Dear Editor Adam Osth,

We appreciate the opportunity to revise our manuscript and thank the reviewers for their insightful comments, which helped us sharpen the theoretical positioning and practical relevance of our contribution. In response, we conducted an extensive revision aimed at clarifying the role and advantages of significance testing based on the Overlapping Index with permutation inference in psychological research.

The manuscript has been substantially strengthened by expanding the theoretical framing and by explicitly situating the method within core research practices in experimental, cognitive, clinical, and developmental psychology. We now more clearly demonstrate how distribution-level analyses can reveal effects that remain obscured when analyses focus exclusively on mean differences.

Each reviewer comment has been carefully addressed. We moderated claims where necessary, provided additional methodological clarification, and implemented a transparent, reproducible workflow accompanied by a real-data example, improving both clarity and usability of the proposed approach. These revisions strengthen both the methodological rigor and the clarity of presentation, making the manuscript more accessible to a broad psychological audience.

We thank you for your time and consideration and believe the revised version offers a stronger and more precise presentation of the contribution and its relevance for psychological research.

With best regards,

Ambra Perugini

on behalf of all co-authors

Reviewer 1

The article introduces the overlapping test which simultaneously test for differences in means, variances, and shape within a single test. The authors argue for several benefits gained from using the overlapping test over more frequently used analyses (e.g., Welch t-test, Wilcoxon-Mann-Whitney test, Kolomogorov-Smirnov test) and conducted a simulation study. Overall, I think that the article could be a good contribution to the literature. However I do have some serious concerns which would require extensive revisions. I will outline my concerns below:

Major comments:

1

The authors argue on page 13: “Its main strengths lie in its easy interpretability as an effect size”. However, I am currently not convinced. Effect sizes should be mainly used to interpret the magnitude of a phenomenon and to infer the practical or theoretical relevance. However, the η index provides the amount of overlap between two or more empirical density functions. How should researchers infer what a meaningful overlap is between two or more empirical density functions? I don't think that this is an intuitive effect size, at least not how it is currently presented in the article. Some real data examples would be beneficial in interpreting this.

Response 1

We thank the reviewer for the observation regarding our claim that the η index offers easy interpretability as an effect size. We understand the concern: η, as currently presented, may not fit the conventional expectations of an effect size that directly quantifies the magnitude of a mean difference or its theoretical/practical relevance (e.g., as Cohen's d does).

However, we argue that η serves a preliminary and meaningful role, particularly when interpreted within the framework of global tests. The ζ-Overlapping test (based on $ζ = 1 − η$) can be intended as operating as a global test, which addresses a very general null hypothesis: that two distributions are the same in all aspects - mean, variance, skewness, and shape. This contrasts with traditional tests that target specific parameters (e.g., mean difference in a t-test). In this context, η serves as a nonparametric distribution-level effect size that quantifies the degree of similarity between two empirical distributions. Such a test is especially appropriate in psychology, where data distributions often deviate from ideal assumptions (e.g., normality, homogeneity of variances), and researchers are frequently interested in more nuanced group differences.

To strengthen this framework, the revised manuscript provides and clarifies examples that highlight the utility of η/ζ as a global effect size across various psychological fields. Please find the changes marked in ted in the revised manuscript (pag. 2)

2

I wonder how many researchers have a main interest in the distribution of their data. That is, oftentimes researchers simply want to examine what the difference are between two means. In this case, as the authors correctly point out, the researchers should take into account the assumptions underlying their statistical test.

I think that the overlapping test would force them to look at how their data is distributed which is good. However, it is currently lacking what the researchers should do after they found a statistical significant effect after conducting an overlapping test because a significant effect could mean that the means are different, or the variances, or the shape. Should researchers conduct follow-up tests such as an independent sample t-test? In this case, the overlapping test is not necessary but researchers simply need to examine the assumptions underling their test. It would be good to provide more information on this.

Response 2

We agree with the Reviewer that more information were needed in the previous version about this important point. Please find in the revised discussion the introduction facing the issue about how many researchers focus on differences in means and may not initially consider the full distribution of their data. One of the strengths of the ζ-Overlapping test is precisely that it encourages a more holistic view, prompting researchers to explore whether groups differ not just in central tendency but in dispersion or shape as well.

We fully agree that our original manuscript lacked specific guidance on what researchers should do after obtaining a significant ζ result. We have now revised the manuscript enriching it with a step-by-step workflow that clarify the specific role played by the proposed significance test. Thanks to the comment of the Reviewer the revised version of the manuscript now clarify that the ζ-Overlapping test should be seen as a first-line, global diagnostic tool, rather than a replacement for targeted tests. When the ζ test returns a significant result, it indicates that something differs between the distributions, but it does not specify what differs.

Therefore, we recommend the following steps:

First the researcher should visually inspect the data, already from data visualization it becomes clear how the two groups/conditions may differ. Then it follows hypothesis testing. In case of a statistically significant result we recommend follow up analysis only if relevant for and justified by the researcher.

3

The authors state that on page 2: “The rationale behind these steps involves first introducing the concept of the Overlapping Index (η), which is crucial because it provides an intuitive measure of similarity between distributions by quantifying the overlapping area of their empirical density functions, a common question in quantitative psychology.”.

Evidence that indeed shows that this is a common question in quantitative psychology is necessary. In my personal experience, I have not seen this as a common question. Another possibility is to argue why this should be a common question.

Response 3

We thank the Reviewer for raising this important point. We examined the citation network of Pastore & Calcagnì (2019), the seed and seminal paper introducing the Overlapping index (η), and found that as of February 2026, it has garnered over 370 total citations, with more than 120 in psychology-related journals alone, reflecting widespread adoption across disciplines.

Recent citations span neuroscience, psychometrics, cognitive science, developmental psychology, and educational research, underscoring η's utility beyond descriptive statistics.

For instance, in neuroscience, Yamauchi et al. (2025) applied η to quantify distributional overlaps in neural reward monitoring signals from macaque frontopolar cortex, revealing subtle differences in response variability during decision-making tasks. In cognitive and behavioral measurement, (DOI: 10.3758/s13428-025-02918-6) cited the Overlapping Index to evaluate distributional similarities emphasizing η's role in assessing non-normal data overlaps across in lab and online data collection. In developmental and clinical psychology (Spaggiari et al., 2024). In language and educational psychology, Conversano et al. (2024) quantified reading behavior overlaps between monolingual and bilingual learners, demonstrating η's sensitivity to shape and tail differences.

These applications across neuroscience, experimental, developmental, cognitive, and clinical fields affirm η's substantive value for analyzing variance, skewness, and tails in behavioral data, likely key for psychological inference. We thank the reviewer that allowed us to highlight this fundamental aspect. Please find the changes marked in red in the revised Manuscript (end of Introduction, page 2).

4

One of the main benefits that the overlapping test brings is that it doesn't rely on the assumptions of normality and homoscedasticity. However, these assumptions are mainly problematic in traditional significance testing (e.g., t-test) with small sample sizes as indicated on page 3. How much benefit is gained from teaching researchers this new test when more and more frequently researchers use larger sample sizes for which these assumptions are not that problematic. Would a simple focus on visualizing the data not suffice?

Response 4

We appreciate this thoughtful question. It is true that many of the assumptions underlying classical tests (e.g., normality, homoscedasticity) become less problematic as sample sizes grow, due to the central limit theorem and increased power. Even though online data collections and multi-lab projects are facilitating sample size increase, there are some fields in which such growth remains limited (De Pieri et al., 2025; Thielemann et al., 2022) for many reasons (i.e. clinical trials, low prevalence populations, MEG and lab studies), and in such studies it remains crucial to carefully evaluate assumptions. However, the value of the overlapping test - particularly in the form we propose, with permutation-derived p-values - goes beyond merely circumventing assumption violations in small samples.

As clarified in the manuscript (see page 3 and Figure 3), the core strength of the ζ-based test is not limited to small samples, but rather lies in its global, distribution-aware nature. The test provides a statistically grounded means of assessing whether two (or more) conditions or groups differ in any distributional property, not only in the mean. It is very common practice in psychology to assess no gender differences or no difference between some conditions of an experiment in a sample before proceeding to more tailored complex analyses. In such cases multiple tests are run inflating type I error (See supplementary materials in the updated OSF repository). In this framework, it is extremely useful to test a global H0, rather than specific parameters.

Moreover, while visual inspection is indeed crucial (and we fully agree with the reviewer that visualizing distributions should be encouraged), it remains inherently subjective. Our permutation-based test offers a complementary inferential tool: it quantifies and tests whether the observed overlap is unlikely to have occurred by chance. This can be especially helpful in large-scale psychological studies with complex, multimodal, or skewed data (as shown in our real data examples), where visual interpretation alone may be insufficient or inconsistent across observers.

We believe this method provides a more systematic and interpretable extension to visual tools, aiding researchers not only in exploring data but also in reporting it rigorously and transparently, which aligns with current trends in open and reproducible psychological science.

Please find the changes marked in red in the revised manuscript (page 15):

5

Also, the authors indicate that the assumptions of normality and homoscedasticity are oftentimes not met in psychological research. It would be good to provide evidence for this especially because the benefits of the overlapping test rests on this.

Response 5

We thank the Reviewer for raising this critical point. We agree that demonstrating empirical evidence for assumption violations is essential to justify the relevance of the overlapping test. Our revisions now explicitly integrate both references within the manuscript and widely acknowledged methodological critiques in psychological science.

As discussed in the manuscript (pp. 2–3), reaction times in cognitive tasks are characteristically right-skewed. The ex-Gaussian distribution—which combines a normal and an exponential component—has long been proposed as a more appropriate model for RT data (e.g., Lacouture & Cousineau, 2008; Spangler et al., 2021). A substantial body of work shows that applying standard *t*-tests or ANOVAs to raw RTs can be misleading, often motivating transformations or outlier trimming merely to approximate normality. More broadly, methodological developments in cognitive psychometrics increasingly emphasize distributional modeling, precisely because classical statistics may fail when applied to skewed data, even with large samples (Heathcote, Popiel, & Mewhort, 1991; Rousselet & Wilcox, 2020).

Meta-analytic evidence further indicates that only approximately 5.5% of psychological datasets approximate normality (skewness and kurtosis between −0.25 and +0.25), even in moderate sample sizes (Blanca et al., 2013). Consequently, standard normality tests (e.g., Shapiro–Wilk, Kolmogorov–Smirnov) frequently reject the normality assumption in realistic datasets, highlighting a systemic mismatch between statistical assumptions and empirical psychological data. Similar concerns arise in electrophysiological research, where ERP components such as P300 amplitudes and latencies often display skewed or heavy-tailed distributions due to individual differences, trial variability, and preprocessing factors; methodological sources (e.g., Luck, 2017) explicitly caution that classical parametric tests may produce biased inferences under these conditions.

We have revised the manuscript accordingly to make these points more explicit. These additions clarify how the Overlapping Index combined with permutation testing can provide informative insights even when applied to existing datasets, including large-scale data collections already available to researchers. By evaluating full distributional differences rather than single parameters, the approach can reveal effects that may have remained unnoticed—or theories that may have been prematurely rejected—because conclusions were previously based on isolated summary statistics. These revisions therefore make more evident both the practical utility and the theoretical implications of adopting a distribution-level perspective in psychological research.

Please find the changes marked in red in the introduction and discussion sections

6

The authors report the following on page 1: "In fact, there is a growing caution against blindly using statistical tools and analytical methods without a deep understanding of their assumptions and implications (Scheel, Tiokhin, Isager, & Lakens, 2021). However, the main point of the article of Scheel and colleagues is that hypothesis testers should spent less time hypothesis testing and conduct more foundational research. However, the current article seems to be contradicting this by simply conducting a different type of hypothesis tests with fewer assumptions. However, this does not take away the issues reported by Scheel and colleagues. I think that the overlapping test/approach could be useful in the way that Scheel and colleagues argued but this needs to elaborated.

Response 6

We agree with the Reviewer that the work of Scheel et al. (2021) emphasizes the need to move away from excessive reliance on traditional hypothesis testing toward more foundational and theory-driven research. In our article, we explicitly acknowledge this shift, and, indeed, our intent is proposing a new global hypothesis test that may fit better theory-driven hypotheses that address complex phenomena that can be rarely reflected by solely considering the mean parameter. We agree with the reviewer that our initial phrasing could be misleading and not fully in line with the ideology of Scheel et al. (2021) and we rewrote the paragraph to emphasize how our work is actually in line with Scheel et al. (2021) by offering a more transparent, assumption-light, and effect-size-focused method for comparing empirical data distributions. Thank to the indication prompt by the reviewer's comment, in the revised version of the manuscript the ζ-overlapping test is clearly presented as a way to foster deeper understanding of data by examining the full shape and overlap of distributions, rather than relying solely on mean comparisons or p-values under strong assumptions (e.g., normality, homoscedasticity). We acknowledge that this connection could have been more clearly articulated in the first version of the manuscript. In the revision, we elaborated on how the ζ-overlapping test supports the broader goals promoted by Scheel et al. - namely, enhancing transparency and focusing on meaningful data patterns rather than mechanistic testing. We also clarified that this method is best viewed not as a replacement of one hypothesis test with another, but as a conceptual and practical step toward richer, theory-aligned inference.

Please find the changes marked in red in the revised introduction.

7

For the simulation study, it seems that the overlapping test is the most powerful one. However, is this not simply driven by the fact that it examines statistical significant differences between the means, variances, and shapes? In other words, is it not simply more powerful because it has more chances to find statistical significant effects compared with the more focused tests (e.g., mean differences). That is, the overlapping test seems to just be a broader test but not more powerful (no simulation was necessary for this as this could have been derived logically). In this case, the only somewhat fair comparison seems to be the Kolmogorov-Smirnov test. This goes along with the question about what the “H0 false” means in figure 7. Is this that there is a difference between means? Difference between variances? Both? If it is a difference in means then it is H0 = false for an independent sample t-test while the H0 = true for the F test when variances are set to be equal. Hence, the comparisons seem to be unfair but maybe I am mistaken. Either way, it would also be interesting to examine how well the overlapping test performs compared to more specific tests (e.g., t-test) when the assumptions hold? This is currently unclear to me from only Figure 7. More information is necessary here.

Response 7

We appreciate the reviewer’s reflection on the interpretation of the simulation results. We agree that the ζ-overlapping test, by design, is sensitive to multiple distributional properties (mean, variance, and shape), which inherently gives it a broader scope than tests focused on a single parameter. However, we believe this broader sensitivity is precisely what makes the test practically valuable for a broad range of psychologists, from experimental to applied psychological research, where violations of parametric assumptions are common and the exact source of manipulation checks and group differences is often unknown. In response to the reviewer’s suggestions, we have taken the following actions in the revised manuscript:

Clarified the definition of “H_0 false” in Figure 7 to explicitly state what kind of distributional difference is present in each scenario. Expanded the simulation discussion to highlight that the overlapping test is more comprehensive, not necessarily more powerful in a strictly comparable sense, and is best used when researchers want to assess any type of difference, not just location shifts. Added a new paragraph to reflect on comparisons with specific tests (e.g., t-test, F-test) under ideal conditions, showing that the ζ-overlapping test performs comparably when assumptions are met, and robustly when they are not.

Please find the changes marked in red in the revised text just after Figure 7

8

This also explains the correlation matrix which seems uninformative to me at this moment. That is, it is unsurprising that different tests that examine different aspects don't necessarily correlate well with one another.

Response 8

We agree that the lower correlations among test statistics reflect the fact that different tests target different aspects of distributional differences - such as mean, variance, or shape - and thus are not expected to be strongly correlated. Our intention in presenting the correlation matrix was not to suggest redundancy or equivalence among tests, but rather to illustrate their divergent sensitivity profiles as well as similarities.

This supports our central argument: that the ζ-overlapping test captures a broader set of distributional differences than standard tests. However, we understand that this may not have been sufficiently clear in the previous text. We have revised the manuscript to clarify the rationale for including the correlation matrix and to temper claims about its interpretive value.

Please find the changes marked in red in the revised manuscript (after the correlation matrix, page 13)

9

Relatedly, should researchers not just conduct the statistical test that best answers their research question/hypothesis and simply take into account the assumptions and deal with them if they are violated? Additionally, I wonder how applied researchers may use the overlapping test and misinterpret it as evidence for a difference between, for example, means while it is actually the variance or shape of the distributions that differ.

Response 9

We fully agree that researchers should select the statistical test that best addresses their specific research question and consider the assumptions involved. However, our contribution is positioned within contexts where differences between conditions and/or groups may arise not just from the mean, but also from variance and shape - conditions that might be common in psychological data. We believe that there is a broad range of scenarios where researchers are not interested in mean differences and can benefit from the use of a global test focused on distributions, as shown by the number of citations received by Pastore & Calcagnì (2019)

In this light, we present the ζ-overlapping test as a global omnibus test, conceptually similar to ANOVA, but not limited to a single parameter. Whereas ANOVA tests for differences in means across groups, the ζ-test evaluates whether two empirical distributions differ in any aspect, including central tendency, dispersion, and shape. Its strength lies in offering a single, assumption-light inferential tool capable of capturing complex, multivariate distributional discrepancies, capable of controlling type I error avoiding multiple tests.

To minimize the risk of misinterpretation, we now have clarified in the manuscript that a significant ζ result indicates that the distributions differ in some way, but data visualization, descriptive statistics and possible follow up analyses are necessary to determine the specific nature of the difference.

Please find the changes marked in red in the revised manuscript “2.4 Illustrative example with real data”

10

It is more frequently recommended to conduct equivalence and minimum-effect tests (Murphy & Myors, 1999; Lakens et al., 2018; Riesthuis, 2024). It would be good to provide researchers with ways to conduct such tests using this approach. Especially, as the authors argue that the η is an intuitive effect size.

Response 10

We agree that equivalence and minimum-effect testing are increasingly advocated as more informative alternatives to traditional null hypothesis significance testing (NHST), particularly when researchers are interested in ruling out trivial effects or demonstrating the presence of meaningful ones (e.g., Murphy & Myors, 1999; Lakens et al., 2018; Riesthuis, 2024). Given that the ζ-overlapping test is built upon an interpretable effect size (η), it is well-suited for the adaptation of these approaches.

In the revised manuscript, we have added a paragraph outlining how researchers can perform equivalence and minimum-effect testing using ζ. Specifically, we describe how predefined thresholds for η (or ζ) can serve as bounds of practical equivalence or minimum effect sizes, and how permutation-based inference can be used to statistically evaluate whether the observed overlap is smaller or larger than these thresholds.

Please find the changes marked in red in the revised manuscript (end of Discussion page 17)

11

On page 13 it is stated: “In cases where the test fails to detect a difference between the two distributions, one can conclude that the groups/conditions do not differ significantly in any parameter”. No evidence for this is provided as it is not examined whether the overlapping test outperforms the more focused statistical tests (e.g., t-test). Hence, it could be that the Welch t-test is more powerful than the overlapping test when the focus is on mean differences. A simulation study examining this issue would be beneficial.

Response 11

We appreciate the reviewer’s attention to this important interpretive point. We agree that the statement on page 13 may overstate the inferential scope of a non-significant ζ-overlapping result. The ζ-test indicates whether the empirical distributions differ in any measurable way, but a non-significant result does not allow us to conclude definitively that no difference exists in any parameter - only that no difference was detected across the aspects the test is sensitive to (mean, variance, shape) under the given sample conditions.

To address this, we have revised the text to temper the claim and now emphasize the caution needed when interpreting a Null result of the ζ-test. We also acknowledge that when the research focus is on a specific parameter (e.g., the mean), more focused tests such as the Welch t-test may indeed be more aligned. We now include this point explicitly and suggest that future simulation studies could be conducted to systematically compare the sensitivity of the ζ-test to focused tests under different data-generating conditions useful for specific field and research questions.

Please find the changes marked in red in the revised manuscript (page 13, just after the original statement)

12

Last, I was unable to reproduce the results of the simulation study because the "EngTurk.csv" file was missing which didn’t allow me to render the results. Also, the script was difficult to follow and some clear annotations would be beneficial.

Response 12

We thank the reviewer for highlighting this point, we should have addressed the location and source of the file in the manuscript. The mention of the file "EngTurk.csv" refers to a dataset that we retrieved from an open repository on OSF (Oksuz & Rubschat, 2024)and is not required for reproducing the simulation results presented in the current manuscript. Link to the dataset: https://osf.io/muwjz/files/ubtyq).

We thank the reviewer we enriched the code with informative comments following the indication to increase its reproducibility and added a clear citation in the manuscript that allows to retrieve the data: “To give a real world example we computed the overlap between two conditions of a reading test capturing RTs. The data is taken from an open repository on OSF (Oksuz & Rubschat, 2024).”

Minor comments

1) Information missing in the following sentence on page 5: “Even though the Overlapping Index has a simple interpretation, one could argue that it does not provide information on the significance of η, therefore, we decided to implement permutation testing to offer to the ones interested a value of significance”

We thank the reviewer, we rephrased the text as follow : “Even though the Overlapping Index has a simple interpretation, it does not provide information about the statistical significance of η. Therefore, we implemented permutation testing to provide a measure of statistical significance.”

2) Figure 3 is missing the y axis and x axis titles.

We thank the Reviewer we modified Figure 3 accordingly.

3) There seems to be some errors in the following sentence on page 11: "In figure 7, is represented type I error in panel [A] and power in panel [B] estimated considering as true null hypothesis the situation in which samples are drawn from two exactly equal populations (figure 5, panel 1).".

We thank the Reviewer and modified as follow: " Figure X shows the type I error in panel A and the power in panel B, estimated under the null hypothesis corresponding to the situations described in the section *Definition of Statistical Tests*."

4) There are quite some spelling mistakes throughout the manuscript (e.g., page 6: assumtions). It would be good to proofread the manuscript once again.

We thank the reviewed for the detailed comment, we proof-read the entire manuscript.

Reviewer 2

Summary

This manuscript introduces a permutation-based significance test of the overlapping index for independent samples. The overlapping index quantifies the overlap of two distributions using their empirical density functions. It also describes a simulation study evaluating the Type 1 error rate and power of the overlapping index and other significance tests for independent samples (t, Welch, Wilcoxon, variance homogeneity, and KS test). The results show that the overlapping index maintains good Type I error rates but has much more power than tests solely focussing on central tendency or variance. In fact, it shows a rather high correlation with the KS-test.

Evaluation

Overall, I cannot really see the manuscript's contribution to the psychological literature or how it fits within the scope of PB&R. There are a number of issues with the current manuscript and I cannot see how a revision could overcome them to make it suitable for PB&R or a psychological audience. Maybe the authors might be better served by submitting this to a more statistical rather than psychological audience.

Criticisms

1

I cannot see the benefit of using a test based on the overlapping index for psychology. As the authors themselves state, the main problem with the test is that it does not inform the researcher which feature of the two distributions compared differs. I cannot think of a psychological research question where knowing there is some difference between two distributions, but not knowing which feature differs, is helpful in answering the question. Rather, such a test seems like an invitation for researchers that are interested in going for a fishing expedition in their data. Put more poignantly: The problem in psychology is not too few significant results but too many and the present tests offers even more significant results than presently available.

One very telling result from the manuscript's simulation study is the finding that the largest correlation of the overlapping index test is with the KS test. However, the KS test is hardly used in psychology and for good reasons. Researchers are not interested in knowing if there is some difference between distributions, but are interested in knowing if two distributions differ in their central tendency.

Response 1

We appreciate the reviewer's engagement and the opportunity to clarify the contribution and applicability of our approach. We respectfully disagree with the claim that the ζ-overlapping test is of limited value to psychological research or that it invites unprincipled data exploration.

A similar concern was raised by another reviewer, and we would like to note that we examined the citation network of Pastore & Calcagnì (2019), the seed and seminal paper introducing the Overlapping index (η), and found that as of February 2026, it has garnered over 370 total citations, with more than 120 in psychology-related journals alone, reflecting widespread adoption across disciplines. Recent citations span neuroscience, psychometrics, cognitive science, developmental psychology, and educational research, underscoring η's utility beyond descriptive statistics.

For instance, in neuroscience, Yamauchi et al. (2025) applied η to quantify distributional overlaps in neural reward monitoring signals from macaque frontopolar cortex, revealing subtle differences in response variability during decision-making tasks. In cognitive and behavioral measurement, (DOI: 10.3758/s13428-025-02918-6) cited the Overlapping Index to evaluate distributional similarities emphasizing η's role in assessing non-normal data overlaps across in lab and online data collection. In developmental and clinical psychology (Spaggiari et al., 2024). In language and educational psychology, Conversano et al. (2024) quantified reading behavior overlaps between monolingual and bilingual learners, demonstrating η's sensitivity to shape and tail differences. These applications across neuroscience, experimental, developmental, cognitive, and clinical fields affirm η's substantive value for analyzing variance, skewness, and tails in behavioral data, likely key for psychological inference. We thank the reviewer that allowed us to highlight this fundamental aspect. Please find the changes marked in red in the revised Manuscript (end of Introduction, page 2).

Psychological phenomena often manifest through complex patterns in the data that go beyond mean differences. Numerous studies have highlighted how differences across manipulation checks or groups in variability (e.g., in reaction time), distributional asymmetry (e.g., skewed responses in emotion ratings), or multi-modal responses (e.g., in decision making) are both theoretically meaningful and empirically common. Traditional tests focusing solely on the mean risk missing such differences entirely. The ζ-test offers researchers a transparent and assumption-light omnibus alternative, allowing them to detect whether two groups differ in any distributional aspect, not to replace focused tests, but to complement them when the exact nature of the difference is unknown or when standard assumptions (e.g., normality, homoscedasticity) are violated. There are also very specific cases in psychological sciences when it is of interest of the researchers to exclude major differences between groups or conditions (control variables) and such a test would allow to use only one test instead of multiple ones, lowering the risk of false positives (as shown in Supplementary Materials).

Second, the reviewer argues that such a tool could promote “fishing expeditions.” On the contrary, we argue that the explicit quantification of distributional overlap - grounded in a well-defined effect size (η) - discourages arbitrary or post hoc testing by summarizing complex differences in a single interpretable metric. We are careful in the manuscript to note that a significant ζ-test should not be overinterpreted and must be preceded by extensive data visualization and descriptive statistics and followed if necessary by theory-driven, focused follow-ups. We further clarify this in the revised text to avoid misuse.

Lastly, while the ζ-test is indeed most correlated with the Kolmogorov–Smirnov (KS) test, we respectfully disagree that the lack of use of the KS test in psychology reflects its irrelevance. Rather, it reflects a disciplinary bias toward mean-centric inference, often reinforced by convenience rather than theoretical fit. Our work challenges this limitation by offering a new tool tailored to realistic psychological data that often violate classic test assumptions and express effects beyond location shifts.

Thank to the reviewers comment the revised manuscript now clearly show that psychology - especially in its growing push toward distributional thinking, robustness, and replicable inference - can benefit from a global test that provides valid signals of any statistical difference while maintaining interpretability and computational transparency and that the effect size related to it has been already used in published literature.

2

I am not convinced by the authors statement regarding the non-parameteric or distribution-free nature of their test. In particular, the problem with empirical data is of course that we do not know the distribution function f(). Hence, what the calculation of the overlapping index requires is a way to empirically obtain f(). Most likely this is achieved via a kernel-density estimator. However, kernel-density estimators also have parameters that might need tuning and their performance depends on features of the data. For example, kernel-density estimators have problems with sparse data or bounded data. And especially bounded data is very common in psychology. Hence, the complete absence of a discussion of the limits or dependency of the current method on the validity and suitability of the underlying kernel-density estimator is not satisfying.

Response 2

We thank the reviewer for this important observation. The concern about the dependency of the ζ-overlapping test on kernel density estimation is entirely valid and deserves explicit discussion.

While our test is nonparametric in the inferential sense, meaning that no assumptions are made about the underlying distribution (e.g., normality, homoscedasticity) and p-values are computed via permutation, it does indeed rely on kernel density estimation (KDE) to approximate the empirical density functions. We agree that this estimation introduces smoothing parameters (e.g., bandwidth) that can influence the resulting overlap measure, particularly in sparse samples or when data are bounded, as is common in psychological applications (e.g., Likert scales, percentages, or bounded reaction times).

To address this point, we have added a dedicated paragraph in the manuscript discussing the role of KDE in our method, its limitations (particularly with sparse or bounded data), and recommendations for appropriate bandwidth selection (e.g., boundary-corrected kernels or transformations for bounded data). Importantly, we note that KDE-based estimation is already widely used in psychology (e.g., in estimating response time distributions) and its limitations are well understood and manageable. We also verified through simulation (including non-normal and skewed distributions) that the performance of our test remains robust across common data types in psychology. Nonetheless, we agree that transparency regarding this assumption is essential, and we thank the reviewer for prompting this addition.

Add to the Manuscript (In the section of the Overlapping Index):

Of note, although the ζ-Overlapping test is nonparametric in that it does not assume any specific parametric form for the underlying distributions and relies on permutation for inference, it does require estimating empirical density functions using kernel density estimation (KDE). Like all nonparametric estimators, KDE introduces smoothing parameters (e.g., bandwidth) that can affect the precision of the overlap estimate. While standard bandwidth rules (e.g., Silverman's rule) work well in many scenarios, special care is required in cases of bounded data, where standard kernels may yield biased estimates near the boundaries. In such cases, boundary-corrected KDE methods or data transformations (e.g., logit for proportion data) can be used to improve accuracy. We note that KDE is widely accepted in psychological research (e.g., for modeling reaction time distributions). Nonetheless, researchers should be mindful of KDE characteristics when applying the method in bounded or sparse data contexts.

3

Unfortunately, the manuscript is not very clearly written with several sentences and sometimes even paragraphs difficult to understand. Take for example the first paragraph of the manuscript. None of the three sentences it consists of is particularly clear. What is actually being said here? What is the message that the authors try to convey? There are overall way too many sentences with vacuous/unclear message.

Response 3

We thank the reviewer for pointing out the importance of clarity and precision in writing. We acknowledge that certain parts of the manuscript, particularly the opening paragraph, could have been expressed more directly. In response, we have exstensively rewritten the first paragraph to clearly state the motivation for the study, the problem it addresses, and the rationale for proposing the ζ-overlapping test. Our goal was to present the core contribution in simple and unambiguous terms. Please find the revision marked in red in the Main manuscript.

While the reviewer refers to other sentences or paragraphs as unclear or vacuous, no specific examples beyond the first paragraph are provided. Nevertheless, we have conducted a thorough line-by-line review of the manuscript and revised multiple passages for clarity, focus, and readability. These edits aim to ensure that the structure of the argument and the purpose of each section are transparent and accessible, especially for a psychological audience.

We hope the revised version marked in red in the revised manuscript reflects these improvements and makes the manuscript's contribution more immediately clear to all readers.