Episode Book

EPISODE #6583

Why Comparing Two Things Beats a Rating Scale

PUBLISHED	14 November 2026 at 09:19 UTC
DURATION	25:22
LISTEN	myweirdprompts.com/episode/pairwise-comparison-vs-rating-scales-85194/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	21 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 11:16 UTC

SYNOPSIS

Humans are terrible at holding an absolute rating scale steady, which is why comparing two options side by side almost always wins. From psychophysics and 2AFC to ITU-T P.800 audio testing, Thurstone's 1927 law, and the Bradley-Terry model behind Chatbot Arena, four different fields invented the same trick under different names. Dropping a five-point scale for a one-bit choice cancels out rater fatigue and calibration drift, and adding a hidden reference like MUSHRA keeps your ears honest.

TRANSCRIPT



Corn

Four takes of the same Herman clip, played one after another. That's what Daniel spent his afternoon on. Same sentence, same donkey, four nearly identical EQ curves, and his ears gave up inside ten minutes.



Herman

Which, for the record, is the correct reaction. If you can easily tell them apart in sequence, the clones weren't worth choosing between.



Corn

Right. So Daniel did what he usually does, which is rebuild from scratch something that already has four names. He set it up as A versus B, two clips at a time, four answers: A better, B better, both are bad, both about the same. And it worked, far better than the scale ever did. His actual question is what this method is called, in psychophysics, in audio listening tests, in AI evaluation, and why comparing two things at once beats rating each on a scale. He's framing it as part one of three, so we keep this one on the concept and leave the tooling and the mathematics mostly alone.



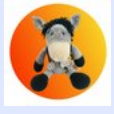
Herman

Thank goodness, because the mathematics is a whole episode of its own.



Corn

It's at least two.



Herman

The thing I love about this prompt is that the forgetting is the finding. Daniel keeps losing the name because there isn't one name — there are four, one per tradition, and the traditions don't read each other's papers. A psychoacoustician calls it two-alternative forced choice. A speech codec engineer calls it Comparison Category Rating. A psychometrician calls it paired comparison and reaches for Thurstone. An ML evaluation person calls it pairwise preference and reaches for Bradley-Terry. Same act, four mental models.



Corn

So give me the headline version before we go tradition by tradition.



Herman

Psychophysics: 2AFC, developed out of Fechner's work, where the observer is not allowed to say "I don't know." The choice is forced between the two alternatives. Statistically it's paired comparison, and Thurstone's law of comparative judgment, from 1927, is the model that turns a pile of pairwise judgements into an interval scale. Audio: the governing standard is ITU-T P.800, and the pairwise procedure is CCR, an annex bolted on in the 1996 revision. AI evaluation: pairwise comparison, aggregated with Bradley-Terry, which is the machinery behind Chatbot Arena.



Corn

And the reason all three landed on the same move?



Herman

Humans are bad at holding an absolute scale steady and good at judging which of two things is better. The noise you delete by asking for a comparison is usually worth more than the resolution you lose by asking for one bit.



Corn

One bit. That's the whole bargain.



Herman

Although — one thing Daniel's answer set has that none of the ITU scales do is "both are bad." Park that. We'll pull the thread later.



Corn

Noted. Start with the psychophysics, since it's the oldest.



Herman

Two variants matter here. Concurrent presentation, both options side by side, which is exactly what Daniel did with two clips on a page. Or sequential, two intervals, and you answer only "first" or "second." The second one is where the forcing really shows, because the observer can't say "I didn't detect anything." The Wikipedia article on 2AFC has a line I like: the observer is not allowed to say "I do not know," or "I am not sure," or "I did not see anything."



Corn

A test design with opinions about my honesty.



Herman

It's earned them! There's a classic example with extremely dim lights. Ask a person "did you see a light, yes or no" and they may truthfully answer no on every single trial. Run the same dim light through a 2AFC task — was it top or bottom — and the same person reliably gets the location right. The information was in them the whole time. The yes-no question just couldn't reach it.



Corn

Because "top or bottom" is a comparison and "did you see it" is an absolute judgement against an internal threshold that drifts.



Herman

Which is precisely Daniel's EQ problem. "Is this clone's timbre right" asks him to hold a memory of my voice steady for as long as the comparison takes. "Which of these two is closer" asks him to hold it for about three seconds.



Corn

Flattering that the reference is my voice and I still can't judge it absolutely.



Herman

Then Thurstone. Nineteen twenty-seven, the law of comparative judgment. The point of it is that from many pairwise judgements you can reconstruct an interval-scale ordering of the things along whatever dimension people were judging. With one caveat he was explicit about: the measurements represent how we perceive entities, not the actual physical properties. You're measuring the perception, not the stimulus.



Corn

The clone that wins is the clone that sounds most like me to the listener. Not the clone with the smallest spectral error.



Herman

Right. And there's a lineage detail I find delightful. Zermelo — the set theory Zermelo — described a pairwise model for ranking chess players in incomplete tournaments in 1929. That model is equivalent to Bradley-Terry, proposed in 1952. So chess ratings and the LLM leaderboard are the same mathematics. Elo, Bradley-Terry, Thurstone, one family.



Corn

Daniel's four Hermans get ranked the way grandmasters do.



Herman

That's the yes. Now the audio tradition, because this is where it gets institutional. P.800 is the telephone-quality standard, and its default listening test is the opposite of what Daniel did. Absolute Category Rating: one clip at a time, five-point scale from excellent down to bad, averaged into a mean opinion score. That's your MOS. The whole industry ran on that for decades, because it fits the use case — rate how a codec sounds, clip by clip, no reference needed.



Corn

One clip, one score, no comparison. The thing that just failed Daniel for an afternoon.



Herman

It failed him because his clips are nearly identical. If the differences are small, absolute rating drowns them in its own noise. So in the 1996 revision they added Annex E, Comparison Category Rating. Paired samples in random order, and you rate the second against the first on a seven-point scale: much better plus three down to much worse minus three. Averaged, that gives a CMOS. And note what that scale has in the middle — zero, "about the same." Daniel's fourth option, sitting right there in the middle of an ITU annex from 1996.



Corn

"Both about the same" is literally their zero point.



Herman

It is. Then the cousins. DCR, degradation category rating — you hear the reference, then the sample, and rate how degraded the sample is. That's the reference-anchored version, and for a voice clone judged against the original recording it's arguably the closest formal procedure to what Daniel was doing. MUSHRA is the multi-candidate version: one open reference, every candidate rated zero to a hundred on the same page, with a hidden reference and low-pass anchors thrown in. And ABX is the different question entirely — hear A, hear B, hear X, say whether X is A or B. Scored with a binomial test on proportion correct.



Corn

That's "can you tell them apart," not "which is better."



Herman

It's the prior question. Before you rank the clones, ABX tells you whether you can even distinguish them. If Daniel cannot tell the original from a given clone at a rate above chance, the ranking below that is aesthetic preference, not fidelity. And people run ABX on exactly this — codec transparency, whether a lossless chain is audibly different from the original.



Corn

Map Daniel's four answers onto the standards, then.



Herman

A better and B better — CCR's better and worse. Both about the same — CCR's zero. Both are bad — nothing. No ITU scale has it. ACR, DCR, CCR all assume every stimulus is rateable. The nearest relatives are P.835, which separates signal quality from overall preference so you can say "the speech is clean and I still hate it," and the way MUSHRA's anchors work, which lets a listener put everything low. But "both are bad" as an explicit button is Daniel's own contribution.



Corn

Which I'd argue is the most honest button on the page.



Herman

Why?



Corn

Because it separates "I can't hear a difference" from "I can hear a difference and I dislike both." Those are completely different data. Collapse them and your ranking quietly becomes garbage — you're scoring indifference and disgust on the same axis. Daniel's protecting the comparison question from the quality question.



Herman

That's a real defense of it. Whether it survives contact with the ranking math is another matter, and we'll get there. But now the mechanism, because we've said "one bit beats a scale" twice without earning it.



Corn

I was about to object on information grounds. A one-to-five score carries more bits than a coin flip. By the data processing inequality, the coin flip cannot be better.



Herman

That's the trap, and it's the intellectual centre of this whole literature. There's a paper from 2014, "When is it Better to Compare than to Score," that takes exactly your objection head on. Their answer is that the data processing inequality doesn't apply when you're comparing cardinal and ordinal work from humans. The inequality assumes a fixed channel. A human isn't a fixed channel. The human's absolute-score channel is noisy in ways the comparison channel isn't — calibration drift, inflation, hedging — and that noise eats the extra bits before you ever get to use them.



Corn

The scoreboard has more pixels but the lens is smeared.



Herman

They ran seven experiments on Mechanical Turk, a hundred subjects each, randomly assigned to the ordinal or cardinal version of the same task. Average per-sample noise was often significantly higher in the cardinal measurements. And the speed: on one task, ninety-eight seconds for the comparison versus a hundred and eighty-one for the score. On another, eighty-four versus three hundred and five.



Corn

Three hundred and five seconds. To produce a number nobody trusts.



Herman

To produce a number that is mostly noise about the rater. Because asking for a score asks the reviewer to do two jobs at once: build a mental reference for what a three looks like versus a five, and then hold that reference steady across a hundred examples reviewed over an afternoon. They don't. Scores cluster in the middle because reviewers hedge, and they drift over a session as fatigue sets in. There's also a lovely invariance argument — when an annotator picks A over B, they communicate real information regardless of whether they'd rate both excellent or both terrible.



Corn

The comparison is immune to the annotator's mood.



Herman

Additive bias cancels in the subtraction. If your ear is having a harsh day, both candidates sound worse to you, and the preference still means what it meant. That's the bridge to the voice-clone case: Daniel, one listener, one afternoon, four clones — the textbook setting for absolute-scale collapse, and the setting where the one-bit question holds up.



Corn

Before we go back to the clones — you know what I'd want, if I were Daniel?



Herman

Tell me.



Corn

Some way of knowing my ear is working today. He's one rater. The entire result rests on one pair of ears with no calibration check.



Herman

That's the gift sitting in MUSHRA. The hidden reference. You embed the true reference among the candidates without telling the listener. The post-screening rule: exclude anyone who rates the hidden reference below ninety on more than fifteen percent of items, or rates a low-pass anchor above ninety on more than fifteen percent. If you can't reliably put the real thing at the top, your ratings aren't measuring what you think.



Corn

Daniel could slip the original Herman clip in as a blind fifth candidate.



Herman

Four clones plus the real me, shuffled, no labels. If the real clip doesn't win most of its pairs, the afternoon is measuring fatigue, not fidelity. It's an ear check disguised as another data point.



Corn

I'd like to say my own clip would beat my clones. I've heard my clones.



Herman

This one — let's hear what our listeners made of all this. Here's a voicemail from Opal. Opal: Hi Daniel and Corn and Herman, I am calling from the bus, the number eighteen in Tel Aviv, standing up because there are no seats and holding the pole with one hand and the phone with the other. So today you are talking about comparing two things instead of rating them on a scale, which is the what, the A B test thing, right? My question is, does it still work if the two things are both sort of bad and I cannot tell which one is less bad, because last week I had to choose between two apartments and both had a mold problem and I chose the one with the smaller mold and I still think about the other one.



Corn

Opal, first — hold on to that pole.



Herman

Second: you landed on the exact edge we're circling. Two bad options is where comparison is weakest, because "which is less bad" still works, but the signal you're extracting is tiny and the regret is enormous. Your apartments question had a hidden third answer — "neither, keep looking" — which is the apartment version of Daniel's "both are bad" button. Sometimes the right output of an evaluation is walking away, and that's a decision the A-versus-B page can't make for you.



Corn

The smaller mold was the correct answer. That you still think about the other one isn't an evaluation problem, it's a mold problem.



Herman

There is an honest caveat coming, though, and it cuts against everything we just said. Pairwise is not universally better. There's a study comparing five crowdworker evaluation methods across three dialogue settings, and the finding was different methods won depending on the models being compared. No clear winner. Pairwise per-turn worked when pairs clearly differ. Pairwise per-dialogue when differences only emerge over several turns. But single-model scoring — rate each one on its own — performed well when comparing models that differ only slightly in quality but are otherwise similar. Two models with different parameter counts.



Corn

Which is almost exactly Daniel's situation. Four near-identical clones, one afternoon.



Herman

Almost. It's stated for dialogue models, not audio, and nobody has run Daniel's exact design — one reference, four clones, a four-answer set including "both bad." The searches come back empty. So the one result arguing his experience isn't universal is a result from a different modality, untested on his case. His afternoon worked; the literature hasn't told us it always will.



Corn

Good. I was starting to feel like the coin flip had a halo.



Herman

It doesn't. Now the jump to models on any task, because this is where the idea went industrial. The pattern: elicit binary preferences on pairs, then fit a Bradley-Terry model to recover a latent score and a ranking. That's Chatbot Arena — over two hundred and forty thousand crowdsourced votes at the time of its paper, and one of the findings worth holding is that the crowd's votes agreed well with expert raters. Random people clicking A or B, aggregated, tracked the specialists.



Corn

And then it breaks.



Herman

Then it breaks, in three specific ways. First, non-transitivity. LLM judges — and humans, given enough items — exhibit rock-paper-scissors cycles. A beats B, B beats C, C beats A. The common cheap design is to compare everything against one baseline model, which silently assumes preferences are transitive. They aren't, and rankings turn out to be sensitive to which baseline you picked. The fix is a round-robin tournament — every model plays every other — with Bradley-Terry over the results. That moved Spearman correlation with the Arena leaderboard from ninety-five percent to about ninety-six and a half, and Kendall from eighty-two to eighty-six.



Corn

A percentage point and a half, bought with every model playing every other.



Herman

Every pair, which is quadratic. So the same paper proposes Swiss-system matchmaking to get most of it for less. Second pathology: position bias. Judges prefer whichever answer came first, or second, depending on the judge. It's the pairwise-specific hazard — absolute scoring doesn't have it, because there's no position. And there's a subtler finding on top: the standard reliability checks — position bias, transitivity, pairwise agreement — are misleading, because each is dominated by close-rank-gap pairs, where inconsistency is expected and errors cost little. Far-gap pairs determine the ranking accuracy but barely move the proxies. So you can have a judge that looks reliable by all the standard numbers and is fine exactly where it doesn't matter.



Corn

The metrics measure the pairs where disagreement is cheap.



Herman

Third, and this is the sharpest one, the identifiability problem. There's a line of work on separating judge quality from judge bias — one model gives each judge a reliability parameter, inferred from the comparisons alone, and it beats naive averaging. But a paper this year argues the quality-bias split is not identified by pairwise data, and the failure is exact. Across forty-eight real judge pools, the likelihood over that coefficient is flat — flat to zero point zero zero zero zero nats — and collecting twenty-six times more comparisons buys nothing.



Corn

Zero information about the thing you're debiasing.



Herman

A "debiased" score is selected by the prior you brought, not recovered from the data. So: absolute scoring has drift. Pairwise has cycles and a bias term you cannot separately identify. Neither is free. The trade is real, not a victory lap.



Hilbert

"Both about the same" is a coward's answer.



Herman

I'm sorry?



Hilbert

The four options Daniel used. Wrong. It should be three. A, B, and "I can't tell." Corn, you were close to it earlier when you split the indifference from the disgust, and then both of you walked past it. I sat in on a listening test for a tape machine, an actual one, in a room in Haifa. Eleven of us. Headphones on, pairs of reels, mark the sheet. The man next to me, Gideon, marked "about the same" on every pair where he could not hear a difference. Eleven. He could not hear a difference, he told me at the tea, since nineteen seventy whatever the machine was. But the sheet said "about the same," which on our scale sat one step above "slightly worse," and the analysis put that machine ahead of one it had no business beating. I kept his sheet. I still have it. The three-answer form — A, B, cannot tell — does not have this problem, and I have run it that way in my kitchen since. Herman, you said one bit. It is one bit only if the third answer is honest.



Herman

The discrimination test and the preference test as separate buttons.



Corn

Gideon couldn't hear the difference and the scale heard a preference.



Hilbert

Yes. And I would rather have a man tell me he heard nothing than tell me two reels matched. He heard nothing. That is a result. The kettle knows more about that than half the sheets did.



Corn

How long did the test run?



Hilbert

Four afternoons. Gideon came to all of them. Ate all the biscuits. That's the man for you.



Herman

The ABX answer — "cannot tell" — folded into the preference scale as its middle point. That's the whole confusion in one sheet of paper.



Hilbert

Keep the three answers separate and the biscuits to yourself.



Corn

He's right, though. And it's worse than he said.



Herman

Go on.



Corn

The endpoints have the same problem at the top of the scale. "Both are bad" — Daniel's honest button — the ITU scales don't have it, and Gideon's "about the same" shouldn't exist. Every answer set is a set of claims about what a listener can distinguish, and the traditions keep getting this wrong in different directions. Which, now that I say it, is the closing argument for the whole afternoon: the comparison question works because it asks almost nothing, and the answer set is where you quietly ask too much.



Herman

And there's one more result I want on the table before we land, because it's the whole argument for the method in one finding. A blind verification study on coding. Human-to-LLM agreement — overlap on what needed fixing — came out at about point three zero, well below the human-to-human agreement of point five two. Standard practice reads that as the LLM coding worse. Then a blind verifier preferred human and LLM coding at indistinguishable rates — fifty-one and a half versus forty-eight and a half, nowhere near significance — and a Bradley-Terry ranking placed two LLMs above two of the three humans.



Corn

The agreement metric said worse. The preference said same.



Herman

Same. The metric measured overlap in judgments; the preference measured what the verifier actually wanted to keep. Those are different questions, and we keep letting the first one stand in for the second.



Corn

The misconception, before we close. Mine's the obvious one: that this is something AI evaluation invented. Chatbot Arena, the leaderboard era, last few years.



Herman

It's Thurstone's 1927 law of comparative judgment, an ITU annex from 1996, and a statistical model from 1952 that Zermelo had already written for chess in 1929. The LLM leaderboard is the newest user of the oldest idea in the room. The mistaken part is thinking pairwise beats scoring universally — that's the halo we already punctured. It's a trade: calibration noise for one bit, and which side wins depends on the candidates and the rater, not on the method's pedigree.



Corn

The thread I'd leave dangling is Daniel's fourth button. "Both are bad" isn't in any scale, nobody has run his exact design, and whether it helps the ranking or quietly biases the Bradley-Terry fit is open. And the result that argues against pairwise on near-identical candidates comes from dialogue models, not audio — so his afternoon is real, and the literature can't yet explain it. He may be the first data point. If the names travelled between traditions, he'd never have had to reinvent the thing — but he did, and he got it mostly right, minus Gideon's button.



Herman

Three answers, Corn. A, B, cannot tell.



Corn

Thanks to Hilbert Flumingtop at the desk, producing this show as ever. If you enjoyed this one, try episode fifteen, AI Gets Personal; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode eleven, How Does Fine Tuning Work Anyway. This has been My Weird Prompts. If the show earns a review from you, we'd take it — the button's right there wherever you listen.



Herman

We'll be back soon. See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **ITU-T Recommendation P.800 (08/96), Methods for subjective determination of transmission quality [primary]**
and https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=3638, retrieved 2026-11-14.
https://www.itu.int/rec/T-REC-U.210/recommendation.asp?lang=en&parent=T-REC-P.800

2. **Wikipedia, Two-alternative forced choice**
last edited 2026-04-26, retrieved 2026-11-14.
https://en.wikipedia.org/wiki/Two-alternative_forced_choice

3. **Wikipedia, Pairwise comparison (psychology)**
(psychology), last edited 2026-10-29, retrieved 2026-11-14.
https://en.wikipedia.org/wiki/Pairwise_comparison

4. **Wikipedia, Law of comparative judgment**
last edited 2025-11-26, retrieved 2026-11-14.
https://en.wikipedia.org/wiki/Law_of_comparative_judgment

5. **Wikipedia, Louis Leon Thurstone**
last edited 2026-10-11, retrieved 2026-11-14.
https://en.wikipedia.org/wiki/Louis_Leon_Thurstone

6. **Potato annotation docs, Listening Tests: MUSHRA, ACR, P.808-Style Studies and ABX**
v2.12.0+ (mentions 2.12.1), retrieved 2026-11-14.
https://www.potatoannotator.com/docs/annotation-types/listening-tests

7. **When is it Better to Compare than to Score?**
(full text via ar5iv), retrieved 2026-11-14.
https://arxiv.org/abs/1406.6618

8. **Human Evaluation of Conversations is an Open Problem**
retrieved 2026-11-14.
https://arxiv.org/abs/2201.04723v1

9. **Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference**
submitted 2024-03-07, retrieved 2026-11-14.
https://arxiv.org/abs/2403.04132v1

10. **Investigating Non-Transitivity in LLM-as-a-Judge**
revised 2025-06-05, ICML 2025 Spotlight, retrieved 2026-11-14.
https://arxiv.org/abs/2502.14074v3

11. **Who can we trust? LLM-as-a-jury for Comparative Assessment (BT-)**
revised 2026-05-28, ICML 2026, retrieved 2026-11-14.
https://arxiv.org/abs/2602.16610v2

12. **Duel-Evolve: Reward-Free Test-Time Scaling via LLM Self-Preferences**
submitted 2026-02-25, retrieved 2026-11-14.
https://arxiv.org/abs/2602.21585v1

13. **LLM Evaluation as Tensor Completion**
revised 2026-09-03, retrieved 2026-11-14.
https://arxiv.org/abs/2604.05460v2

14. **Low Rank for Rank: Uncertainty-Aware Task-Specific LLM Ranking under Sparse Pairwise Comparisons**
revised 2026-09-02, retrieved 2026-11-14.
https://arxiv.org/abs/2605.29395v2

15. **Argument Quality Assessment with Large Language Models: A Pairwise Bradley-Terry Approach**
revised 2026-08-31, EMNLP 2026 Findings, retrieved 2026-11-14.
https://arxiv.org/abs/2605.28313v2

16. **When Can You Debias an LLM Judge?**
revised 2026-07-18, retrieved 2026-11-14.
https://arxiv.org/abs/2607.02104v2

17. **Agreement Is Not Quality**
submitted 2026-07-30, retrieved 2026-11-14.
https://arxiv.org/abs/2607.28890v1

18. **Pair Difficulty Matters: Rethinking Pairwise LLM-as-a-Judge Evaluation and Consistency**
revised 2026-10-27, EMNLP 2026 short paper, retrieved 2026-11-14.
https://arxiv.org/abs/2609.37577v3

19. **Sushant Gundla, Pairwise comparison and human review that scales**
retrieved 2026-11-14.
https://learn.sushantgundla.com/evals/human-review-that-scales

20. **Michael Brenndoerfer, Preference Evaluation: Pairwise Comparisons & Elo Rankings for LLMs**
retrieved 2026-11-14.
https://mbrenndoerfer.com/writing/preference-evaluation-pairwise-comparisons-elo-llm

21. **On Crowdsourcing-Design with Comparison Category Rating for Evaluating Speech Enhancement algorithms**
retrieved 2026-11-14.
https://ar5iv.labs.arxiv.org/html/2306.01538