

Reviewer-1:-

While the manuscript is framed around a very important topic, cultural / contextual consideration for “normative” practices in XYZ domain (i.e., parenting in this case; spanking, specifically), the study design and data employed rigorously explore such a nuanced, complex topic and research question.

We thank Reviewer 1 for the thoughtful and constructive feedback. Below, we provide detailed responses to each comment and describe the changes made to the manuscript.

The title is sufficient; however, greater specificity would be beneficial (see highlighted) – “Moderators of maternal spanking and externalizing problems in children”. The article has good English quality and is well-written.

The title has been revised for clarity and specificity. The new title is: **“Moderators of the Association Between Maternal Spanking and Externalizing Behavior in At-Risk Black and White Children.”**

Greater detail in the Methods/Results sections is needed. Additionally, the Discussion could benefit from further discussion of confounds related to the study /date itself. For example, in the Methods, including a description of exactly how “caregivers reported on the race and gender of the child” is necessary (e.g., provide the exact item wording), given these variables are a main study focus. Furthermore, how was parental race assessed? This information is needed. For “child gender,” were parents asked about the child’s biological sex or “gender,” specifically? Or, as is often the case in research, was it presumed based on the child’s outward physical appearance? This information is important because “gender” is not the same as “biological sex” (e.g., chromosomal make-up).

We added the exact wording for demographic items, including those assessing child race, caregiver race, and child gender. For child gender, we clarified that the item is presumably referring to biological sex and included the historical context regarding gender constructs at the time the study was initiated.

Similarly, it would be beneficial for the reader if the “spanking” item (i.e., exact wording of question) was provided, as this would allow for greater accuracy of the current study’s conceptualization (e.g., for comparison with other “spanking”

conceptualizations). (Related, it would be beneficial to explore the validity of this item for spanking, or describe how it is indeed valid, given it is a sole focus of the study. Currently, on p. 5, the authors describe the validity of the CTS-OC scales but that is not the same psychometric information as the single item itself.)

Thank you for your thoughtful feedback. We have addressed this by adding the exact wording of the spanking item in the **Measures** section to enhance the study's transparency and facilitate comparisons with other studies. We also added acknowledgment of the limitations of using a single-item measure and its potential impact on validity. We hope these additions address the reviewer's concerns and provide a clearer conceptualization of the spanking measure used in this study.

For the Results section, while the Tables present this information, including the statistical findings (e.g., r statistics and accompanying p values) in-text along with the narrative description (p.6 / 7) for the significant findings is needed (per APA requirements / research reporting and publishing standards).

In the results narrative, effort has been made to elaborate and clarify the range of findings presented in Tables, now A4 and A5.

For example, provide the significant statistics for the spanking – externalizing correlations. Further, the reader is not provided descriptive statistics for Externalizing Behavior at each age (e.g., mean, standard deviations) – this would be very helpful, given it would provide insight into the data distributions themselves.

Descriptive statistics has been provided for Externalizing Behaviors across ages 4, 6, 8, and 12. See the new **Table A2**.

This comment is also important in regard to the analyses involving the “clinical range on the CBCL” only (p. 7). The authors simply say, “the same pattern is also found,” but provide no data. How many children met “clinical range” T-scores for the CBCL? Was it many? Only a few?

Additional data has been provided for the clinical range CBCL Externalizing Behavior score distribution.

Might the relationship (in both samples, full and “clinical-only”) be driven by a handful of children with outlier externalizing scores? Or outlier households with high levels of reported spanking?

Notably, none of the distributions of externalizing behavior scores were notably skewed (i.e. skewness/kurtosis values consistently between +/-1.0 at each age). While a few outliers were detected at the high end of the scale, this is expected for those scoring in the clinical range relatively to the rest of the sample.

More information about the data distributions are needed for the reader to better assess the quality of the data. Providing more narrative detail (in-text) about the results of the moderation analyses (e.g., what values are being provided in “Spanking x Race” column in Table A5) is needed. The moderation analyses are not well-described.

We have edited/revised this information to try to improve the clarity of the narrative pertaining to the moderation analyses. In addition, sub-headings have been inserted into the results narrative to further improve the clarity of presentation.

Furthermore, in the Results, added discussion about the “practical” significance (e.g., effect sizes) of these findings would be helpful – e.g., while significant, correlations are “weak” (Table A4) and differences across Black and White households (Table A2) are “small” (eta squared); therefore, perhaps, although statistically significant, the “effect” is not all that important (i.e., “small” effect “size”) or clinically meaningful.

The practical or clinical significance of the effects obtained is extremely difficult to determine with the data available. This consideration was added to the **Discussion** section.

Finally, given the multiple analyses run to examine multiple relationships at varying ages, did the authors perform any statistical correction(s) to avoid Type I or Type II error? Perhaps the findings are not truly significant (Type I) or not truly

not significant (Type II error)? In Table A4, with adjusted p-values, would those relationships still be significant at the... .01 level, for example?

Virtually all of the significant correlations, for example, were significant at $p < .001$; some at $p < .01$. Nevertheless, it is possible that some of the otherwise significant effects obtained might not be regarded as significant if Bonferonni-type adjustments were made. We have noted this in the manuscript and recommended caution in interpreting these results.

Last, in the Discussion section, there seems to be no discussion of the (potential in)accuracy of reporting by White households – that is, the possibility that “White households” are dishonest in their reporting of spanking (due to stigma) and spank more than reported. Could this be a confound at play? (Generally speaking, it seemed assumed that spanking was accurately being reported.)

Thank you for raising this critical point regarding the potential inaccuracy of spanking reports. To address this, we have revised the **Discussion** section to explicitly acknowledge this limitation and explore its implications for the interpretation of our results.

Two minor mistakes that the Reviewer wishes to point out for the authors – For example, “Achenbach (1991)” from in-text was mis-cited in the References section as “Achenbach (1999).” This needs corrected – Achenbach 1999 is not the correct CBCL reference. Additionally, in-text, Mackenzie et al 2011 is cited but no reference provided in the References. Additionally, while all references seemed proper recent and sufficient, as an additional comment...

Thank you for catching these errors in our references. We have corrected the citation for Achenbach to reflect the proper reference for the Child Behavior Checklist (CBCL), which is *Achenbach (1991)*. Additionally, we have reviewed the manuscript and updated the References section to include the missing citation for Mackenzie et al, now 2012 to reflect the journal publication date as opposed to the online publication date.

With any controversial, nuanced and multifaceted topic, there is always some concern about biases, including whether “all sides” and “all voices” on the topic

are being represented (e.g., confirmation bias, cherry picking), especially given that the literature itself is inherently biased and rooted in historical, systemic oppression.

While no ethical issues are explicitly noticed, it is true that the authors explore a highly controversial topic and provide findings that (could potentially) be groundbreaking and/or add further noise to this controversy – in this regard, it feels especially important to ensure that the study conducted is of sound, rigorous high-quality scholarship.

We appreciate the reviewer's thoughtful observations regarding the importance of addressing biases, representing diverse perspectives, and ensuring the rigor of our study on such a nuanced and controversial topic. These points have been carefully considered and integrated into the manuscript to strengthen its scholarship and transparency.

First, we acknowledge that spanking is a culturally and historically charged topic, with systemic biases shaping much of the existing literature. To provide context for our study, we have added a section in the *Introduction* explicitly addressing these biases and their potential influence on how corporal punishment has been conceptualized and studied. This inclusion helps frame our findings within a broader understanding of how cultural and societal norms intersect with disciplinary practices.

Second, we have taken steps to ensure that all relevant perspectives are represented, striving for a balanced and comprehensive review of the literature. Our discussion incorporates findings that demonstrate both positive and negative associations between spanking and child outcomes, as well as studies reporting null effects.

We also emphasize the methodological rigor of this study, recognizing the critical need for high-quality scholarship when exploring sensitive topics. Our approach includes the use of validated measures, careful control of confounding variables (e.g., histories of physical abuse), and transparent reporting of limitations, such as the use of single-item measures for spanking and the potential for social desirability bias in self-reports. These limitations are discussed in the **Discussion**

section, where we advocate for future research using multi-item measures and alternative methodologies to further enhance reliability and validity.

We are committed to upholding ethical research standards and advancing a respectful and inclusive dialogue on this important topic. We appreciate the reviewer's feedback, which has been invaluable in enhancing the clarity, depth, and balance of our work. We hope the revisions, in light of the word limit, reflect these efforts and provide a meaningful contribution to the field. Please let us know if further clarification is needed.

Reviewer-2

While the manuscript is written clearly and easy-to-follow, and the authors are applauded for having the start to an interesting project, my recommendation is to "major revision" this manuscript.

We appreciate the reviewer's thoughtful and detailed feedback, which has provided valuable guidance to improve our study's rigor, clarity, and contextualization. We value the opportunity to revise this work and will address the edits below.

This paper explores the relationship between spanking and externalizing behavior at varying ages, as moderated by "race" and "(child) gender." Investigating potential "cultural" and "contextual" differences in salient correlates of problematic behaviors in childhood, like the current study attempted to do, represents a very important and ever-timely topic. However, this domain is also incredibly nuanced and complex, which thus necessitates great thoughtfulness regarding the importance of ensuring the scholarship conducted is of the highest quality possible.

We agree that the topic of spanking and its relationship to externalizing behavior, particularly as moderated by race and child gender, is indeed nuanced, complex, and critically important. This study aims to contribute to the understanding of these relationships while recognizing the challenges inherent in addressing such a multifaceted topic. We have incorporated some additions to the manuscript that we hope continues to reflect this thoughtfulness and commitment to quality.

This research is well-aligned with the scope of The Journal of Mental Health and Social Behaviour.

The manuscript is well-written and explores an interesting topic, with hardly any hiccups (e.g., APA-style formatting issues, typos) (see Minor Comments, below). The Introduction and Discussion sections, specifically, are thoughtfully presented and (seemingly) theoretically supported – although, of course, there are always concerns about whether Authors "cherry picked" studies to tell "their story." That

said, concerns arose about whether the literature review was biased at all, given the inherent biases of the social science literature itself (i.e., systemic racism and oppression within scientific reporting and publishing), it is difficult to ensure that “all” voices / perspectives are represented and avoid committing “confirmation (selection) bias.”

We appreciate the acknowledgment of the manuscript’s clarity and the thoughtful presentation of the **Introduction** and **Discussion** sections. We fully agree with the concerns regarding inherent biases in the social science literature, particularly related to systemic racism and oppression, which can influence how corporal punishment is studied and interpreted. To address this, we ensured the inclusion of studies with varying findings to represent a broader spectrum of perspectives and outcomes. We have explicitly acknowledged these biases in the introduction to emphasize their potential influence on the conceptualization and interpretation of spanking research. By incorporating this revision, we aim to increase transparency and demonstrate our commitment to presenting a balanced and critically reflective perspective on this important topic.

Despite these strengths, I have major concerns regarding the sufficiency of the research design, data employed, and, consequently, the methodological rigor of the study itself. While these comments are not exhaustive, efforts were made to provide specificity for guidance. Please see “Comments regarding Appropriateness” (textbox above) and “Major Comments #1-4” (below). Therefore, although the overall presentation (extant literature; supporting theoretical framework) is strong, the design of the study and the research methodology employed, as well as the presentation of the data/analyses, do not meet the level of rigor and quality needed to thoroughly investigate such a nuanced, complex topic, like “cultural” racial* differences in parenting.

Major comments:

1. Operationalization / Conceptualization Issues

We appreciate the thoughtful feedback regarding the operationalization and conceptualization of “race” in our study. Below, we address these concerns and outline the steps we have taken to improve the clarity and contextualization of this critical variable:

i. Race

1. Unclear how “race” was measured beyond “parent-report.”

We acknowledge the concern about the lack of clarity regarding how “race” was measured. In the revised manuscript, we have explicitly clarified that race was assessed through caregiver self-report as part of the LONGSCAN protocol. The relevant demographic items allowed caregivers to select their racial or ethnic background from predefined categories. This information has been detailed in the

Materials and Methods section to provide greater transparency about the measurement process.

2. Binary (Black vs. White) categorization oversimplifies the lived experiences and systemic inequities faced by Black families.

We agree that the binary categorization of “Black” versus “White” oversimplifies the rich diversity of experiences within and across racial groups, including the systemic inequities that shape parenting practices and outcomes. However, our decision to limit the analysis to Black and White households was guided by the structure of the LONGSCAN dataset and the need to focus on meaningful comparisons within the scope of this study. In the **Limitations** section, we now explicitly address the constraints and limitations of this binary categorization, acknowledging that it does not capture the full complexity of racial identity, systemic inequities, or the diverse lived experiences of Black families.

We have also added a discussion in the **Introduction** and **Discussion** sections emphasizing the importance of systemic and historical contexts, particularly as they pertain to racialized parenting practices and child outcomes. By situating our findings within the broader framework of systemic inequities, we aim to avoid oversimplifying the relationships we examine and provide a more nuanced interpretation of the results. We believe these revisions improve the clarity and rigor of our manuscript and ensure a more thoughtful and context-sensitive treatment of “race” as a construct. Thank you for highlighting this important area for improvement.

ii. Spanking

1. Single-item measure with no exploration of context or nuance - undermines the depth of the investigation.

We acknowledge that the use of a single-item measure for spanking limits the depth of the investigation, as it does not capture the broader context, severity, or nuances of this disciplinary practice. To enhance transparency, the revised draft now includes the exact wording of the item used to assess spanking in the **Measures** section, as well as an explicit discussion of its limitations in the **Discussion**.

2. To better understand the data distributions, reporting additional statistics (beyond mean and stand deviation) would be helpful

Please see the added table of descriptive statistics (**Table A2**) and edits to corresponding narrative.

b. Limited Scope: “Cultural” and “Contextual” Factors

i. Weak support for framing the study as examining "cultural" differences

1. On page 9/10, paragraph 5 - the authors acknowledge that within-group comparisons (e.g., variability among Black households) might provide more meaningful insights than cross-racial comparisons.

ii. Framing the study as “racial” differences may be more accurate than describing it as “cultural” (although “race” may also represent an overly simplified “conceptualization” of a complex, nuanced “construct”)

We thank the reviewer for this recommendation and have incorporated the addition of identify-based cultural considerations and race.

c. Ambiguity: “Gender” Measure

i. Unclear how "gender" is measured. Is it assumed? Was it actually “biological sex” being assessed? The distinction is critical, given that “gender” and “biological sex” are not the same thing.

We agree that the distinction between “gender” and “biological sex” is critical and appreciate the reviewer highlighting this important point. In the revised manuscript, we have clarified that the measure of "gender" in the study likely reflects the caregiver's report of the child's biological sex, as the dataset used for this research (LONGSCAN) was collected in the early 1990s. At that time, research instruments commonly employed binary categorizations for gender, equating it with biological sex as perceived by the caregiver. To address this ambiguity, we have added the following clarification to the **Materials and Methods** section under the **Demographic Information** subsection.

2. Research Design/Methodological Weaknesses

a. Simplistic Design

i. Overall, the design/analyses are too simplistic for such a nuanced topic

We acknowledge that our study design and analytical approach may not capture the full complexity of the topic. However, we believe that the use of longitudinal data from the LONGSCAN dataset, while limited, provides valuable insight into the relationships among spanking, externalizing behaviors, and moderators such as race and gender. To address this concern, we have explicitly acknowledged the limitations of our analytical approach and the potential benefits of employing more advanced methodologies in future research. These limitations and suggestions for future research are included in the revised **Discussion** section.

b. Limited / No Contextualization:

i. Black and White families are compared without accounting for systemic inequities that may shape parenting practices and child outcomes - this "apples-to-oranges" comparison risks perpetuating biased narratives.

We agree that systemic inequities and historical contexts play a crucial role in shaping parenting practices and child outcomes. We include explicit discussion of how systemic factors, such as structural racism, socioeconomic disparities, and

societal norms, may influence both the prevalence of spanking and its associations with externalizing behaviors. Additionally, we have clarified in the *Introduction* and *Discussion* sections that the study findings should not be interpreted as direct comparisons of parenting practices across racial groups in isolation but rather as an exploration of how these practices are moderated by broader systemic and cultural influences. Furthermore, we have emphasized the need for caution in interpreting these results, acknowledging the potential for biased narratives and the limitations of binary racial categorizations. These revisions aim to contextualize the findings more effectively and highlight the importance of examining systemic inequities in future research.

3. Statistical Concerns

a. No Corrections (for Multiple Analyses)

i. No corrections for family-wise error rate (FWER) or false discovery rate (FDR) were applied, increasing the risk for Type I/II errors

1. Corrected p-values should be calculated and reported

a. Would the findings still be significant at the .01 level?

Virtually all of the significant correlations, for example, were significant at $p < .001$; some at $p < .01$. Nevertheless, it is possible that some of the otherwise significant effects obtained might not be regarded as significant if Bonferroni-type adjustments were made. This has been acknowledged in the revised manuscript.

b. Insufficient Statistical Reporting

Statistical narrative results have been included.

We thank this reviewer for the constructive feedback and hope these revisions, in light of the word limit, address the reviewer's concerns and enhance the overall quality of the manuscript.