On the number of trials needed to distinguish similar alternatives

Flavio Chierichetti[a,1,2], Ravi Kumar[b,2], and Andrew Tomkins[b,2]

Edited by William Press, University of Texas at Austin, Austin, TX; received February 4, 2022; accepted June 17, 2022

A/B testing is widely used to tune search and recommendation algorithms, to compare product variants as efficiently and effectively as possible, and even to study animal behavior. With ongoing investment, due to diminishing returns, the items produced by the new alternative B show smaller and smaller improvement in quality from the items produced by the current system A. By formalizing this observation, we develop closed-form analytical expressions for the sample efficiency of a number of widely used families of slate-based comparison tests. In empirical trials, these theoretical sample complexity results are shown to be predictive of real-world testing efficiency outcomes. These findings offer opportunities for both more cost-effective testing and a better analytical understanding of the problem.

discrete choice | statistical testing | sample complexity

Significance

Consider the process of testing two vintages of wine, two TV manufacturing processes, or two recommendation algorithms to determine whether one is preferred. Under the standard model of discrete choice, we study a wide range of A/B testing approaches to determine how many samples are required to pick a winner. We observe that, as quality (and level of investment) increases, the distinctions between alternatives become increasingly fine grained. We analyze the setting where the degree of difference between alternatives shrinks toward zero, and compute closed-form expressions for the asymptotically exact sample complexity of each test type. From this characterization, we are able to make specific recommendations for testing methodology at all target levels of error.

Many different forms of tests are used to compare two classes of items. As an example, a company might wish to compare its current cola offering to a new variation it has developed. In the Specified Tetrad test, a rater might be given two samples of the original and two samples of the new variation, in unknown order, and asked to identify the top and bottom two, along some axis such as sweetness, acidity, or overall quality. As a second example using a different test format, Tomonaga and Imura (1) showed chimpanzees $n-1$ generic pictures and one picture distinguished by past experience, in unknown order, and trained the chimpanzees to select the distinguished picture. And as a final example, users of online streaming video services may be offered a carousel of recommended videos assembled in some way from two sources: a production algorithm and a new candidate algorithm undergoing testing; the stronger algorithm may be identified based on the pattern of videos the users choose to watch. More generally, in our setting a rater is given some slate of samples, drawn somehow from two different alternatives, and asked to answer some question: What is the best, or what is the worst, or which are the top two? It is of paramount interest to identify tests capable of teasing out distinctions between classes as efficiently as possible.

Tests of this form are commonly used. The standard formulation was introduced by Thurstone (2) in 1927 and has been widely adopted in psychology for sensory testing (3) and used in animal behavior (1, 4) and animal training (5). The same setup also occurs frequently in the development of popular online experiences that show sets of results such as search ranking or recommendations for movies, books, or mobile apps (6–8). Perhaps more prosaically, the same tests are employed outside commercial and academic settings by enthusiast home brewers and coffee aficionados (9).

Despite broad usage, the current state of the art in understanding the statistical properties of such tests is quite limited: No tight bounds are available for the number of trials needed for a target accuracy, and so statistical power is typically estimated by simulation for specific problem settings. An analytical characterization of sample complexity would unlock more principled approaches to design of testing and would deepen our understanding of this critical area.

In this work, we take a step in this direction. We note a common progression in many testing scenarios, in which the objects to be tested get better and better over time, and the difference between them becomes smaller and smaller as low-hanging fruits are exhausted and diminishing returns set in. By focusing on this scenario of shrinking degree of difference between alternatives, we are able to develop closed-form analytic expressions for the sample complexity of several broad families of common tests. These expressions also demonstrate a number of structural regularities that allow us to make statements about the optimal and pessimal parameter settings for each family of tests we study.

Formal Model

In our setting, the goal is to compare two alternatives, each of which should be viewed as a source of items, rather than a single item. The experimenter may construct tasks for the rater that employ more than one item from each alternative. We adopt the Bradley–Terry

Author affiliations: [a]Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy 00185; and [b]Google, Mountain View, CA 94043

Author contributions: F.C., R.K., and A.T. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: flavio@di.uniroma1.it.

[2] F.C., R.K., and A.T. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2202116119/-/DCSupplemental.

Published July 28, 2022.

model (10), perhaps the most well-studied theoretical formulation of discrete choice. In this model, an item drawn from a particular alternative will be scored by a rater as the sum of a base score for the alternative plus an additive noise term drawn independently from a standard Gumbel distribution.[*] We say that the alternative with the higher base score is strong, and the other is weak. We define ϵ as the difference between the strong and weak base scores, also called the degree of difference between the alternatives. It is well known that the likelihood that a user (or rater) prefers a strong item to a weak item under standard Gumbel noise is a logistic function of ϵ, $f(\epsilon) = \frac{1}{1+e^{-\epsilon}}$. In sensory testing, this is the Thurstonian model with logistic link function (2, 11). In machine learning, it is the multinomial logistic regression model (12, 13) if the base score for an alternative is assumed to be linear in features of the item or a deep neural network with softmax layer (14, 15) if the base score is nonlinear in features of the item.

Families of Tests

The tests we consider all present the rater with a slate of items constructed by drawing s items from the strong alternative and w items from the weak one. The rater is then asked some question about the overall slate of $n = s + w$ items. We consider four families of tests, each characterized by the type of question the rater must answer:

- Permutation test: The rater is asked to order the set of items from worst to best.
- Ordinal(k) test: The rater is asked to return the kth best item.
- Prefix(k) test: The rater is asked to return the set of the k best items (in no particular order).
- Partition test: The rater is asked to partition the items into two groups of equal size, without specifying which group is better; this test is defined only for $s = w$.

These and other forms of tests are commonly studied in the sensory-testing literature (3, 16).

Sample Complexity Results

Perhaps the most fundamental theoretical question in testing is to determine the sample complexity of a test. Sample complexity is defined as the smallest number of trials that are sufficient to determine with error probability at most δ whether the samples were drawn under the null hypothesis in which strong and weak alternatives are identical ($\epsilon = 0$) or, under the alternate hypothesis, in which some small nonzero degree of difference ϵ exists between the base scores of the alternatives.

For the Bradley–Terry model with arbitrary degree of difference ϵ, there is no known closed-form solution for this problem, and it is unlikely that one exists. However, when ϵ approaches zero, we are able to provide closed-form results due to the interplay between the structure of Gumbel noise and the dynamics of shrinking ϵ. (If the noise is instead derived from the Gaussian distribution, which is also studied in sensory testing, we do not see a similar path to a closed-form result even for vanishing ϵ.) We describe our theoretical results next, and then to complement these results, we also present a series of experiments on real data, which confirm that our findings apply in practical settings.

The derivations and background for all results are given in *SI Appendix*, section 2. Our theorems are stated in terms of the harmonic number $H_t = \sum_{i=1}^{t} 1/i$ and the inverse error function $\mathrm{inverf}(y)$, the inverse of the error function $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$. These functions are discussed in *SI Appendix*, section 1.

Theorem 1. *Let the number of strong and weak alternatives s and w be two fixed positive integers, let the degree of difference $\bar{\epsilon}$ between the two alternatives be chosen to be either 0 or some sufficiently small known positive constant ϵ, and let the error probability $0 < \delta < \frac{1}{2}$ be a fixed real number. For Permutation, Ordinal, and Prefix tests, showing a slate of size $n = s + w$, the minimum number of samples to correctly output $\bar{\epsilon}$ with probability $1 - \delta \pm O(\epsilon)$ has the following limiting form as ϵ vanishes*:

$$(1 \pm O(\epsilon)) \cdot C \cdot \frac{1}{\epsilon^2} \cdot \mathrm{inverf}(1 - 2\delta)^2.$$

The coefficient C depends only on the type of test and the parameters s and w and is given by

$$C = \frac{8 \cdot (n-1)}{s \cdot w} \cdot \begin{cases} \frac{n}{n - H_n} & \mathit{Permutation} \\ \frac{n-1}{(H_n - H_{n-k} - 1)^2} & \mathit{Ordinal}(k) \\ \frac{k}{(n-k) \cdot (H_n - H_{n-k})^2} & \mathit{Prefix}(k). \end{cases} \qquad \textbf{[1]}$$

All our discussion below assumes that ϵ is sufficiently small (with respect to $n = s + w$, to the test, and to δ) for the $O(\epsilon)$ term to be negligible. We give empirical results to support this assumption in our experiments. There are a number of direct consequences of our functional form:

- The form of *Theorem 1*, which is a tight characterization of the number of samples, allows robust comparisons of tests. If one test requires twice as many samples as another for a particular δ and a small enough ϵ, it will require twice as many for every δ and every smaller choice of ϵ. Note that merely an upper bound on the sample complexity would be insufficient to compare and rank tests in terms of their efficiency.
- Further, since the value of C depends on the number of strong and weak samples s and w through a $1/(sw)$ term for all tests, the same robustness also applies for construction of the slate: The ratio of samples required between tests with a slate of a given composition is robust to any choice of δ and any small enough ϵ. Therefore, the relative power of these tests can be understood entirely through the behavior of the coefficient C. Consider any pair of tests for which the theorem holds: for instance, the Ordinal(n) test and the Prefix($n/2$) test. No matter what our error tolerance δ or our (sufficiently small) degree of difference ϵ is, these two tests will always have the exact same ratio in number of samples required.
- If δ and ϵ are not fixed, but vary according to the needs of the experimenter, the impact on sample complexity is well behaved. As the degree of difference ϵ is cut in half, the sample complexity grows by a factor of 4. And the dependency on the error probability δ is given by the same function of δ, for all tests.
- The sample complexities predicted by *Theorem 1* have no hidden constants. They exactly reproduce the number of samples required to attain the desired ϵ and δ with a particular test.

Due to these consequences, an experimenter may estimate the per-trial rater cost and then multiply this cost by the sample complexity to determine the cheapest test. The best test will be unchanged for every target accuracy.

[*] Gumbel is a standard noise distribution, defined formally in *SI Appendix*, section 1.

We now discuss the implications of the form of the expression for coefficient C. As we observed above, the value of C depends on s and w only through a $1/(sw)$ term. Thus, for fixed slate size n, and for any parameter k of the Ordinal and Prefix tests, the leading coefficient is minimized and the optimal sample complexity attained with $\{s, w\} = \left\{\left\lfloor \frac{n}{2} \right\rfloor, \left\lceil \frac{n}{2} \right\rceil\right\}$, so an equal slate of strong and weak elements is ideal, even for tests that ask only for the best or worst item. Specifically, and somewhat surprisingly, even when asked to choose the best of 100 samples, no matter the target accuracy, the 100 samples should always be constructed with 50 each from the strong and the weak alternative. The sample complexity increases monotonically with increased imbalance in the composition of the slate. This happens for all δ, all ϵ, all n, and all k.

Based on the expressions in Eq. **1**, it is also possible to determine the settings of k that minimize the sample complexity and likewise to study some other common test settings. We provide such a study in *SI Appendix*, section 3.1. We also provide in *SI Appendix*, section 3.2 a discussion of the simplified sample complexities that result when the slate size n grows large.

For our model as given, we have fully characterized the Permutation, Ordinal, and Prefix tests. For the Partition test (which, in the case $n = 4$ is known in the literature as the Unspecified Tetrad test) the asymptotic sample complexity is as follows:

Theorem 2. *For the Partition test, the number of required samples to provide correct output with probability at least* $2/3$ *has the following limiting form as* $\epsilon \to 0$:

$$\Omega\left(\frac{1}{\epsilon^4}\right).$$

Note that the Partition test has a sample complexity that asymptotically grows worse and worse without bound relative to any of the other tests, as the sample complexity grows as ϵ^{-4} for Partition, rather than ϵ^{-2} for all the other tests we study.

Leading Coefficients Compared

Fig. 1 compares the leading coefficient C for various tests at slate sizes from 2 to 16. Ordinal(n) returns the worst element of the slate and is always the best Ordinal test, although for increasingly large slates its sample complexity becomes unboundedly worse (as a function of n) than the other tests in Fig. 1. Prefix(αn) with $\alpha \approx 0.8$ is always the optimal Prefix test, while Prefix($n/2$) is slightly worse in sample complexity, by an amount that tends to about 35% as n grows. Finally, the Permutation test contains a superset of the information from all other tests and is therefore optimal in sample complexity over all “rank-based tests” (tests like ours that consider just the ordering of elements, rather than the actual scores), asymptotically requiring slightly fewer than half the samples of Prefix($n/2$). However, Ordinal(n) requires the rater to return just a single item, while the other tests require additional levels of cognitive burden. These and related theoretical results are derived and discussed in *SI Appendix*, section 3.1.

Experiments and Discussion

Fig. 2 shows the predicted values of coefficient C from Eq. 1 for a set of 11 different tests. These predictions are plotted against the empirical value of C found in experiments. We show these results for two large datasets, Movies and Books, described in detail in *SI Appendix*, section 4.2. Fig. 2 shows that our predictions correctly order the 11 tests by sample complexity except for two small deviations for Movies and one small deviation for Books. The details of experiment design and datasets, along with discussion of the deviations and the sources of discrepancy between theory and experiments, are given in *SI Appendix*, section 4.

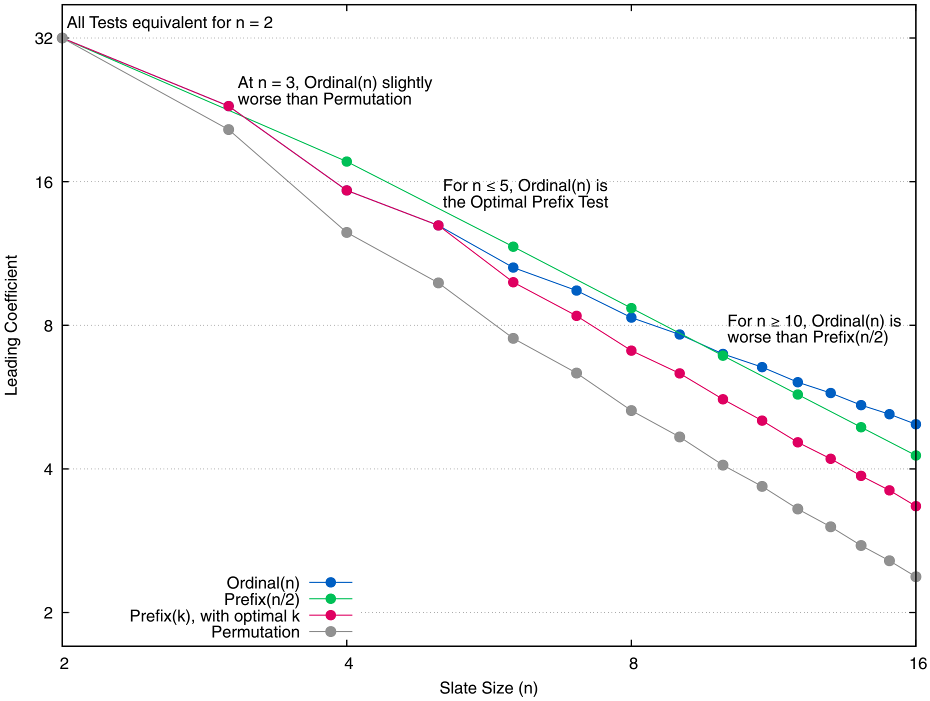

Fig. 1. Comparison of sample complexity of different tests. Shown is the leading coefficient C of the minimum number of samples for various tests using, for a given slate size, the optimal split between weak and strong elements and with small enough ϵ. Observe that the Prefix($n/2$) test is defined only for even n. For $n = 4$, the Prefix($n/2$) test (i.e., the Specified Tetrad test) requires more samples than the Ordinal(n) test (“return the worst item”). Note also that Ordinal(n) is identical to Prefix($n - 1$), which for slates of size five or less is the sample-optimal Prefix test. Conversely, for slate sizes $n \geq 6$ the optimal Prefix test outperforms Ordinal(n), and for $n \geq 10$, the Prefix($n/2$) test also outperforms Ordinal(n).

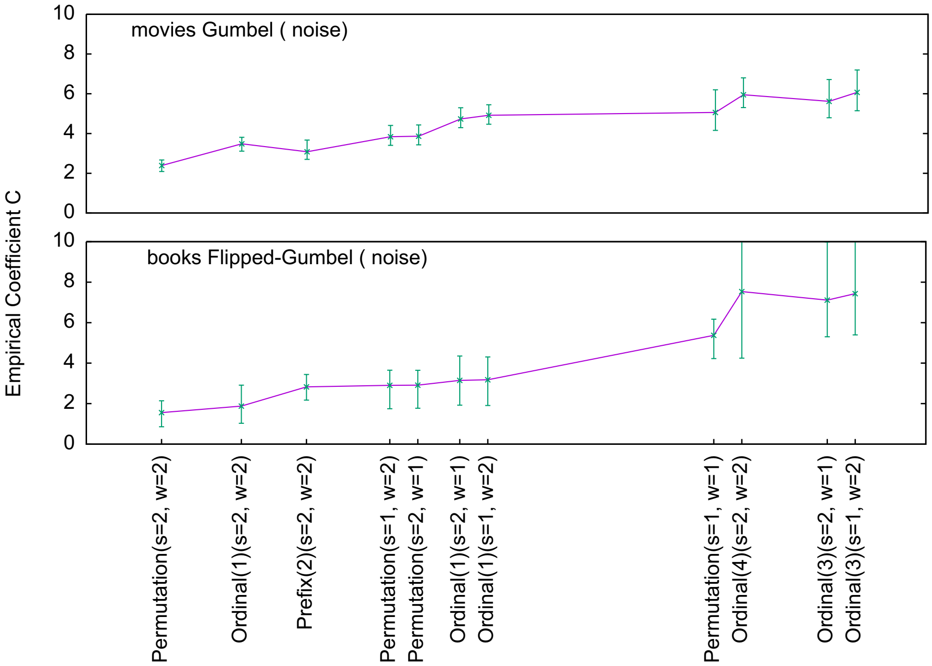

Fig. 2. Theoretical sample complexity leading coefficient *C* for each test is plotted on the *x* axis against empirical sample complexity leading coefficient on the *y* axis. A monotonically increasing line means that the theoretical predictions perfectly order the tests according to sample complexity. Error bars indicate the interquartile range (25th to 75th percentile) of the estimated coefficient *C* across different pairs of alternatives within the dataset.

In *SI Appendix*, section 4 we also show over a range of datasets that the theoretical dependence on ϵ described in *Theorem 1* holds robustly, with most tests and datasets deviating by at most 3% from the form of *Theorem 1*, across a range of values of ϵ. This experiment also shows that our results hold for realistic values of ϵ in practice despite being developed for the setting $\epsilon \to 0$.

Finally, we show in *SI Appendix*, section 4.8 that our predictions in *Theorem 2* regarding the poor asymptotic sample complexity of Partition tests hold strongly in practice, resulting in sample requirements that are 30 times larger in all cases and for some scenarios 2,000 times larger than in all other tests we studied.

We also observe that, while empirical score distributions may not be Gumbel distributed as required by the Bradley–Terry model, they are nonetheless often skewed in one direction or the other. Our results predict asymmetries in tests that ask raters to identify, for instance, the top or bottom element of a slate, depending on the direction of data skew. Fig. 2 shows this distinction in practice: Movies is well-modeled by Gumbel noise while Books is better modeled using flipped-Gumbel noise. Test selection can therefore be informed by the easily observed property of asymmetric tail skew in rater scores.

We close by observing that our results in *Theorem 1* give exact limiting sample complexities that can then be multiplied by empirically observed costs in rater time or budget of performing particular experiments, to find the cost-minimizing approach to a particular desired testing outcome.

Data Availability. Previously published data were used for this work (17–20).

ACKNOWLEDGMENTS. F.C. was supported in part by the European Research Council Starting Grant DMAP 680153, by a Google Focused Research Award, by the Progetti di Rilevante Interesse Nazionale project 2017K7XPAN, by Bertinoro International Center for Informatics, and by the "Dipartimenti di Eccellenza" grant awarded to the Dipartimento di Informatica at Sapienza. We were fortunate to receive some comments from our colleague Professor V. S. Subrahmanian of Dartmouth University.

1. M. Tomonaga, T. Imura, Efficient search for a face by chimpanzees (*Pan troglodytes*). *Sci. Rep.* **5**, 11437 (2015).
2. L. L. Thurstone, A law of comparative judgment. *Psychol. Rev.* **34**, 273–286 (1927).
3. D. Ennis, *Thurstonian Models–Categorical Decision Making in the Presence of Noise* (The Institute for Perception, Richmond, VA, 2016).
4. J. Taubert, L. A. Parr, The perception of two-tone Mooney faces in chimpanzees (*Pan troglodytes*). *Cogn. Neurosci.* **3**, 21–28 (2012).
5. J. H. Bak, J. Y. Choi, A. Akrami, I. Witten, J. W. Pillow, "Adaptive optimal training of animal behavior" Advances in Neural Information Processing Systems 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (NIPS, 2016), pp. 1947–1955.
6. Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, H. Li, "Learning to rank: from pairwise approach to listwise approach" in *Proceedings of the 24th International Conference on Machine Learning*, Z. Zoubin Ghahramani, Ed. (Omni Press, 2007), pp. 129–136.
7. C. Burges *et al.*, "Learning to rank using gradient descent" in *Proceedings of the 22nd International Machine Learning Conference*, L. De Raedt, S. Wrobel, Eds. (ACM Press, New York, NY, 2005), pp. 89–96.
8. J. J. Jeon, Y. Kim, Revisiting the Bradley–Terry model and its application to information retrieval. *J. Korean Data Inf. Sci. Soc.* **24**, 1089–1099 (2013).
9. C. Coffee, *What is triangle testing?* (2022). https://www.chriscoffee.com/blogs/main/refining-our-palate. Accessed 27 May 2022.
10. R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).
11. L. L. Thurstone, The measurement of social attitudes. *J. Abnorm. Soc. Psychol.* **26**, 249–269 (1931).
12. D. R. Cox, "Some procedures connected with the logistic qualitative response curve" in *Research Papers in Probability and Statistics (Festschrift for J. Neyman)*, F. N. David, Ed. (Wiley, 1966), pp. 55–71.
13. H. Thiel, A multinomial extension of the linear logit model. *Int. Econ. Rev.* **10**, 251–259 (1969).
14. L. Boltzmann, Studies on the balance of living force between moving material points. *Wiener Berichte* **58**, 517–560 (1868).
15. J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition" in *Neurocomputing: Algorithms, Architectures and Applications*, F. Soulié, J. Hérault, Eds. (Springer, 1989), pp. 227–236.
16. H. S. Lee, M. O'Mahony, The evolution of a model: A review of Thurstonian and conditional stimulus effects on difference testing. *Food Qual. Prefer.* **18**, 369–383 (2007).
17. B. C. Franczak, R. P. Browne, P. D. McNicholas, C. J. Findlay, Product selection for liking studies: The sensory informed design. *Food Qual. Prefer.* **44**, 36–43 (2015)
18. F. M. Harper, J. A. Konstan, The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.* **5**, 19:1–19:19 (2015).
19. T. Kamishima, "Nantonac collaborative filtering: recommendation based on order responses" in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2003), pp. 583–588.
20. C. N. Ziegler, S. McNee, J. Konstan, G. Lausen, "Improving recommendation lists through topic diversification" in *Proceedings of the 14th international conference on World Wide Web* (Association for Computing Machinery, New York, NY, 2005), pp. 22–32.