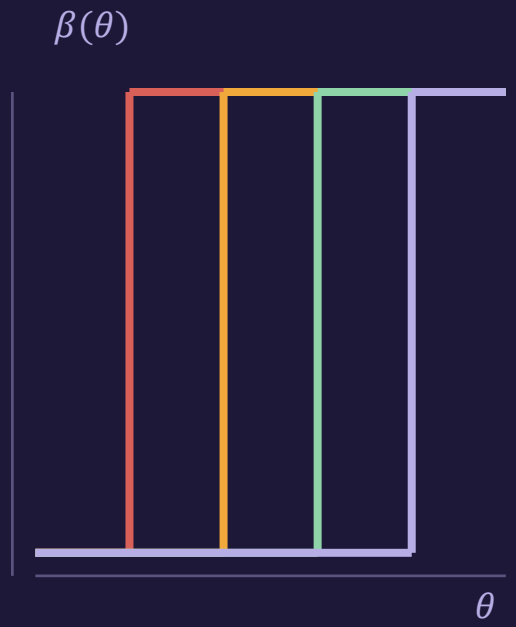
Two is better than one

Optimizing binary rating scales for online platforms

Alec Sun (University of Chicago)

Joint work with **Liren Shan · Jingyan Wang**

NeurIPS 2026



01

Motivation

Ratings on online platforms are inflated



Online ratings are inflated

74%

of Airbnb reviews give 5 stars

Fradkin et al., 2021

100%

median positive-rating share, eBay sellers

Nosko & Tadelis, 2015

99.3%

mean positive-rating share, eBay sellers

Nosko & Tadelis, 2015

Why that's a problem



Tiny, noisy differences

Users compare restaurants on rated 4.8 vs. 4.9 which are mostly noise



Bimodal extremes

Ratings pile up at extremes, hiding the middle



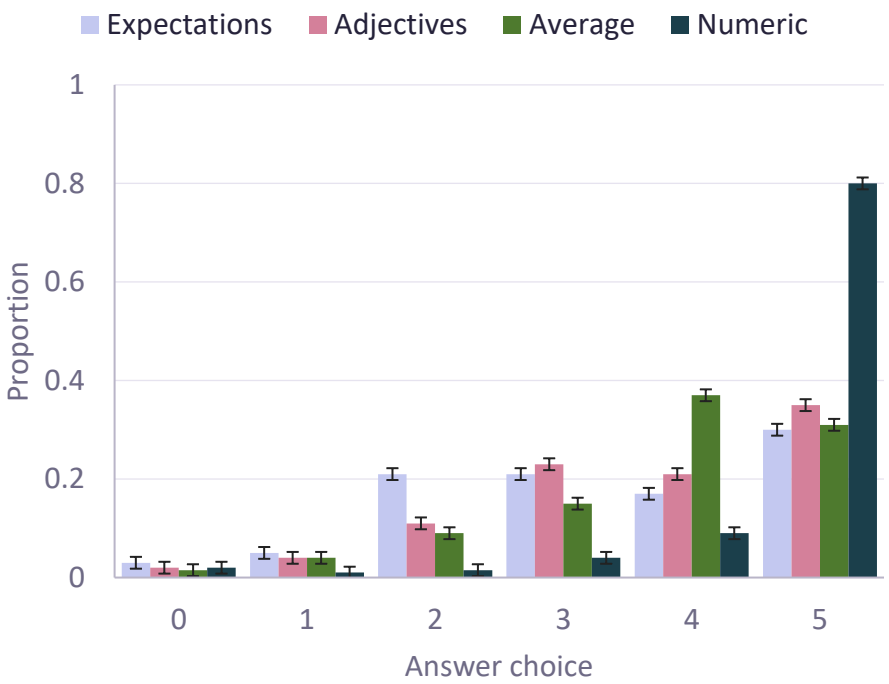
Weaker recommendations

Algorithms estimating quality from ratings inherit noise

Verbal rating scales

Treatment	Question	Answer choices
Expectations	How did this freelancer compare to your expectations?	Much worse than I expected, Worse than I expected, About what I expected, Better than I expected, Far better than I expected, Beyond what I could have expected
Adjectives	How would you rate this freelancer overall?	Terrible, Mediocre, Good, Great, Phenomenal, Best possible freelancer!
Average	How does this freelancer compare to others you have hired?	Worst Freelancer I've Hired, Below Average, Average, Above Average, Well Above Average, Best Freelancer I've Hired
Numeric	How would you rate this freelancer overall?	0, 1, 2, 3, 4, 5

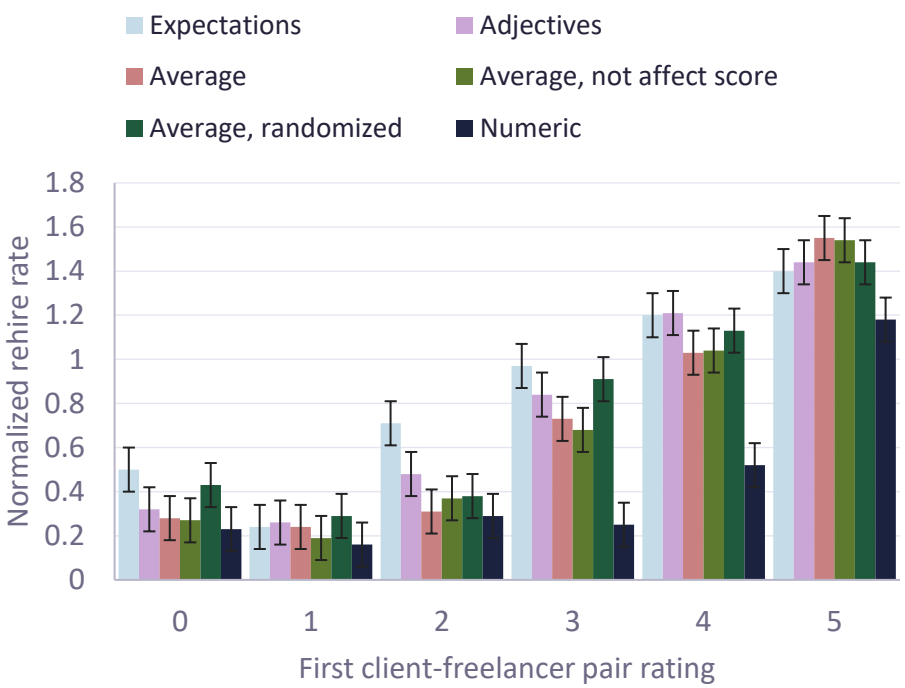
Verbal scales deflate ratings



Expectations	Compared with expectations
Adjectives	Terrible to Best possible freelancer
Average	Compared with others hired
Numeric	0 to 5, the standard scale

About 80% of numeric ratings are a 5, vs about a third on verbal scales

Deflated ratings are more informative



The outcome

Did the client rehire the freelancer?

Verbal scales

Rehire rate climbs steadily from rating 2 to 5

Numeric scale

Most get 5 stars; a 5 predicts rehiring less than top verbal ratings

Deflated ratings predict rehiring better

Why a binary rating model?



Simple to deploy

Easy to ask, answer and randomize



Matches human limits

People can't reliably use fine-grained numbers (Miller, 1956)



Tractable to optimize

One curve $\beta(\theta)$ summarizes a question; scales can be optimized

An item of quality θ gets



yes

with probability $\beta(\theta)$

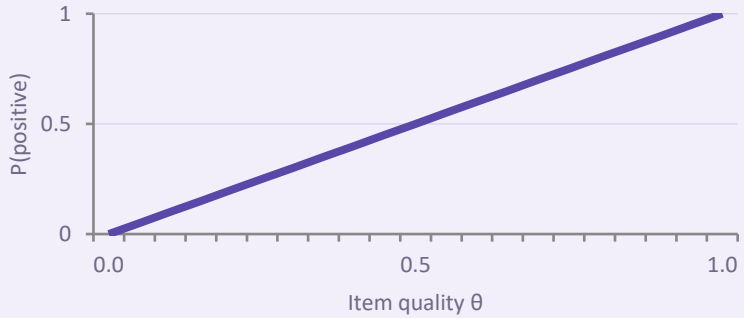


no

with probability $1 - \beta(\theta)$

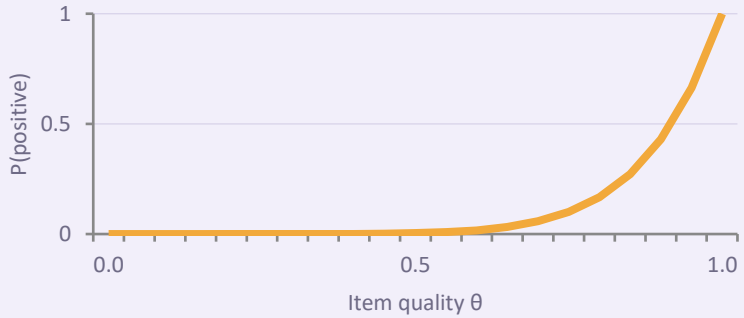
Changing the question changes the scale

“Did you have a positive experience?”



Lenient: positive rate tracks quality ($\beta \approx \theta$)

“Was it the best experience you've ever had?”



Strict: $\beta(\theta)$ stays near 0 unless θ is close to 1

Each question sets curve $\beta(\theta)$; the question is the curve

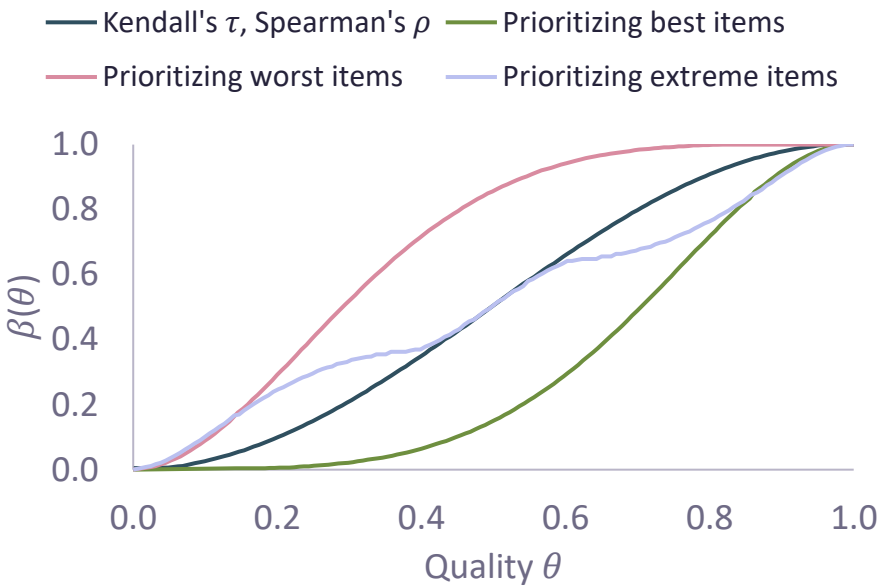
Best scale depends on the goal

A commodity marketplace wants to separate its bottom 5% of sellers

Which rating system?

An inflated one:

The “prioritizing worst items” curve gives most sellers a yes; only weak ones get a no



Inflation isn't always bad; the best curve depends on which comparisons matter. Weight $w(\theta_1, \theta_2)$ sets this.

02

The model

Formalizing binary ratings and the ranking goal



A rating scale is a function β



Item quality θ

Hidden quality $\theta \in [0, 1]$; higher θ is better



Binary question

Yes/no question with response curve β



Response

Positive with probability $\beta(\theta)$; β rises with θ



Aggregate

After k ratings, rank by positive count

Multiple scales

User i has own scale β_i, and number $n_k(\theta)$ of positive ratings is a sum of k independent flips with biases $\beta_1(\theta), \ldots, \beta_k(\theta)$

Why $\theta \in [0, 1]$?

Any quality distribution rescales so θ is the quality quantile, uniform on [0, 1]

Goal: rank every pair of items correctly

Pairwise accuracy for items $\theta_1 < \theta_2$

+1 Better item has more positives: correct

0 Equal positives: a tie

−1 Worse item has more positives: wrong

Which pairs matter? A weight $w(\theta_1, \theta_2)$

Kendall's τ: $w = 1$

All pairs count equally

Spearman's ρ: $w = \theta_2 - \theta_1$

Far-apart pairs matter more

Objective W_k: weighted average pairwise accuracy, where perfect ranking is $W_k = 1$

Benchmark: best single scale

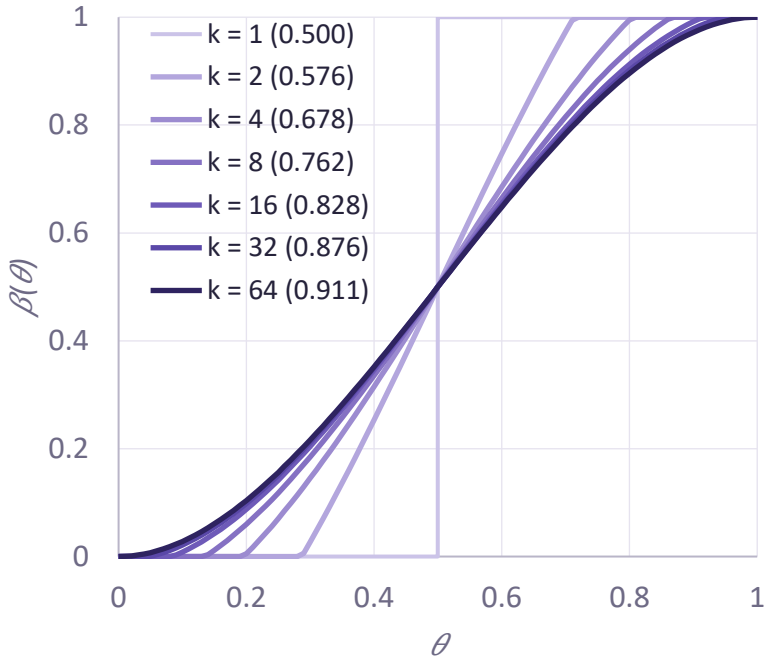


Figure 5 · Best discretized single scale (Kendall's τ); objective in parentheses in legend

One scale for all users
Prior work (Garg & Johari, 2019) optimizes a single scale β

No closed form
Solved numerically on a 50-step grid, plus an analytic upper bound

Accuracy grows slowly with *k*

0.576 at $k = 2$ → **0.911** at $k = 64$

What if different users get different questions?

Prior work: one scale for all

Same question for all, so each rating is an identical noisy coin flip

vs

This paper: heterogeneous scales

Half of users get “Was it above average?”, half “Was it the best you've ever had?”

How should platforms design heterogeneous rating scales, and how much do they gain over one scale?

Multiple scales never hurt (set them equal), but the quantitative gain is unknown

Three results at a glance

1 LOCAL

Two is better than one

Split any scale in two, each used by half of users: better ranking

strict gain

2 GLOBAL

Evenly spaced steps

Per-user scales: optimum is k sharp threshold questions, evenly spaced

$$W = 1 - \frac{1}{k+1}$$

3 ADAPTIVE

Binary search

Adaptive questions: the optimum binary-searches quality

$$W = 1 - \frac{1}{2^k}$$

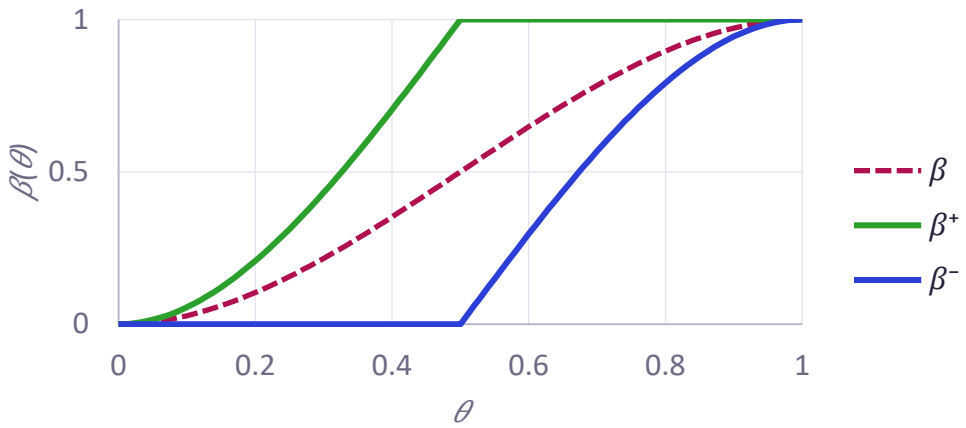
03

Two is better than one

Local result: splitting any scale in two always helps



Splitting one scale into two



$$\beta^+(\theta) = \min\{2\beta(\theta), 1\}$$
$$\beta^-(\theta) = \max\{2\beta(\theta) - 1, 0\}$$

Use β^+ for half of k ratings and β^- for the rest

Figure 1 · Splitting an S-shaped β into β^+ and β^-

Theorem 3 For any scale β, any even k and any weight, the split pair ranks at least as well as β alone, strictly better unless β is a 0-1 step function

Why splitting helps

Specialization

β^+ doubles the slope for weaker items; β^- doubles it for top items: steeper scales separate similar items better

Same signal

Expected positives unchanged: $\beta^+ + \beta^- = 2\beta$

Less noise

Lower variance, same mean: fewer ranking mistakes

Example: an item with $\beta(\theta) = 0.3$, two ratings

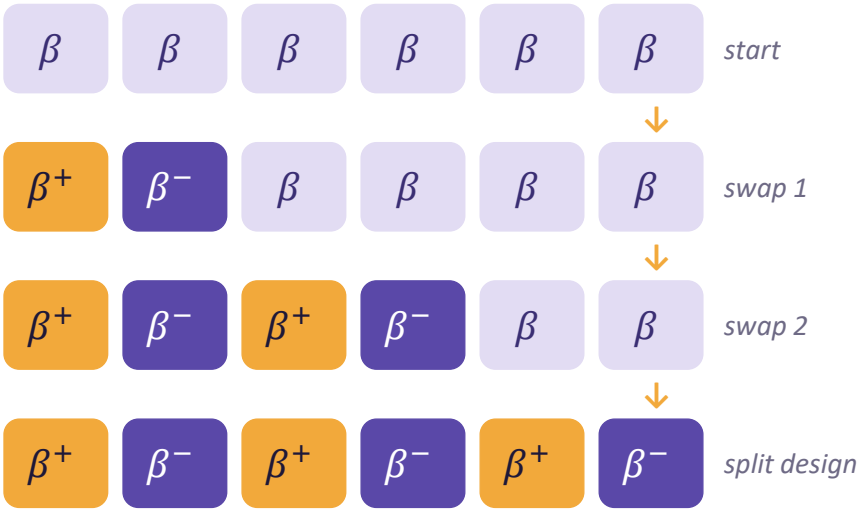
Original	Split
Bern(0.3) + Bern(0.3)	Bern(0.6) + Bern(0)
0.42	**0.24**
variance	variance

Same mean (0.6), 43% less variance

In general, for $p < \frac{1}{2}$: $2p(1-2p) < 2p(1-p)$

Keep splitting: every split helps

How the proof works (idea only)



Each swap trades two β ratings for one β^+ and one β^-, shifting probability toward correct orderings

Recursive splitting

Split a scale used by 2m users into two of m users each: accuracy rises every time

More scales, never worse

More scales are always better: one per user

04

The optimal set of scales

Global result: the best design when each user gets their own scale



The optimum: evenly spaced step functions

THEOREM 4

Best k scales are 0-1 steps, one per rating

Threshold scales: $\beta_i(\theta) = \mathbf{1}[\theta \geq \tau_i]$

Thresholds for Kendall's τ: $\tau_i = \frac{i}{k+1}$

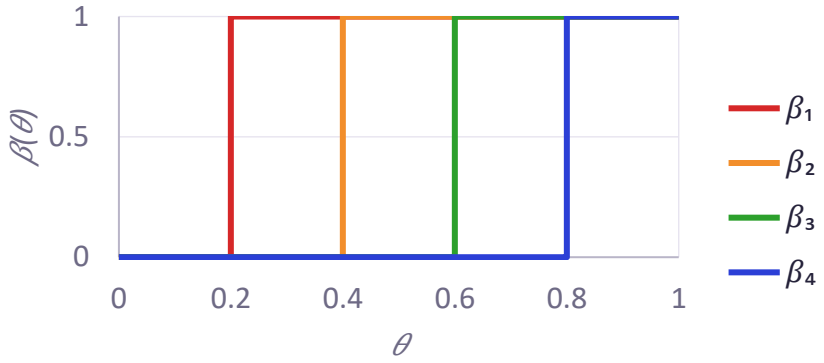


Figure 2 · Optimal configuration for $k = 4$ (Kendall's τ)

User i answers one sharp threshold question, "Is this item better than the $\frac{i}{k+1}$ quantile?"

Resulting objective $1 - \frac{1}{k+1}$

Threshold questions sort items into buckets



Different buckets

Always ranked correctly

Same bucket

A tie: no information to separate them

The only loss

Same-bucket probability $1/(k+1)$

Why step functions are optimal

1

Randomness never helps

Accuracy is linear in an item's outcomes: deterministic is optimal

2

Respect the order

Worse item with more positives: swap. Outcomes rise with θ.

3

Equal-size buckets

Maximizing $\sum x_i x_j$ over lengths summing to 1 forces all $k+1$ buckets equal

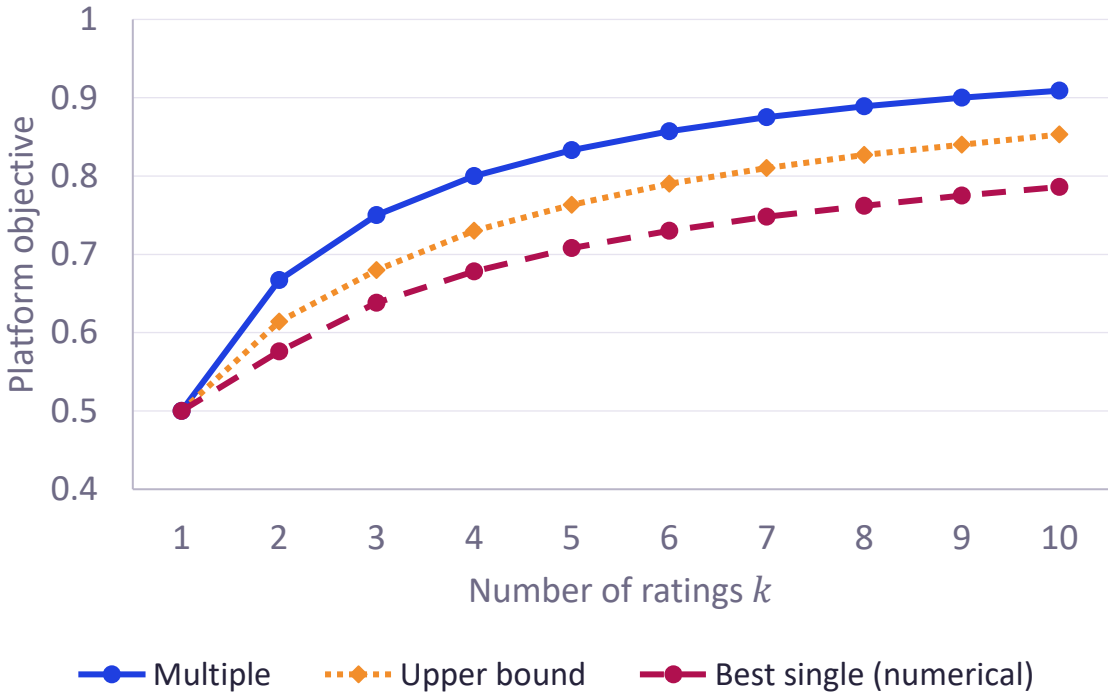
4

Build it with steps

k step-function scales at thresholds $i/(k+1)$ produce this assignment

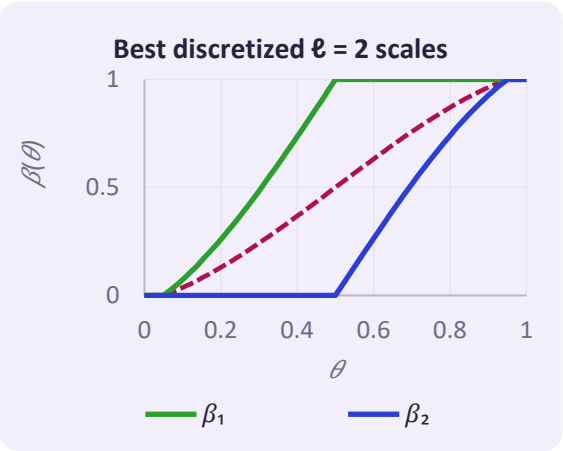
Steps 1, 2, 4 hold for any weight; only step 3's bucket spacing depends on w

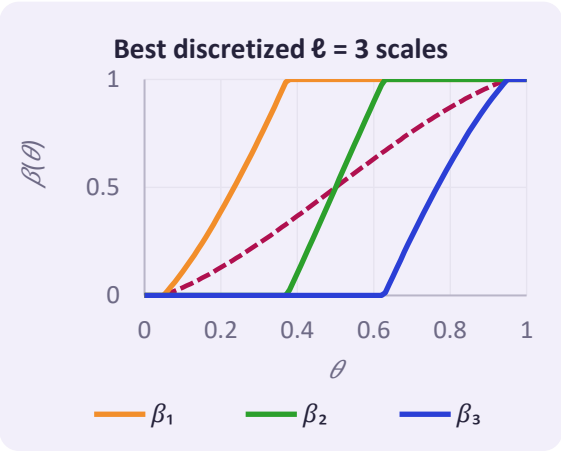
How much do multiple scales gain?



Multiple scales approach perfect ranking at rate $1/(k+1)$

In between: a few scales, shared by users





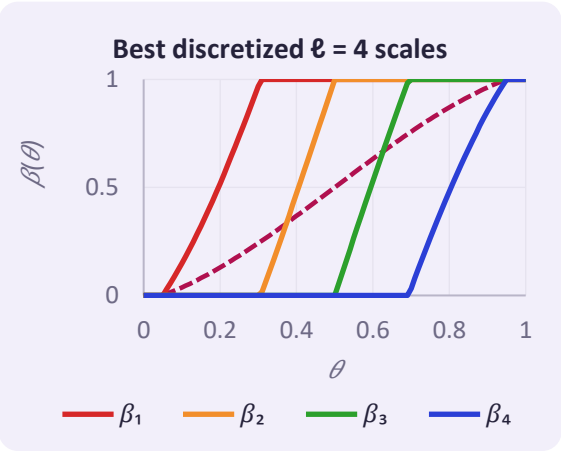


Figure 4 · What the numerics show

- $k = 24$ ratings split evenly across ℓ scales, dashed β = best single scale
- Scales specialize: each is steep over its own quality range
- As ℓ approaches k, graphs approach step functions

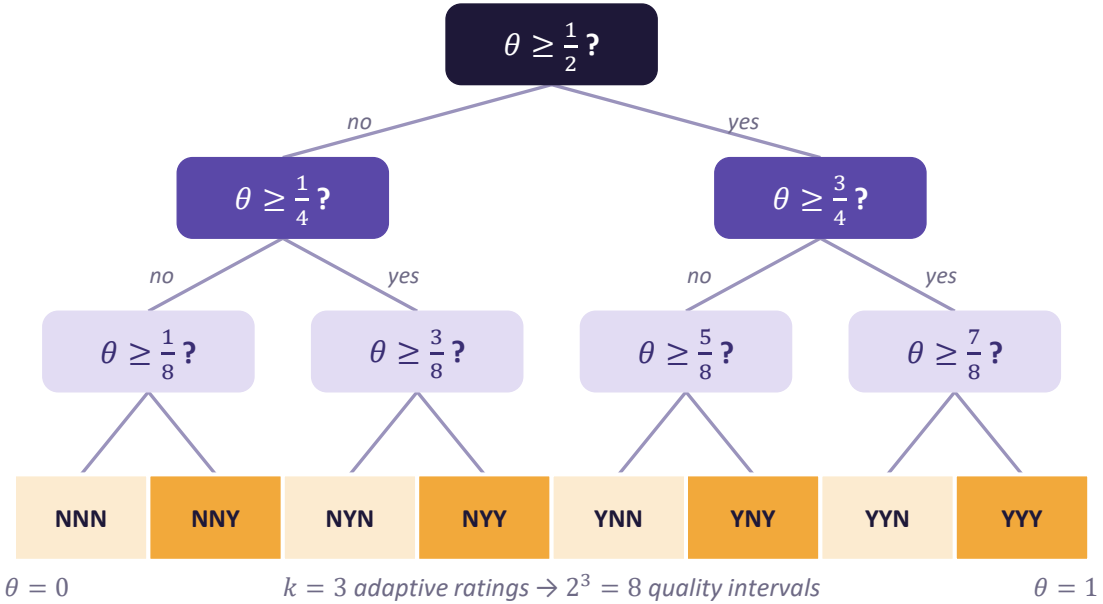
05

Adaptive scales

What if the next question can depend on earlier answers?



Adaptivity turns rating into binary search



Users arrive one at a time; user j's scale depends on the first $j - 1$ ratings

A design is a depth-k decision tree of rating scales

Each answer halves the range; ratings spell out θ's binary digits

Adaptivity yields exponentially better ranking

2^k

equal-length intervals, vs. $k+1$ without adaptivity

$1 - \frac{1}{2^k}$

optimal objective under Kendall's τ

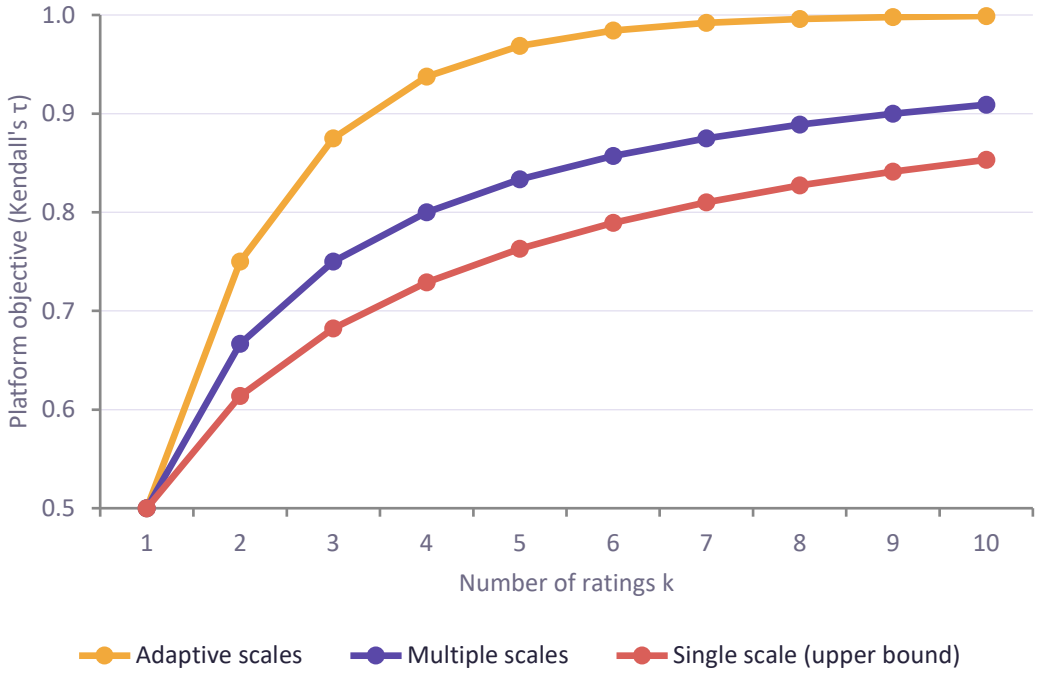
$2^k - 1$

non-adaptive ratings needed to match k adaptive ones

Example: 10 adaptive ratings rank as well as 1023 non-adaptive ones

Still 0-1 steps; thresholds now depend on earlier answers

Three designs, side by side



Ratings needed to reach 0.9

4	**Adaptive**
9	**Multiple**
≥ 17	**Single**

provable bound; numerical single scale is much greater

06

Discussion

Related work, implementation, and open questions



Where this fits in the literature

Rating inflation & design

Rewording labels and asking the right question helps: Garg & Johari (2019) optimize a single scale

Miscalibration

Earlier methods correct ratings afterwards; scale design changes calibration at collection

Crowdsourcing

Dawid–Skene methods learn annotator expertise from data; here, variation is by design

Item response theory

Like adaptive testing, but derives optimal response functions instead of picking from a question bank

From theory to practice

Limitation

- Perfect 0-1 steps assume users judge thresholds exactly: unrealistic
- Real questions in practice cannot distinguish low-quality items

Implementation

- Randomize each user's question
- Garg & Johari (2019): test candidate questions on reference items of known quality, then pick how often to show each so the mix matches the target scale

Mechanical Turk workers rated TOEFL paragraphs

10 TOEFL writing samples of varying quality

5 expert scores, two paragraphs each

9 adjectives, from Awful to Phenomenal

The question each worker saw

“Would you consider this paragraph of quality [adjective] or better?”

What it produced

The share of workers saying yes, for every paragraph and adjective

Simulated with 500 items and 100 buyers

The adjective mix chosen from this data beat showing every adjective equally

Source: Garg & Johari (2019), Designing Optimal Binary Rating Systems, AISTATS; ETS (2005), TOEFL iBT Writing Samples

Open questions



Better aggregation

Go beyond ranking by positive rating count



Joint design

Optimize rating scales and aggregation rule together



Recommender co-design

Co-design online recommenders with (possibly adaptive) rating questions

Key takeaways

1. Splitting any rating scale in two strictly improves item ranking
2. Optimal scales are threshold questions at evenly spaced quantiles
3. How a platform asks users matters: heterogeneous questions make ratings far more informative