Qeios

Peer Review

Review of: "Are Bidding Behaviors of Jeopardy! Gameshow Contestants Influenced by Gender/Sex?"

Andrea Morone[1]

1. Department of Economics, Management and Law, University of Bari, Italy

Summary

This manuscript asks whether contestants' Daily Double wagering in Jeopardy! differs by perceived sex/gender, using a retrospective dataset of episodes spanning 1984–2022. The core outcome is the wager expressed as a percentage of the contestant's "current amount," and the paper reports no statistically significant differences between male- and female-presenting contestants (overall, by Daily Double, and by decade).

The question is interesting, and the setting is appealing: Jeopardy! provides repeated, incentive-relevant wagering decisions in a naturally occurring environment. That said, several design and analysis choices currently make it difficult to interpret the results—especially any conclusion framed as "no difference."

Strengths

Motivating question with clear relevance to debates on gender and risk-taking.

High-visibility naturalistic setting with real stakes and repeated decisions.

The manuscript is **readable and well organized**, and the limitations section already gestures at key challenges.

doi.org/10.32388/XZEDG4

Major comments (most important to address)

1) Outcome definition conflicts with game rules (and dropping >100% is consequential)

The paper defines risk-taking as *wager / current amount* and **omits observations above 100%**. However, wagers above 100% are not anomalies in Jeopardy!; they arise mechanically from the rules (e.g., minimum wager thresholds such as $1,000/$2,000 or current score, whichever is larger). Dropping these observations risks:

Non-random selection (you disproportionately keep cases where contestants have higher scores),

Distorting the distribution by removing some of the most aggressive/strategic wagers,

Making the "risk" interpretation less faithful to how contestants actually face constraints.

Suggested fix (high priority):

Use a **rule-consistent denominator**, e.g., `wager / max_allowed_wager` (where max_allowed depends on the round and the "$1,000/$2,000 or score" rule).

Alternatively, model the wager amount directly with the correct constraints, or treat the outcome as **censored/limited** by rule-based bounds rather than deleting valid observations.

2) Statistical independence is unlikely (episode and contestant clustering)

Daily Double observations are plausibly clustered:

within **episodes** (shared opponents, board, pace, strategy, context),

and possibly within **contestants** (repeat champions; stable individual tendencies).

Simple independent-sample t-tests can therefore yield **miscalibrated inference**.

Suggested fix:

Use a regression framework with **cluster-robust standard errors** (at minimum by episode), and if identifiers allow, consider **mixed-effects models** (random intercepts for episode and contestant).

3) Bounded, non-normal outcome: t-tests/ANOVA may be a poor match

A wager ratio (or percent) is **bounded** and likely not normally distributed. It may also show mass points at salient strategies (e.g., "true Daily Double," round numbers), violating assumptions behind t-tests and standard ANOVA.

Suggested fix:

Consider **fractional response models** (e.g., fractional logit) if using ratios, or **beta-type models** if appropriate (with proper handling of 0/1 boundaries).

Report **effect sizes and confidence intervals**, not only p-values.

4) The ANOVA design (decade as "within-subject") seems conceptually mismatched

The manuscript describes decade as a within-subject factor. Unless the same unit is repeatedly observed across decades (rare), decade is typically **between-unit**, not within-unit. This threatens the interpretability of the reported F-tests.

Suggested fix:

Treat year/decade as **between** (categorical) or **continuous** time trend with an interaction with sex/gender, estimated in the main regression model.

5) Interpreting wagers as "risk preferences" requires situational controls

Daily Double wagers are not pure "risk attitude." They are strategic choices that depend heavily on:

score relative to opponents (lead/lag, distance to leader),

round (Jeopardy vs Double Jeopardy),

timing within the round,

perceived probability of answering correctly (category strength / skill proxies).

Without controlling for these state variables, it's difficult to interpret differences (or non-differences) as risk preference differences.

Suggested fix:

Include key situational covariates: current score, distance to leader/second, round indicator, perhaps a proxy for contestant performance/confidence (e.g., correctness rate up to that point if reconstructible).

Consider stratifying analyses by "being ahead vs behind" or by "close game vs not," since strategic incentives change sharply.

6) Sampling frame and reproducibility need strengthening

Episodes were selected based on availability (specific platforms) and subsampling rules. For a strong empirical contribution, readers need:

an **explicit episode list** (and what was excluded and why),

clear **random sampling procedure** (and seed),

handling rules for missing or ambiguous information,

at least a minimal assessment of **coding reliability** (inter-rater agreement) for a subset, especially given manual coding and "perceived sex/gender" classification.

Suggested fix:

Provide a supplementary file with episode identifiers, variables, and code.

Double-code a subset and report agreement (e.g., kappa for gender coding; ICC for wager/score extraction).

7) "No difference" claims should be supported by precision/equivalence, not only non-significance

A non-significant p-value does not establish "no effect." If the intended message is that differences are negligible, the paper should show that the data rule out effects of practical relevance.

Suggested fix:

Report **confidence intervals** for group differences.

Consider **equivalence testing (TOST)** with a pre-defined smallest effect size of interest (SESOI).

Minor comments / clarity improvements

Be consistently explicit about the meaning of **"perceived sex/gender"** and avoid slipping into language that implies biological sex if the variable is presentation-based.

Clarify exactly what "current amount" means (score immediately before the DD? any adjustments?).

Add **Ns in figure captions** for each group and each Daily Double/decade bin.

If multiple comparisons are conducted (overall + by DD + by decade), state whether the analysis is exploratory and how multiplicity is handled.

What would make this paper much stronger (concrete checklist)

Redefine the outcome to respect Jeopardy!'s wager constraints (do not drop >100% cases).

Re-estimate using an appropriate model for bounded outcomes + clustering by episode (and contestant if possible).

Add key strategic controls (lead/lag, round, closeness of game, skill proxies).

Provide transparency materials: episode list, sampling method, code, and basic coding reliability.

Emphasize effect sizes, CIs, and equivalence/precision when discussing "no differences."

Overall assessment

The manuscript has a promising question and a potentially strong setting, but the current empirical implementation makes the conclusions hard to interpret. I would encourage a substantial revision focused on rule-consistent measurement, appropriate modeling, and situational controls. With those changes, this project could become a useful contribution.

Declarations

Potential competing interests: No potential competing interests to declare.