a fixed softmax and rescaling the survivors performs that same arithmetic on a score vector, so wherever the target really is the conditional law, that operation is right by construction and we will not consider it further.

However, linear renormalization of probabilities is very often the default in the much more general setting where a transport is needed: the obvious choice is reached for, little thought goes into its justification, and empirical results might not support it after the fact. In that setting somebody intervenes on the feasible set. The horse is scratched, the option is disallowed, or the action is removed, and the experiment that generates the choice is no longer the one that generated the original shares. In general the old vector does not determine the new one, and any

[1]The Gumbel construction is due to Holman and Marley [39, 37]; the uniqueness is Yellott's [43].

answer is a modelling assumption. The same masking arithmetic is often applied here too, as one such assumption.

Writing $Y(S)$ for what a chooser facing menu S selects, and $P_S(i) = \Pr\{Y(S) = i\}$ for the share it produces, the cases separate as below. The table adds a fourth, a model that recomputes its scores with the permitted set as an input; its output is not a transport of the old vector, and it too is set aside.

Operation	Target	Status
Outcome information	$\Pr\{Y(S) = i \mid Y(S) \in T\}$	linear by definition
Mask on a fixed score vector	the same quantity, computed	linear by construction
Feasible-set intervention	$P_T(i)$	not identified by P_S
Menu-aware recomputation	$P_T(i)$, estimated directly	no transport needed

The first two condition the law of $Y(S)$; the third asks about $Y(T)$, a different random variable produced by a different experiment, and that is what P_S alone cannot determine. We write $P_S(i)$ rather than $P(i \mid S)$ throughout, because the conditioning bar is exactly the confusion at issue.

A race shows the intervention plainly. When a horse scratches, the win, exacta and trifecta odds on the field must be recomputed. Finishing times have a joint distribution, and scratching a horse removes its time from the race. Nothing is conditioned on: the survivors' times keep the law they had, with the scratched horse's coordinate integrated out, and the new odds follow by re-running the race among the survivors. That alone is insufficient information to settle the answer.

Re-running the race requires a law for the performances, and that law might be decided upon implicitly, even unwittingly: under independent Gumbel performances re-running the race is exactly linear renormalization. A race, however, supplies a physical story for the latent draws, and under that story Gumbel performances might not be the most natural a priori choice. Here the convenience of proportionate scaling of probability can lure the modeller into a possibly unnatural choice. This is so well appreciated that we will not dwell on it.

In contrast, a market share, or the popularity of an item on a menu, is a revealed preference with no such story behind it. There the choice of renormalization, made by reflex, might be equally misleading yet not so plainly contradicted by any obvious generative model, unless one is given to thinking that the mechanism of choice inside our brains might itself be a race of sorts.

When a menu item is removed, quite a lot of machinery exists for carefully estimating the probabilities that should be ascribed to the reduced menu. Depending on the data available, probit models or other tools can be brought to bear. Those need restricted-menu data, model selection and tuning, and comparing them with the instinctive act of rescaling probability would be unfair to both. The comparison here is between defaults that need none of that.

To emphasize this last point, the scope of our inquiry is limited to a single competitor to linear renormalization. There is one alternative response that is only slightly more involved analytically but not methodologically complicated. It is the Gaussian special case of Thurstone's Case V family, and we argue for its place as a tuning-free default renormalization to set against the linear one. Both stand on equal footing philosophically.

Nothing has been observed about the smaller menu, so the answer must come from an assumption about how the original shares arose. The standard rescaling 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. 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*. Testing for departures from it directly, rather than through a restriction protocol, is its own literature; Corrêa et al. [35] give a recent Bayesian treatment for similarity choices.

We suggest that Gaussian renormalization is equally natural, if not more so. Read the shares as the winning probabilities of a contest among independent Gaussian performances, which has heritage back to Thurstone's Case V [41] 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,

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

How should we choose between them? We content ourselves with a practical comparison of out-of-the-box linear renormalization and Gaussian renormalization, without being sidetracked by, and without minimizing, the fact that they are merely two points in a family that is itself a gateway to many modelling possibilities: the beginning of the modelling process, not the end by any means. Our justification for holding only a two-horse match race lies in the ubiquity of linear renormalization of probability. In short, two horses are better than one, and we present evidence that the less popular horse might in fact be the more reliable.

This race is a fair one. Linear renormalization holds the ratios of the shares fixed and Gaussian renormalization holds the latent locations fixed. Neither carries a parameter fitted to the data being predicted, which is the sense in which both are free of tuning; the pipeline around them does use one shared convention, the continuity correction $\alpha = 1/2$, and Section 5.4 varies it. The underlying physics is quite different, so even small persistent statistical advantages of one over the other may be suggestive of structure, and of interest to the psychology community but also, we hope, to the broader machine learning audience whose choice of linear renormalization is often tacit. The tacit assumption is not harmless: Xu et al. [33] trace perverse incentives in reinforcement learning from human feedback to exactly this assumption, since the reward models it fits inherit Independence of Irrelevant Alternatives from the same constant-ratio rule discussed below.

Each competitor in this two-horse race is elementary and, not surprisingly, has assumed many names. We treat them as synonyms. We equate 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.

Conversely we equate 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

we present has not been made broadly or crisply 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 places both maps in the independent additive random-utility class, and Proposition 1 makes the direction of their disagreement exact.

2 Scoreboard

Figure 1 and Table 1 give the full result before the populations behind them are introduced one by one in Sections 3 and 5.

The figure asks where each population falls between the two kinds of renormalization, on a scale where linear renormalization is one point and Gaussian renormalization another. Of the twenty it can place, nineteen lie nearer the Gaussian point, the median sits at 1.13, and fifteen lie beyond it.

The single population nearer linear renormalization comes from one experiment in which subjects chose among gambles whose values had to be computed, which takes on a bit of the flavour of a quiz and so might be classified separately. The figure is suggestive rather than decisive, since the statistic it plots is not a proper scoring rule and its ratio is unstable where the Case V slope is small.

Table 1 scores by held-out log loss the thirty-nine comparisons that have full-menu shares to calibrate both maps and restricted-menu choices to score them. It gives linear renormalization no natural advantage.

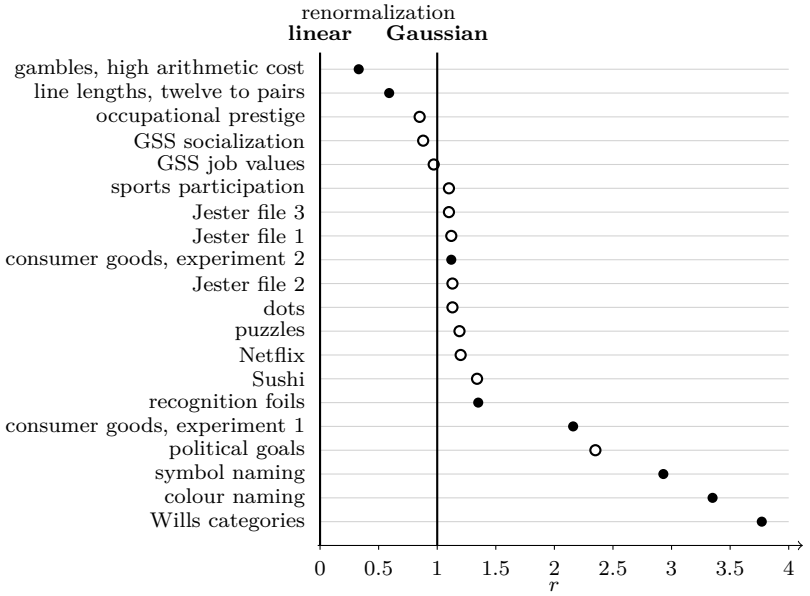


Figure 1: Where each population sits between the two maps. The coordinate is $r = \lambda_{\text{observed}}/\lambda_{\text{Case V}}$, the observed contraction slope divided by the slope Case V implies for that population's own shares; λ is defined formally in Section 4.2 (Table 3), fitted here on pairs. 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. A population appears when a restricted pairwise share q_{ij} exists on the alternatives that calibrated the map. The tone, Getty and Townsend rows and the eight- and four-way line restrictions restrict to sets of three or more, so no q_{ij} exists; the news, ballot and Wikipedia rows have no matched pair on fixed alternatives. Each consumer experiment is plotted once, all subjects, and the gambles once, at high arithmetic cost, which leaves twenty populations. Netflix draws on five separate PrefLib files of three pairs each, and combining them requires calibrating each on its own and pooling the fitted numerator and denominator, not pooling raw shares. Pooling the shares directly would mix five unrelated three-film comparisons into one and put Netflix off the scale at $r = 6.07$. Calibrated file by file, it sits at $r = 1.20$, $[0.92, 1.52]$, indistinguishable from the Gaussian point at this sample size. The ratio is ill-conditioned wherever the Case V slope is small, so the extremes of this axis are its least certain points. The statistic is not a proper score, and Table 1 orders the populations differently.

Two numbers are reported for each population. The *gain* is the held-out log loss of linear renormalization minus that of Gaussian renormalization, in nats per prediction, so a positive gain means Gaussian renormalization forecast the withheld choices better. The *null frequency* is obtained by generating a population that obeys the constant-ratio rule exactly, with worths set to the observed shares, pushing it through the identical pipeline, and recording how often that population produces a gain at least as large as the observed one. It is a parametric bootstrap frequency under one particular joint law, Plackett–Luce for the ranking collections and independent redraws for the menu experiment, and not a model-free test of the constant-ratio restriction. It is reported as a diagnostic and should not be read as a p-value.

Two of these constructions, the recognition foils and the news slates, hold the fitted linear

and Gaussian predictions fixed and redraw only the outcome given those predictions, rather than refitting calibration inside every replicate as the ranking-collection and menu-experiment nulls do. Held at the fitted linear prediction q_L, the expected simulated gain is $\mathrm{E}_{q_L}[\log q_G(Y) - \log q_L(Y)] = -D_{\mathrm{KL}}(q_L \,\|\, q_G) \leq 0$ by Gibbs' inequality, so a negative median is partly guaranteed by the construction and not solely evidence that shrinkage would otherwise inflate the gain. These two null frequencies should be read as diagnostics under a fixed-calibration null, not as measuring the calibration uncertainty that the refit nulls elsewhere in the paper do.

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 and the families contribute unequal numbers of them.[4]

Eight of the nine rows where linear renormalization wins have a feature in common. Four are tone identification, three are line identification, and one is the Getty condition whose four surviving sounds are easily mistaken for one another. In all three studies each alternative is confused mostly with its neighbours. Two of them also place the alternatives on a single perceptual dimension, but Section 5.12 shows that dimension alone does not decide between the maps. The ninth is Wikipedia, whose magnitude is -0.0001.

The two columns answer different questions. A positive gain says Gaussian renormalization predicted better from estimated shares, and part of that can be shrinkage rather than structure. A small null frequency says the population departs from the constant-ratio rule in the direction of contraction, and says nothing about which map forecasts better.

Two rows show the questions coming apart. Wikipedia has a negative gain and a small null frequency, so it departs from the rule in the right direction while still predicting worse. Sports participation has a positive gain and a null frequency of 0.51, so its gain is shrinkage and it carries no evidence about the rule. Section 5 takes up the first question and Section 5.10 the second.

Each consumer experiment appears twice because a gain and a null frequency have to come from the same people, and the forced-choice subgroup is not the whole sample. Three further populations appear in Section 5.11 instead of here, since published aggregates fix the direction of their failure but do not support this protocol.

2.1 The same evidence counted by source

Thirty-nine rows are not thirty-nine tests. Four of them are tone conditions from one experiment, three are restrictions of one line experiment, three are Jester files drawn the same way, and the consumer rows are nested subgroups of two studies. The row count is a description of Table 1 and nothing more. Table 2 gives one row per source, one primary specification each, and a verdict that turns on whether the resampling interval excludes zero rather than on the sign of a point estimate.

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

[3] The twelve-line data is unpublished laboratory data with no licence file and no associated publication, so it is cited by its public repository rather than by author. See the data section.

[4] As a finer point, the collections were assembled by searching for restriction data. It could be argued that this is therefore not a fair sample from a population of applications.

Population	Restriction	Gain	Null freq.	Family
News slates	observed impressions	+.0496	.016[2]	clicks
Sounds, eight to four, $\{1, 3, 5, 7\}$	observed menus	+.0490	$\leq .005$	Getty
Sounds, eight to four, $\{1, 2, 5, 6\}$	observed menus	+.0453	$\leq .005$	Getty
Consumer goods 1, forced choice	observed menus	+.0443	.005	Costa-Gomes
Occupational prestige	ranking-induced	+.0390	$\leq .005$	prestige
Wills categories, three to two	observed menus	+.0299	.002	categorization
Symbol naming, six to two	observed menus	+.0277	$\leq .005$	Yeon–Rahnev
Colour naming, four to two	observed menus	+.0147	$\leq .005$	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	$\leq .005$	preference
Gambles, high arithmetic cost	observed menus	+.0101	.035	Aguiar
Letters, five to three, $\{F, H, X\}$	observed menus	+.0092	.005	Townsend–Landon
Political goals	ranking-induced	+.0076	$\leq .005$	preference
Gambles, medium arithmetic cost	observed menus	+.0069	.124	Aguiar
Consumer goods 2, all subjects	observed menus	+.0059	.194	Costa-Gomes
GSS socialization	ranking-induced	+.0057	$\leq .005$	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	$\leq .005$	Utochkin
Jester file 3	ranking-induced	+.0024	$\leq .005$	Jester
Jester file 1	ranking-induced	+.0020	$\leq .005$	Jester
Jester file 2	ranking-induced	+.0018	$\leq .005$	Jester
GSS job values	ranking-induced	+.0018	$\leq .005$	GSS
Netflix	ranking-induced	+.0018	$\leq .005$	film
Dots	ranking-induced	+.0014	$\leq .005$	perceptual
Puzzles	ranking-induced	+.0009	$\leq .005$	perceptual
Letters, five to four, $\{A, E, F, H\}$	observed menus	+.0006	.134	Townsend–Landon
Gambles, low arithmetic cost	observed menus	+.0002	.672	Aguiar
US state ballots 2024	observed ballots	+.0001	.005	ballots
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	−.0056	1.000	lines[3]
Lines, twelve to four	observed menus	−.0089	1.000	lines[3]
Sounds, eight to four, $\{3, 4, 5, 6\}$	observed menus	−.0128	1.000	Getty
Tones, narrow, ten to six	observed menus	−.0132	n/a	Stewart
Tones, wide, ten to six	observed menus	−.0172	n/a	Stewart
Lines, twelve to two	observed menus	−.0322	1.000	lines[3]

Table 1: Held-out log loss of linear renormalization minus that of Gaussian renormalization, in nats per prediction. Here we read a positive gain as a win for Gaussian renormalization. The null frequency is defined in the text. It is a parametric bootstrap diagnostic for departure from the constant-ratio rule, not a p-value and not a statement about predictive superiority. Replicates are two thousand for the twelve ranking-induced rows and two hundred elsewhere, except for the news row, sixty, and the Wills row, four hundred; the tone matrices are participant-averaged, so no null frequency is computed.

Source	Primary specification	Gain	95% interval	Verdict
Controlled: the same subjects or stimuli before and after				
Costa-Gomes et al.	all subjects, both experiments	+.0100	[+.0048, +.0246]	Gaussian
Yeon and Rahnev	four colours, revealed after	+.0147	[+.0117, +.0178]	Gaussian
Townsend and Landon	38 restricted rows	+.0042	[+.0022, +.0064]	Gaussian
Utochkin et al.	360 nested foil cells	+.0037	[+.0027, +.0049]	Gaussian
Getty et al.	all 72 cells	+.0272	[−.0015, +.0618]	draw
Stewart et al.	four tone conditions	−.0041 to −.0172	none possible	unresolved
Twelve-line laboratory data	all restricted blocks	−.0134	[−.0176, −.0103]	linear
Designed, across groups				
Wills et al.	Experiment 2, 39 cells	+.0299	[+.0175, +.0716]	Gaussian
Aguiar et al.	high arithmetic cost	+.0101	[+.0046, +.0246]	Gaussian[5]
Ecological: the menu changed outside the design				
Microsoft News	6,000 nested slate pairs	+.0496	not computed	unresolved
US state ballots 2024	263 state pairs	+.0001	[+.0001, +.0002]	Gaussian[6]
Wikipedia clickstream	3,544 pages	−.0001	not computed	unresolved
Ranking-induced				
Occupational prestige	all proper restrictions	+.0390	[+.0200, +.0759]	Gaussian
Sushi	all proper restrictions	+.0111	[+.0088, +.0142]	Gaussian
Political goals	all proper restrictions	+.0076	[+.0060, +.0096]	Gaussian
General Social Survey	two items pooled[7]	+.0038	[+.0026, +.0070]	Gaussian
Jester	three files pooled[7]	+.0020	[+.0016, +.0027]	Gaussian
PrefLib perceptual	dots task	+.0014	[+.0008, +.0030]	Gaussian
companion task	puzzles task	+.0009	[+.0005, +.0024]	Gaussian
Netflix	five files pooled[7]	+.0018	[+.0006, +.0037]	Gaussian
Sports participation	all proper restrictions	+.0050	[+.0027, +.0167]	excluded[8]

Table 2: One row per source, except that the companion dots and puzzles tasks get a row each. A verdict of Gaussian or linear means the interval excludes zero; a draw means it covers zero; unresolved means no interval could be computed. Counting sources, Gaussian renormalization is ahead in fourteen, linear renormalization in one, one is drawn, three are unresolved and one is excluded. Of the three unresolved rows, the tone matrices are participant-averaged and digitized from figures, so there is no unit to resample; the news and Wikipedia clickstream inputs are large public archives that are not redistributed, though the scripts that would produce an interval from them are in the repository.

Counting by source gives the same picture as counting rows. Linear renormalization is ahead in exactly one source where the interval settles it, the unpublished twelve-line data, and it has the sign in a second, the tone matrices, where nothing settles it. Getty is drawn, its three conditions splitting in sign inside a single experiment. Wikipedia is unresolved, at a ten-thousandth of a nat.

3 Prior tests of the constant-ratio rule

Some experiments have a very simple setup. A stimulus is presented that is supposed to be matched to one of a fixed set of responses. Then the response list is reduced and the same experiment is repeated. Each row of the resulting confusion matrix is the response distribution for one stimulus, and the rule predicts the reduced matrix from the full one.

In the first of three experiments, Clarke [12] had listeners identify consonant-vowel syllables heard against noise, and obtained three six-by-six master matrices together with six three-by-three submatrices drawn from them. Clarke and Anderson [11] repeated the test on a ten-item set using naive rather than practised listeners, so that practice could be considered as a possible confound.

Pollack and Decker [13] used eight word-initial consonants in noise, with three overlapping four-way subsets. They used different signal-to-noise ratios. They reported average deviations near four per cent.

Hodge [20] is similar in spirit but used lifted weights and visual size. An original eight-item list was then cut to four, and then two. Rich [21] used a different kind of imperfection, our short-term memory. Getty, Swets, Swets and Green [19] kept all eight stimuli in play and cut only the permitted responses, to four at a time in three different sets. Here nothing about the stimulus was changed.

Morgan [22] built a likelihood-ratio test and rejected the constant-ratio rule on the earlier data.

Townsend and Landon [23] ran one master set of five letters with three nested response sets. They used a blocked and counterbalanced design. They published the full matrices per subject.

Rouder [24] finds that accuracy on a reduced choice set improves by more than a guessing correction predicts and by less than the ratio rule predicts, which is the direction of contraction found here.

Wills, Reimers, Stewart, Suret and McLaren [25] also reject the constant-ratio rule in categorization and conclude that it should be replaced by a system based on Thurstonian choice. Their route to this was via a fitted four-parameter winner-take-all model, and they also examine a parameter-free version that simply picks the largest noisy magnitude. They say that Gaussian noise “produces comparable results” to the rectangular noise they use.

Their setting differs from this one in what is available at the outset. They have a model

[5]The low-cost frame of the same experiment is a draw, $+.0002$ with an interval of $[-.0037, +.0180]$, which is the gradient discussed in Section 5.2.3.

[6]The interval excludes zero and the gain is negligible. Third-party candidates carry almost no mass, so the two maps agree to within a rounding error whatever the truth.

[7]Where a source has more than one file or item at the same design, every file or item is calibrated and scored on its own and the scores are pooled at the observation level. None is singled out, since picking one would invite choosing the file with the larger gain.

[8]Sports participation sits exactly on its shrinkage null, frequency 0.513, so its gain carries no evidence about the restriction map and it receives no verdict.

of the stimulus, so their magnitudes are a function of how many category-appropriate symbols a display contains, and a stimulus model of that kind motivates one noise law over another. Here there is no stimulus, only a probability vector, and the noise law has to be chosen. Their Experiment 2 withdraws a response category and their trial data is deposited, so it is scored below with the rest, where it agrees with them: Gaussian renormalization has the lower held-out loss by 0.0299 nats.

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. Townsend and Landon attribute that to Debreu's similarity objection, and Rouder brackets the same failure from the other side, finding real behaviour better than a guessing correction predicts.

Every claim of conformity that was later retested with a test that had power was overturned. Holloway's 1971 reanalysis found a dependency his earlier analysis had not [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 5.11: "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.

Beyond Clarke's comparison on percent correct, the literature contains no test of the constant-ratio rule against a second parameter-free assumption on the matrices themselves. Lee [38] 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 one hundred and thirty-three sources, recording which of them printed or deposited enough to score either assumption.

4 Theory

4.1 Why these two maps

Renormalization does not follow from probability theory. It is the correct posterior for mutually exclusive events only when learning that one of them is impossible says nothing about which of the rest obtained. Monty Hall is the familiar case where it does not: conditioning is still correct there, but the host's choice of door carries more information than the bare fact that the car is not behind it, so the event conditioned on is not simply $Y \in T$. Withdrawal is neither of those two cases, because no information about a fixed experiment has arrived at all. An alternative is gone, which is a different problem. Predicting the smaller menu 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) [39, 37], and Yellott [43] showed that the Gumbel is the only such noise law consistent with the axiom once menus of three or more are included, though pairwise comparisons alone do not single it out. So the constant-ratio rule is one point of a family.

One need not stand 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 renormalization 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 unconsidered default 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.[9]

Kempton and Burrell [34] quantify a consequence of per-step renormalization in language-model decoding: because top-k, nucleus and temperature sampling renormalize locally at every token, the relative probabilities of whole completions are distorted relative to the globally normalized distribution, in half of sampled pairs by a factor of several hundred.

Interpolating between them, or choosing the standardized shape of some third law, requires a number that the shares themselves do not supply. The overall scale of the noise is not such a number, because any common scale is absorbed exactly by the fitted locations and so changes no prediction. We call a quantity of that sort a *gauge*. Both maps 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.

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*

$$\frac{P_S(i)}{P_S(j)} \;\geq\; \frac{P_T(i)}{P_T(j)} \;\geq\; 1,$$

with the first inequality strict when $\log r$ *is strictly concave and* $S \setminus T$ *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

[9] 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.

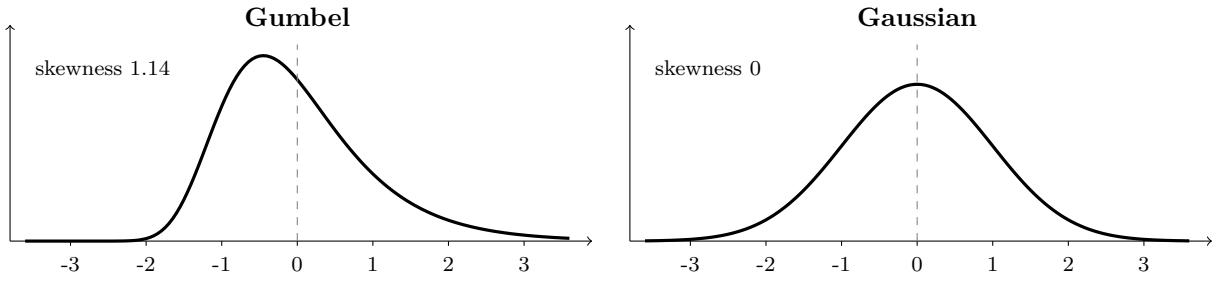


Figure 2: The two noise laws, each standardized to mean zero and variance one, so that only shape distinguishes them. Give every alternative a location, add a draw from the left-hand density to each, and take the highest: the winning probabilities are Equation (1), and restricting the field renormalizes 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 and long-tailed to the right. Neither shape is privileged *a priori*.

the items removed. Then $P_T(i) \propto \int AC$ and $P_S(i) \propto \int ACH$, and likewise for j with B in place of A. Now

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

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 $P_T(i)$ while lowering $P_T(j)$, and the two are equal at $a_i = a_j$ by exchangeability; hence $a_i > a_j$ gives $P_T(i) \geq P_T(j)$. □

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

The Gumbel sits on the boundary of the concavity condition and the Gaussian strictly inside it. 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 [39, 37]. 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.

A deep network is not committed to the axiom in the way that matters here. Any single positive vector over a fixed vocabulary is trivially in the independent additive random-utility class: Gumbel noise with $a_i = \log p_i$ generates it exactly, which is the same construction behind linear renormalization itself. Depth lets the vectors a network produces for different inputs differ in ways that no single context-independent noise law explains. That does not matter at inference time. Once a context is fixed, so is the logit vector, and masking and rescaling it 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 constant-ratio rule does not require recomputing scores with the menu as input. Gaussian renormalization escapes it with a fixed vector, calibrated once to the full vocabulary

and transported by a different noise law. Recomputing the scores is needed only to leave the independent additive random-utility class altogether, and so express dependence that no fixed-vector transport of any single noise law can produce, such as a menu-dependent covariance or context-dependent attention. A model trained to take the permitted set as input can express that; 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 where a score vector is computed once and then restricted *and* the restriction is an intervention on what may be chosen rather than news about a fixed experiment. Constrained decoding, vocabulary masking, restricted action spaces and post-hoc class removal are all candidates, but which of them qualifies depends on the intended estimand and not on the code. Where the target is the conditional law of the unrestricted model, masking is correct and nothing here applies.

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. We provide a plausibility argument applying only to aggregate shares.

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 $\beta > 0$,

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

which by the Holman–Marley construction [39, 37] 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. The proposition adds only the link between the high-heterogeneity limit of that familiar object and the restriction map tested here.

The Gaussian assumption on taste can itself be derived from a limit theorem.

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 therefore 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 not innocuous. 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.

No individual need be non-Lucean for the aggregate to approach the Gaussian transport. 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. At any finite ratio the exact law is Gaussian plus Gumbel, not Gaussian, so Case V is the limit and not the aggregate itself. That mechanism also says where the fit should be worst: in populations whose members are nearly alike. It explains movement from linear renormalization toward Case V and no further. Proposition 2 establishes only the limit, $r \to 1$ as $\sigma/\beta \to \infty$; Table 3 shows the path reaching that limit from below at every σ tested but 16, where it reads 1.01, so the approach is not shown to be monotone and should not be read as bounding r by 1 in general. Either way the mechanism does not account for the median r of 1.13 in Figure 1, nor for the fifteen of twenty populations that contract past Case V. That residual is unexplained here. Dependence, non-isotropic heterogeneity, menu-dependent attention and some third standardized shape are all candidates.

σ	Top share	Observed λ	Case V λ	Ratio
0.5	0.255	0.052	0.224	0.23
1	0.300	0.128	0.225	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	0.230	1.00
16	0.375	0.233	0.231	1.01

Table 3: 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 $\delta_{ij} = \operatorname{logit} q_{ij} - \log(p_i/p_j)$ 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 contest 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`.

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, showing up as a parameter that has to move with the size of the menu.

Their design is the nearest to this one that we have found. Three differences remain. 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. Their result shows that one parameter has to move with m and the other does not, so contraction is present, without measuring how large it is.

4.3 Correlated probit and the unidentified covariances

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 5.11 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 5.11 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 developments make the comparison 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, the *excess* of the observed gain over that null median, and a null frequency. 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 5.10 details the construction.

5.2 Restrictions that were observed

Seven sources in this section 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. Of the other four, the gambles are a designed comparison across groups, and the ballots, news slates and Wikipedia links are ecological; Section 5.2.3 says where that weakens them.

5.2.1 Consumer goods, all subsets of five

In the ranking-and-menu benchmark of Tables 4 and 7, 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.

	n	Choices	Linear	Gaussian	Gain	Excess
Experiment 1, all	230	5,134	0.9438	0.9298	+0.0140	+0.0158
Experiment 1, forced	76	1,900	0.9639	0.9196	+0.0443	+0.0467
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 4: 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 5.10. Subject-bootstrap intervals for the gains are $[+0.0072, +0.0335]$, $[+0.0171, +0.0868]$, $[+0.0009, +0.0307]$, $[+0.0034, +0.0560]$ and $[+0.0048, +0.0246]$ respectively.

Gaussian renormalization 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 null frequencies of 0.010 and 0.005; experiment 2, with 134 subjects against 230, does not, at 0.194 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

In the consumer experiment the subject sees the reduced menu before deciding, so preferences could in principle be formed with the restriction already known. The two perceptual experiments below do not allow that.

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 is told only *after* it has gone which two named alternatives the answer lies between. The two conditions were run in separate blocks of trials, so the observer knew which condition they were in and had practised it. What they could not know, while looking at the display, was which two colours the pair would turn out to be. They could not attend selectively to those two colours in advance, though they did know that only two would be offered. 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.

The six-symbol gain is the larger of the two arms. The four-colour 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 falls by about three quarters, from +0.0147 to +0.0038, though its interval, $[+0.0006, +0.0072]$, still excludes zero. Accuracy rises from 0.780 to 0.850, linear renormalization becomes close to exactly right, and Gaussian renormalization begins to over-contract.

So the advantage is not a property of these pairs, these colours or these observers. It is large

Arm	Cells	Linear	Gaussian	Gain	Excess
Four colours, menu revealed after	384	0.4942	0.4795	+0.0147	+0.0156
Six symbols, menu revealed after	300	0.6276	0.5999	+0.0277	+0.0313
Same pairs, announced in advance	384	0.3967	0.3928	+0.0038	+0.0047

Table 5: Mean nats per cell, a cell being one observer, one dominant item and one surviving alternative. Subject-bootstrap intervals for the gains are $[+0.0117, +0.0178]$, $[+0.0233, +0.0324]$ and $[+0.0006, +0.0072]$. The null here resamples the calibration row as well as the pair counts, so it charges Gaussian renormalization for calibration noise as well as for sampling noise in what it predicts; all three null frequencies are at the 0.005 floor.

when the surviving pair is selected only after the percept has formed, so that attention cannot be tailored to that pair, and it is small, though still present, when the pair is named in advance.

In both genuine restriction arms *both* maps over-predict the favourite: linear renormalization by 0.069 and Gaussian renormalization 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 map is fitted to restricted-menu data 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. Gaussian renormalization wins by $+0.0037$ per cell over 360 nested cells, interval $[+0.0027, +0.0049]$. 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 5.11: 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 each person at most two disjoint menus within each condition. It is a hard case by the authors' own analysis, since they report that random utility cannot explain the population behaviour while a logit attention model survives.

The same participants completed all three conditions, which differ 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 the gradient as consistent with the taste-heterogeneity mechanism of Section 4.2 rather than as instability, though increased random responding, attention failure or a change

in what subjects consider would produce the same ordering. Because the same people supply all three conditions, the comparison is within subject: whatever tracks arithmetic cost is not a difference in who was recruited. The lotteries were designed so that different risk preferences justify different choices at every cost level, so no condition, including the one-operation frame, has a single correct answer to converge on. Cost changes how much calculation separates a stated preference from the choice made under it. At one operation the calculation is close to immediate, and shares concentrate on whichever lottery it favours. At five operations the calculation is no longer one most people complete unaided, so more of what is observed is noise around it. A model of preference under noise should do more work where that noise is larger, and here it does.

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 [42]. 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 several 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	Null freq.
Lotteries, high cost	3,928	+0.0101	+0.0020	+0.0081	.035
Lotteries, medium cost	3,928	+0.0069	+0.0023	+0.0045	.124
Lotteries, low cost	3,928	+0.0002	+0.0014	−0.0012	.672
News slates	6,000 pairs	+0.0496	−0.0088	+0.0584	.016
US state ballots 2024	263 pairs	+0.0001	−0.0000	+0.0001	.005
Wikipedia links	3,544 pages	−0.0001	−0.0002	+0.0001	.005

Table 6: 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 rows are the same participants under three cost frames; 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 in this table, about five times the next, on the only ecological 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 ballot and Wikipedia rows 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 Sections 5.2.1 and 5.2.2 for that reason.

The lottery gradient runs with arithmetic cost, and the mechanism of Section 4.2 predicts its direction if computation error is per-alternative and roughly Gaussian, since loading more of it onto each alternative raises σ against β.

On the Wikipedia rows, 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.

Five of the twenty populations in Figure 1 put the truth between the two maps, contracting when an alternative goes but by less than Case V says: occupational prestige, the two General Social Survey collections, the high-cost gambles and the line data. They are the exception. In the other fifteen the observed contraction slope lies *above* Case V's, so 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.

5.4 Smoothing and menu weighting

The pipeline makes two choices that the maps themselves do not. Shares are smoothed by add-α with $\alpha = 1/2$, and the aggregate weights every subset of size two or more equally. Both are conventions rather than results, so both are varied.

Dataset	n	K	Subsets	Linear	Gaussian	Gain
Occupational prestige	143	10	1,012	1.0170	0.9780	+0.0390 [+0.0200, +0.0759]
Sushi	5,000	10	1,012	1.3719	1.3607	+0.0111 [+0.0088, +0.0142]
Political goals	2,262	4	10	0.8431	0.8355	+0.0076 [+0.0060, +0.0096]
GSS socialization	5,000	5	25	0.7786	0.7729	+0.0057 [+0.0027, +0.0104]
Jester file 3	5,000	10	1,012	1.5182	1.5158	+0.0024 [+0.0013, +0.0031]
Jester file 1	5,000	10	1,012	1.5280	1.5260	+0.0020 [+0.0013, +0.0031]
Jester file 2	5,000	10	1,012	1.5240	1.5223	+0.0018 [+0.0014, +0.0032]
GSS job values	5,000	5	25	0.8407	0.8389	+0.0018 [+0.0013, +0.0049]
Dots	3,183	4	40	0.7812	0.7797	+0.0014 [+0.0008, +0.0030]
Netflix	4,379	3	15	0.6030	0.6012	+0.0018 [+0.0006, +0.0037]
Puzzles	3,180	4	40	0.7726	0.7718	+0.0009 [+0.0005, +0.0024]
Sports participation	130	7	119	1.2519	1.2469	+0.0050 [+0.0027, +0.0167]

Table 7: Held-out log loss per prediction over every proper restriction, $2 \leq |T| < K$. Datasets are capped at five thousand respondents for runtime. Above the line, Gaussian renormalization is lower and the interval excludes zero in every one of the eleven collections. Occupational prestige is scorable here only because the shared smoothing removes a zero training cell. Sports participation, below the line, also has an interval excluding zero, but Section 5.10 shows its gain is entirely shrinkage: the exclusion is not evidence against the constant-ratio rule for this collection, and it receives no verdict. Intervals resample respondents and refit the whole pipeline inside every replicate, calibration and both maps and all subsets, so they carry calibration uncertainty; two thousand replicates for every collection.

	smoothing			menu weighting		
Collection	$\alpha = 0$	$\alpha = \frac{1}{2}$	$\alpha = 1$	subsets	by size	pairs
Occupational prestige	n/a	+.0390	+.0153	+.0390	+.0299	+.0286
Sushi	+.0113	+.0111	+.0109	+.0111	+.0121	+.0412
Political goals	+.0076	+.0076	+.0075	+.0076	+.0075	+.0079
GSS socialization	+.0059	+.0057	+.0055	+.0057	+.0048	+.0130
Sports participation	+.0056	+.0050	+.0046	+.0050	+.0045	+.0058
Jester file 3	+.0024	+.0024	+.0024	+.0024	+.0024	+.0064
Jester file 1	+.0020	+.0020	+.0019	+.0020	+.0019	+.0053
Jester file 2	+.0018	+.0018	+.0017	+.0018	+.0017	+.0048
GSS job values	+.0018	+.0018	+.0018	+.0018	+.0016	+.0021
Netflix	+.0019	+.0018	+.0016	+.0018	+.0018	+.0018
Dots	+.0015	+.0014	+.0014	+.0014	+.0012	+.0023
Puzzles	+.0010	+.0009	+.0008	+.0009	+.0007	+.0015

Table 8: Held-out gain under three smoothing conventions and three menu weightings. Smoothing moves six collections by 0.0001 or less across the three values and the others by at most 0.0010, except occupational prestige, which is unscorable at $\alpha = 0$ and falls to +0.0153 at $\alpha = 1$. Weighting changes magnitude and not sign. Uniform weighting by surviving size is smaller than the subset count in nine collections, equal to four decimals on Jester file 3 and on Netflix, where $K = 3$ leaves only pairs, and larger on Sushi (+0.0121 against +0.0111). Pairs alone are larger everywhere except occupational prestige, where they are smaller, and Netflix; Sushi runs from +0.0111 to +0.0412 depending which is chosen. Every cell that can be computed is positive.

The equal-subset weighting reported throughout is neither the smallest nor the largest of the three. Counting every subset of size two or more once puts most of the weight near $|T| = K/2$, and pairs are where the two maps differ most, so the reported aggregate usually sits between the size-weighted and pairs-only figures. It is the smallest of the three only on Sushi, and the largest on occupational prestige. No weighting changes any sign.

5.5 The estimand, the weighting and the score

Three features of the score are choices rather than results. Each is varied here.

The first is which menus are scored. The estimand throughout is proper restrictions, $2 \leq |T| < K$. The full menu is excluded because it is not a restriction: both maps are calibrated on those shares and both return them, so it contributes an identical zero. Its weight is one subset in $2^K - K - 1$, which is one in 1,013 at $K = 10$ and one in four at $K = 3$. Including it would therefore shrink small collections and leave large ones alone. That changes the estimand in a K-dependent way and distorts the cross-collection comparison this paper makes. Including it moves Netflix from +0.0018 to +0.0013 and political goals from +0.0076 to +0.0069, leaves everything with $K \geq 7$ unchanged to four decimals, and changes no sign.

The second choice is the weighting across menu sizes, and the three weightings of Table 8 are three summaries of one curve. Table 9 gives the curve itself.

Dataset	$\|T\| = 2$	3	4	5	6	7	8	9	10
Occupational prestige	+.0286	+.0516	+.0549	+.0454	+.0313	+.0178	+.0076	+.0016	+.0000
Sushi	+.0412	+.0263	+.0150	+.0081	+.0040	+.0017	+.0005	+.0000	−.0000
Political goals	+.0079	+.0070	+.0000						
GSS socialization	+.0130	+.0011	+.0004	−.0000					
Sports participation	+.0058	+.0061	+.0051	+.0035	+.0018	+.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
Jester file 2	+.0048	+.0037	+.0024	+.0015	+.0009	+.0005	+.0002	+.0001	+.0000
GSS job values	+.0021	+.0020	+.0008	−.0000					
Dots	+.0023	+.0002	−.0000						
Netflix	+.0018	−.0000							
Puzzles	+.0015	−.0000	−.0000						

Table 9: Held-out gain by the size of the surviving menu, all twelve ranking collections. The last entry in each row is zero to the precision of the pipeline, since nothing has been removed. Ten collections fall monotonically in $|T|$. Occupational prestige and sports participation do not: prestige peaks at $|T| = 4$ with +0.0549, roughly twice its pairwise figure, which is why pairs-only weighting understates it while overstating every other collection except Netflix. Reproduced by `python estimand.py`.

The third is the score itself. Log loss is unbounded above, so a single confident miss can outweigh many small wins, and Table 10 shows the advantage assembled from rare large gains against frequent small losses, which is the shape log loss rewards most. The multiclass Brier score $\sum_j (q_j - y_j)^2$ is bounded and proper, and applied to the same subsets under the same folds it agrees in sign on all twelve collections. Its interval covers zero on the three collections with the smallest log gains: Netflix at +0.00049 $[-0.00024, +0.00122]$, GSS job values at +0.00069 $[-0.00005, +0.00146]$ and puzzles at +0.00033 $[-0.00049, +0.00114]$. The largest Brier gains are occupational prestige at +0.0104, political goals at +0.0060 and Sushi at +0.0046. So the

ordering of collections is preserved and the advantage is not an artifact of an unbounded score, though on the three weakest collections a bounded score cannot separate the maps.

5.6 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.

Its calibration is clean. Calibration row and target row come from the *same subject* in different blocks, as in the colour and symbol arms of Section 5.2.2, 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 in its between-person form 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 Gaussian renormalization 0.9412, a gain of $+0.0042$ with a row bootstrap of $[+0.0022, +0.0064]$, against a fitted-Luce null median of -0.0010, so the excess is $+0.0052$, null frequency 0.005. The subsets give $+0.0006$ $[-0.0002, +0.0015]$ for $\{A, E, F, H\}$, $+0.0040$ $[-0.0003, +0.0078]$ for $\{A, E, X\}$ and $+0.0092$ $[+0.0051, +0.0141]$ 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 9 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 $+0.0092$. Those designs are structurally symmetric, so the difference between them is not attributable to similarity.

The row-level output shows 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 0.163/0.199/0.216 on X. Gaussian renormalization moves the favourite down and the near-substitute up. Both corrections have the right sign, but the near-substitute moves far too little, and the dissimilar letter moves 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 [38] proposed this comparison analytically in 1968; what is added here is the out-of-sample test.

5.7 A category withdrawn between groups

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 Gaussian renormalization 0.6541, a gain of $+0.0299$ against a null median of -0.0015, so the excess is $+0.0314$, null frequency 0.002. Graded and dummy stimuli agree, $+0.0286$ against $+0.0327$.

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 $+0.0484$. The participant bootstrap is wide and right-skewed, $[+0.0175, +0.0716]$.

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.

5.8 How the difference is distributed

5.8.1 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 Sushi and the three Jester files, shows what the average is made of.

Rank	1	2	3	4	5	6	7	8	9	10
Share of outcomes	.173	.130	.127	.104	.099	.101	.079	.072	.072	.044
Mean gain	−.061	−.045	−.034	−.004	+.009	+.019	+.050	+.058	+.100	+.171

Table 10: 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 two rows are pooled by summing outcome counts and gain contributions across the four collections, so they dot to the aggregate of +0.0043, which the script asserts.

Gaussian renormalization pays 0.061 nats whenever the favourite wins, which happens 17 per cent of the time, and collects 0.171 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. That accounts for three results above. 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.

5.9 The exact ordering laws

Ordering is a different question from restriction, and the two should not be confused: removing the observed winner conditions on the outcome of a fixed contest rather than withdrawing an alternative from it, and learning who won is informative about the latent draws. We ask it anyway, because here both models make an exact prediction and the truth is observed, so the comparison is against fact rather than against each other.

For the Gumbel model the sequential rule is not an approximation. Sorting $\log u_i + G_i$ with independent Gumbel G_i yields a Plackett–Luce ranking, in which the next item is drawn from the survivors in proportion to their remaining worths, so removing the winner and renormalizing proportionally *is* that model's ordering law. This is the Gumbel top-k identity, and it is the same construction the null of Section 5.10 uses to draw rankings. Harville's formula, which applies this rule to win probabilities, is therefore exact for Luce and not a heuristic.

The Gaussian side has no such sequential identity, so its law is computed directly. Conditioning on i's draw at x, item i is second when exactly one rival exceeds x and the rest fall below, giving another one-dimensional integral,

$$P(\text{rank}(i) = 2) = \int \varphi(x - a_i) \sum_{j \neq i} [1 - \Phi(x - a_j)] \prod_{k \neq i,j} \Phi(x - a_k)\, dx. \tag{6}$$

Collection	Observed	Plackett–Luce	Case V	PL error	Case V error
Political goals	0.123	0.318	0.284	+0.195	+0.161
GSS job values	0.186	0.303	0.252	+0.116	+0.066
GSS socialization	0.191	0.302	0.257	+0.110	+0.065
Sushi	0.206	0.250	0.196	+0.045	−0.009
Jester file 1	0.138	0.160	0.142	+0.022	+0.004

Table 11: Probability that the full-field favourite finishes second, against each model's own exact ordering law: Plackett–Luce by the Gumbel top-k identity, Case V by Equation (6). Both are verified against direct simulation, four million Gumbel races and three million Gaussian ones, agreeing to within Monte Carlo error. Plackett–Luce overstates the probability in all five collections. Case V removes between 18 and 121 per cent of that error, overshooting slightly on Sushi. The sequential Gaussian rule, which is *not* the Gaussian ordering law, removes only 5 to 12 per cent and is reported in `results/decompositions.txt` for comparison. Reproduced by `python decompositions.py`.

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. For a practitioner it matters more than the average gains: the error is large, replicable and signed, and Harville's rule is exactly what the Luce model prescribes rather than a shortcut somebody chose. Case V removes most of it on Jester file 1, 82 per cent, and around two fifths on GSS job values and socialization, 43 and 41 per cent; on Sushi it removes all of it and is the wrong sign there by 0.009. What produces the remainder on political goals, where 0.161 of the 0.195 survives, 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.

5.10 The shrinkage null

A positive gain has two possible sources and log loss cannot tell them apart. Gaussian renormalization may be closer to how the population chooses. Or it 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 receives no verdict. Everywhere else the mechanism runs against Gaussian renormalization, and the excess over the null exceeds the raw gain in every remaining collection. The menu experiment, the one collection in this table that observes menu choice directly, clears the null for both the forced-choice and the all-subject samples.

Dataset	Observed	Null median	Excess	Null freq.
Menu experiment, forced choice	+0.0266	−0.0041	+0.0307	0.010
Menu experiment, all subjects	+0.0100	−0.0023	+0.0123	0.015
Occupational prestige	+0.0390	−0.0220	+0.0611	0.0005
Sushi	+0.0111	−0.0065	+0.0176	0.0005
GSS socialization	+0.0057	−0.0069	+0.0126	0.0005
Political goals	+0.0076	−0.0013	+0.0089	0.0005
GSS job values	+0.0018	−0.0049	+0.0067	0.0005
Jester file 3	+0.0024	−0.0015	+0.0039	0.0005
Jester file 1	+0.0020	−0.0014	+0.0033	0.0005
Jester file 2	+0.0018	−0.0014	+0.0032	0.0005
Dots	+0.0014	−0.0012	+0.0026	0.0005
Puzzles	+0.0009	−0.0010	+0.0019	0.0015
Netflix	+0.0018	−0.0016	+0.0033	0.0005
Sports participation	+0.0050	+0.0051	−0.0001	0.513

Table 12: The same gain when the axiom holds by construction. Two thousand replicates per ranking row and two hundred for the menu rows. The frequency is $(b+1)/(B+1)$ for b exceedances of the observed gain, so 0.0005 is the floor 1/2001 reached when no ranking replicate matched it. The diagnostic is the excess of the observed gain over the null median; the null probabilities are indicative rather than exact tests, and for the menu rows they are optimistic, because the null redraws each subject's choices independently given the aggregate worths and so omits the within-subject dependence that the bootstrap intervals of Table 4 do preserve. Every null median except sports participation's is negative, because contraction of well-estimated shares is a bias rather than useful insurance, so comparing an observed gain with zero understates it.

5.11 Boundary conditions

Eight of the nine negative rows in Table 1 belong to three closely related collections: four tone conditions, three line restrictions, and the Getty condition whose survivors are mutual confusions. Wikipedia is the near-zero ninth. Three further populations fail both maps in ways that published aggregates settle without this protocol.

Those eight are fewer independent observations than the row count suggests. The obvious reading of what they have in common is that the alternatives lie on a perceptual continuum. Section 5.12 shows that reading failing for two separate reasons, each of which can be demonstrated with three alternatives.

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 renormalization wins all four conditions, by 0.0132 and 0.0051 nats at narrow spacing and 0.0172 and 0.0041 at wide. One row of the wide ten-to-eight condition is excluded: it contains an exact zero, and the location after flooring is unstable enough to move that condition's gain. The other fifteen ten-to-eight rows agree between two independently written calibrators 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 Gaussian renormalization fails here too, and only slightly less. Neither map can produce a reversal.

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. Boissonnet and Ghersengorin [32] axiomatize a choice model identified from reversals of this kind, and it is the theoretical account this row calls for.

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 Gaussian renormalization wins by $+0.0272$ nats, and by $+0.0110$ on the rows whose stimulus survives in the menu, cell bootstrap $[+0.0021, +0.0224]$. It loses one of the three conditions, and not at random.

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

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	-0.0128

Table 13: 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 Gaussian renormalization loses is the one whose survivors are each other's confusions.

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 $\{0, 1, 2, 3\}$, $\{4, 5, 6, 7\}$, $\{8, 9, 10, 11\}$, $\{0 \dots 7\}$ and $\{4 \dots 11\}$; 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 $[-0.0176, -0.0103]$, and by size of the surviving menu it is -0.0056 at eight, -0.0089 at four and -0.0322 at two, so the loss grows as more of the continuum is cut away. The twelve-to-two restriction is the largest single loss for Gaussian renormalization in the paper.

The dependence diagnostic of Section 5.13 predicted this loss. Line length is the unidimensional continuum of the absolute-identification literature, and the diagnostic was written down 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 dependence diagnostic 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.

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 0.9078 against a renormalized 0.9091, a difference of 0.0013. Removing the 0.90 favourite instead gives the surviving 0.09 alternative 0.8027 against a renormalized 0.9000, a difference of 0.0973. 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 line continuum and the high-cost gambles; the collections furthest past Gaussian renormalization are categories, colours, symbols, political goals and consumer goods.

5.12 Two false explanations of the failures

A continuum is not the culprit. Let the stimulus fix a scalar percept s, and give response label i, positioned at x_i on the same physical dimension, the utility $U_i = -c\,(s - x_i)^2 + \varepsilon_i$ with ε_i independent standard normal. Neighbouring labels are more confusable than distant ones and the confusion matrix is strongly banded. Conditional on s this is exactly Case V with locations $a_i = -c\,(s - x_i)^2$, so withdrawing labels and re-running the contest is exactly right, and linear renormalization is wrong. With $c = 1/2$ and labels at $-2, \ldots, 2$, removing the second and fourth leaves a true favourite share of 0.866 at $s = 0$; Gaussian renormalization returns it exactly and linear renormalization returns 0.924. Replace the Gaussian errors with independent Gumbels and the verdict reverses exactly, with the same one-dimensional geometry and the same banding. Dimension does not identify the noise law. If instead s varies from trial to trial, every label shares that variation and the contest is no longer independent; that is the dependence of Section 5.13.

Near-substitution does not fix the sign. Take full-menu shares $p = (0.50, 0.30, 0.20)$ and let the last fifth of the population rank $C \succ B \succ A$, so that C's mass is a perfect substitute for B. Removing C gives a true pair share of 0.50 for A. Linear renormalization predicts 0.625 and Gaussian renormalization 0.616, which by Proposition 1 must lie below it, so Gaussian renormalization has the lower loss, 0.7210 against 0.7254. Now change only that fifth to $C \succ A \succ B$. The full-menu shares are unchanged and the perfect substitute is still there, but the truth becomes 0.70 and linear renormalization is the better of the two, 0.6233 against 0.6261. The

same share vector and the same statement that a removed alternative has a near-substitute admit opposite answers. The deciding factor is where the displaced mass goes: toward a lower-share survivor the odds contract and Gaussian renormalization gains, toward the favourite they expand and it loses.

5.13 Three diagnostics

So the pattern in the failures is not a rule about what kind of thing the alternatives are. The failures do suggest three heuristic checks on the assumptions, of different logical types. Only the line-length collection was scored after they were written down.

Dependence. A shared percept, local substitution, clustered tastes or a common shock makes the covariance materially non-isotropic on the contrast subspace, and independent Case V is then structurally inadequate. Continua and near-substitutes are possible symptoms of this rather than rules in themselves, and the tone, line and Getty collections are read here as its instances: what they share is a single noisy percept moving neighbouring labels together, not their dimension. Inadequacy of Case V does not make proportional renormalization correct, and in the Scottish verdicts neither map survives.

Transport failure. If the permitted set changes locations, attention, criteria or covariance, then a_T depends on T and no fixed-latent map applies, whichever it is. The exam distractors, the phoneme labels and the Getty retuning are of this kind, and the recommendation there is to re-estimate rather than to transport.

Low leverage. If the withdrawn alternatives carry negligible mass and the leader survives, the two forecasts are practically indistinguishable. This settles what the choice is worth, not which map is right.

Meyer-Grant and colleagues [36] distinguish a Gumbel-min construction, in which the observer selects the item believed *new*, from the ordinary Gumbel-max construction this paper uses throughout, in which the highest draw is chosen. These are mirrored models with different predictions, not two names for the same one. They report accuracy invariance as the recognition set grows under the Gumbel-min task, which is Yellott's condition independently confirmed, and by their own account it does not carry over unchanged to a maximum-selection task. The foil experiment of Section 5.2.2 is a maximum-selection task, and Gaussian renormalization wins on it. We have not compared it with their maximum-selection condition. A Gumbel-min model fitted to a minimum-selection task is a different restriction model from linear renormalization, so their result and ours are not interchangeable evidence for or against the Gumbel family.

6 Scope

The practical recommendation is narrow, and Section 5.11 makes it conditional. When a probability vector is transported to a smaller support, Gaussian renormalization should be tried alongside linear renormalization rather than passed over.

Four questions bear on whether the independence the Gaussian contest assumes is tenable. The first two split the dependence check of Section 5.13 in two, and the third and fourth are its transport-failure and low-leverage checks. They are heuristics, neither necessary nor sufficient for either map to win, formed partly from the collections they describe, and Section 5.12 constructs populations where the first two point the wrong way. None is settled by the share vector alone. Distinct unordered items, such as goods, articles, jokes or occupations, usually answer no to all

four, which is a rule of thumb about where the two maps have separated in our data rather than a statement about where they must.

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 warns that independent Case V may be misspecified, and no more. It does not identify proportional renormalization as the better of the two, and a no is not a guarantee of anything either. Section 5.12 exhibits a one-dimensional continuum on which Gaussian renormalization is exactly right, and a pair of populations with identical full-menu shares and a perfect near-substitute on which the two maps win one apiece, the sign turning on whether the displaced mass moves toward a lower-share survivor or toward the favourite. A yes to the third says neither transport map applies, since the surviving alternatives are not the ones that were there before, and the model has to be re-estimated. A yes to the fourth says the choice is not worth making.

A confusion matrix bears on the first two, which is the diagnostic Townsend and Landon [23] used, and the within-set confusion mass of Table 13 makes it numerical. It does not settle the first outright: Section 5.12 constructs an independent Case V model, with ordinary per-stimulus locations and no covariance at all, whose confusion matrix bands exactly like a dependent one. Banding is consistent with dependence and does not by itself identify it. The matrix does supply 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.

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. Evaluating the winning probabilities for a fixed location vector, the forward direction, is one pass over a shared grid, following the lattice algorithm of [2]; inverting from shares to locations solves for that vector by repeated forward passes, immediate at the field sizes here, twelve alternatives at most. The restricted predictions are then one-dimensional quadrature on the same grid. The method is designed for and tested at much larger fields [2].

The recommendation needs no additional restricted-menu observations and no appreciable additional computation, although the boundary conditions of Section 5.11 require structural information about similarity, ordering or task geometry that the shares do not contain. Away from those conditions Gaussian renormalization had the lower loss on every row examined here. At them 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. The tables reject a representative-agent map.

The experiment compares two restriction maps, not two fitted likelihoods. A fitted Plackett–Luce model [40] 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 14 degrees of freedom, 17 with the calibration menu, 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 12 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 11 is a second such check, on a different question. Both models have an exact ordering law there, Plackett–Luce by the Gumbel top-k identity and Case V by Equation (6), and the observed frequency is available, so the two are compared against fact. Case V is the closer of the two in all five collections, by margins from 0.018 to 0.050, having overshot on one. Ordering is not the restriction question and the agreement of the two exercises is not automatic. Wherever a restricted-menu observation exists, a fitted shape parameter should improve on both maps. The case made here is for the situation where none does.

7 Conclusion

We have set the Thurstone Case V $N(0, 1)$ model, which we call *Gaussian renormalization*, against the ubiquitous linear rescaling of probability. The question is not confined to menus of goods; it arises wherever a probability vector is moved to a smaller support, which includes masked softmax vocabularies, scratched fields, delisted catalogues and disallowed response sets.

Linear renormalization is a theorem when the support shrinks by conditioning and a modelling choice when the feasible set is intervened on; only the second case is at issue. A mask often encodes an intervention while performing the arithmetic of conditioning.

Three things are established by the tables. Gaussian renormalization has the lower held-out log loss in thirty of the thirty-nine comparisons. Counting by source instead, which is the right unit because the rows cluster, it is ahead in fourteen of nineteen, with one going the other way,

one drawn and three unresolved. On the contraction axis of Figure 1 it is the nearer of the two maps in nineteen of the twenty populations that axis can place. Neither count estimates a rate, since the corpus was assembled by searching for restriction data. Linear renormalization is ahead where an interval settles it in exactly one source, the unpublished twelve-line data, and it carries the sign in the tone matrices, where nothing settles it.

Where the restricted probability is itself observed, which is the after-stimulus arms of Section 5.2.2, both maps over-predict the favourite and Gaussian renormalization removes between a fifth and a third of the error. That is the ceiling on what either default can claim, and it is well below what a fitted shape parameter should reach wherever a restricted-menu observation exists to fit one.

Section 5.11 lists three conditions under which Gaussian renormalization should not be expected to help: the alternatives are dependent, the task changes when an option is withdrawn, or too little mass is withdrawn for the choice to matter. They are heuristics, and none of them can be determined from a share vector alone. A confusion matrix bears on the first when one is available, though banding alone does not identify dependence, since an independent model with the right locations produces it too (Section 5.12); otherwise the judgement is made from the alternatives themselves and can be wrong.

Table 11 compares the exact ordering laws on a single event, the favourite finishing second. A comparison over complete rankings is left to future work.

Theory does not decide between the two maps. Both are zero-parameter members of one family, and nothing about a masked softmax expresses a belief that choice errors are double exponential. The evidence assembled here does not install a new default either. It supports computing Gaussian renormalization alongside linear renormalization. It costs one inversion, it fits nothing, and in thirty of thirty-nine comparisons it forecast the withheld choices better than the reflex everybody already uses.

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 and is therefore cited by repository rather than by author. All are public.

That last deposit is not redistributed, because it carries no licence. What the repository holds is the derived confusion counts, one array of per-block stimulus-by-response tallies, which is what the analysis reads and is a measurement of the experiment rather than a copy of the deposit. `python derive_rouder.py -fetch` clones the original, and `python derive_rouder.py` rebuilds the counts from it, reproducing every figure reported here exactly. `MANIFEST.tsv` records the

upstream git blob hash of every file the reduction was built from.

Analysis code and every input needed to reproduce the tables are under `research/restriction` in the `microprediction/winning` repository, 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` recording the 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, ConsRank 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-checking: 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.

We assembled a census of one hundred and thirty-three 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 14 lists the eight 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	$6 \to 3$, two more	library access
Pollack and Decker [13]	consonants	$8 \to 4$, three sets	not scoreable[10]
Hodge and Pollack [44]	auditory, unidimensional	nested	library access
Lacouture et al. [45]	line lengths	nested	library access
Kent and Lamberts [46]	dot separation	$N = 2, 4, 6, 8, 10$	library access
Negoias et al. [47]	odor descriptors	$16 \to 8 \to 4$	never published
Jaeger et al. [48]	flavour terms	nested, $n = 735$	library access
Hörberg et al. [49]	odor naming	free, then cued list	data request

Table 14: Restriction experiments not scored above. "Library access" means the matrices appear in the article and need transcription, as Townsend and Landon's were for Section 5.6. The unpublished twelve-line laboratory data found by the same search is scored in Section 5.11 and so does not appear here. 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. Five need only library access to a printed table rather than a data

[10] 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.

request, which is the cheapest way anyone could extend this comparison.

References

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

[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. Proportion of correct responses as a function of the number of stimulus-response alternatives. *Journal of the Acoustical Society of America*, 31(6, Suppl.):835, 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] C. M. Holloway. Consonant recognition with two levels of decision complexity. *Quarterly Journal of Experimental Psychology*, 22(3):467–474, 1970.

[17] Charles M. Holloway. A test of the independence of linguistic dimensions. *Language and Speech*, 14(4):326–340, 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, Mark 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] Niels Boissonnet and Alexis Ghersengorin. Grabbing the forbidden fruit: restriction-sensitive choice. arXiv:2509.11673, 2025.

[33] Wanqiao Xu, Shi Dong, Xiuyuan Lu, Grace Lam, Zheng Wen, and Benjamin Van Roy. RLHF and IIA: perverse incentives. arXiv:2312.01057, 2023.

[34] Tom Kempton and Stuart Burrell. Local normalization distortion and the thermodynamic formalism of decoding strategies for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22216–22231, 2025.

[35] Hugo Sales Corrêa, Suryanarayana Sankagiri, Daniel Ratton Figueiredo, and Matthias Grossglauser. Measuring IIA violations in similarity choices with Bayesian models. arXiv:2508.14615, 2025.

[36] 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.

[37] 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.

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

[39] 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 [43] cites at pp. 110 and 123, rather than [37].

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

[41] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

[42] 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.

[43] 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(2):109–144, 1977.

[44] Milton H. Hodge and Irwin Pollack. Confusion matrix analysis of single and multidimensional auditory displays. *Journal of Experimental Psychology*, 63(2):129–142, 1962.

[45] Yves Lacouture, Shu-Chen Li, and A. A. J. Marley. The roles of stimulus and response set size in the identification and categorisation of unidimensional stimuli. *Australian Journal of Psychology*, 50(3):165–174, 1998.

[46] Christopher Kent and Koen Lamberts. An exemplar account of the bow and set-size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):289–305, 2005.

[47] S. Negoias, C. Troeger, P. Rombaux, S. Halewyck, and T. Hummel. Number of descriptors in cued odor identification tests. *Archives of Otolaryngology–Head & Neck Surgery*, 136(3):296–300, 2010.

[48] Sara R. Jaeger, Michelle K. Beresford, Amy G. Paisley, Lucía Antúnez, Leticia Vidal, Rafael Silva Cadena, Ana Giménez, and Gastón Ares. Check-all-that-apply (CATA) questions for sensory product characterization by consumers: investigations into the number of terms used in CATA questions. *Food Quality and Preference*, 42:154–164, 2015.

[49] Thomas Hörberg, Jonas K. Olofsson, Rohan Raj, Erika J. Laukka, and Maria Larsson. Free odor identification engages domain-general cognitive abilities in old adults. *Chemical Senses*, 50:bjaf049, 2025.