

Peer-Review

Kolesnikov, Alex. 2026. "A Color-Coded Urgency Bar and Intervention Thresholds in Sequential Health Decisions." *Journal of High School Science* 10 (3): 394–413.
<https://doi.org/10.64336/001c.166227>.

Please address my concerns below

The manuscript does not describe how the numerical urgency score was presented to participants. Specifically, it is unclear whether participants were informed of the scale range or maximum possible score. If participants were shown only raw values (e.g., "14") without contextual information (e.g., 14/21 or an explained scale), the numerical-score condition may have been disadvantaged relative to the color-coded display. Consequently, the conclusion that color coding outperforms numerical presentation may reflect differences in interpretability rather than an inherent advantage of color coding. The authors should clarify the participant-facing interface and justify why the numerical presentation constituted a fair comparator.

Experimental urgency scores lack validation. The urgency scores are entirely author-generated. For example, fever = 3, headache increases score, vomiting increases score, but there is no rationale explaining why these weights are appropriate. The manuscript explicitly states that these are experimental quantities rather than validated clinical scores. Although acceptable for a proof-of-concept study, the paper should discuss this more prominently because the numerical thresholds (e.g., intervention at score 14.37) have no external clinical meaning.

The authors should report the GEE working correlation structure (e.g., exchangeable, independence, or autoregressive) and briefly justify its selection. Given that the repeated observations arise from sequential decision opportunities within scenarios rather than conventional longitudinal measurements, the choice of working correlation deserves explanation. Without this reporting, your work cannot be reproduced.

Sample size reporting is confusing. Initially the manuscript reports 277 participants but later reports, 1,083 observations and 3,047 decision rows. These are legitimate derived observations, not independent participants, but the manuscript should either explicitly derive these numbers or briefly explain how these came to be.

Color effects may not generalize. The conclusion is that colour-coded bars improve intervention. However, the experiment actually tested one particular color scheme, interface and graphical implementation. The discussion occasionally sounds broader than the evidence supports. This may not necessarily generalize to different color palettes, alternative visual encodings or presentation formats.

Add error bars to Figure 2 to facilitate interpretation

Holm adjustment is mentioned. However, the manuscript should specify exactly which contrasts, family definition and adjusted p-values, rather than simply stating Holm correction was applied.

The manuscript does not adequately describe the implementation of the colour-coded urgency bar. The Methods should explicitly state the color palette used, the mapping between urgency scores (or stages) and colors, and the thresholds at which colors changed. If the interface used conventional

green–yellow–red signaling, this should be stated explicitly, as learned semantic associations with these colors may partly explain the observed behavioral effect. Including an example image or screenshot of each information format would greatly improve reproducibility and interpretation. Please include screenshots of all 6 display formats in a supplementary file or as an appendix. Currently, the reader is asked to infer what “visual bar,” “coloured bar,” and “multimodal” looked like, making it difficult to interpret or replicate the experiment.

Response to the Editor and Reviewers

A Colour-Coded Urgency Bar and Intervention Thresholds in Sequential Health Decisions

Alex Kolesnikov

22 July 2026

General response

Dear Editor and Reviewers,

Thank you for the careful and constructive assessment of the manuscript. I have revised the manuscript, supplementary materials and article-specific OSF reproducibility package substantially. In accordance with the Editor’s instruction, each comment is reproduced verbatim below and followed by my response, the changes made and the location of the corresponding revision.

Editor’s instruction

Editor comment

Dear author,

Please address all the comments of the reviewer and revise your manuscript accordingly.

When you resubmit, please also submit separately a word file that lists verbatim each comment by the reviewer and then describe how and where in the revised manuscript you have addressed that particular comment. Do this for all comments in one separate file.

Please note that if the the concerns/comments/questions of the reviewer are addressed inadequately, incompletely or insufficient address is paid to detail, your manuscript may be rejected. Please therefore thoroughly review your responses before resubmission.

We look forward to reviewing your revised manuscript.

Best,

Shireesh Apte

Response

I have prepared the requested point-by-point response, with every reviewer comment reproduced verbatim and linked to the relevant revisions in the manuscript, supplementary materials or OSF repository. A dedicated article-specific OSF repository was created to support this revised manuscript and contains the de-identified data, participant-facing materials, analysis code, generated outputs and reproducibility documentation. Anonymous peer-review link: https://osf.io/3k8jt/overview?view_only=7d51a6100eef4644b35a945c2445abc4.

Summary of major revisions

1. **Complete documentation of the tested interfaces.** The Methods now describe all six participant-facing formats, including the numerical-score wording, category boundaries, colour mapping, ordinary and coloured bars, and the multimodal display.
2. **Participant-facing screenshots and source materials.** Supplementary Figure S1 presents all six formats at the same underlying scenario stage, while the complete original Part B stimuli are retained in the dedicated article-specific OSF repository.

3. **Fuller explanation of the experimental urgency score.** The revised Methods explain the score's construction, methodological precedents, distinction from established clinical instruments, author-assigned weights and lack of external clinical interpretation. Supplementary Tables S1 and S2 report the complete symptom-weight mapping and scenario trajectories.
4. **Explicit specification of the primary model.** The manuscript identifies the participant-clustered logistic GEE, exchangeable working-correlation structure, robust sandwich standard errors, predictors, reference categories, estimated working correlation and software environment.
5. **Exact derivation of the analytical sample.** The revision traces 279 submitted surveys to 277 consenting participants, 1,108 potential Part B sequences, 25 ambiguous sequences, 1,083 clean scenarios and 3,047 displayed stage-level at-risk rows.
6. **Complete multiplicity reporting.** All seven prespecified Holm-adjusted contrasts are reported with their directions, estimates, confidence intervals and raw and adjusted p-values. The multimodal-versus-coloured-bar comparison remains separate and is labelled exploratory.
7. **Improved uncertainty reporting.** Figure 2 now includes 95% participant-cluster bootstrap confidence intervals based on 20,000 resamples, with the resampling method and fixed seed reported in the Methods and caption.
8. **Additional robustness analyses.** The revision reports working-independence, score-independent, conservative reconstruction of ambiguous sequences, demographic-adjusted and categorical-urgency analyses.
9. **Explicit analytical assumptions and scope.** A dedicated subsection now explains the risk-set representation, fixed five-stage endpoint, censoring treatment, continuous-score assumption, working-correlation specification, participant clustering, prior calibration and the hypothetical nature of the outcome.
10. **Narrower interpretation.** The article now refers specifically to the tested green-yellow-orange-red bar within this vignette task and avoids claims about general colour effects, clinical improvement, causal mechanisms or real-world behaviour.
11. **Dedicated article-specific reproducibility repository.** A new anonymous OSF repository was created to support this revised article and contains de-identified data, participant-facing materials, scripts, generated outputs, software-environment information, a release audit, manifest and manuscript-output crosswalk.
12. **Transparent assistance disclosure.** The Acknowledgments and AI Assistance section identifies Ellie Fassman's project-mentor guidance and review of the results and overall paper structure, together with the limited use of OpenAI ChatGPT for spelling, grammar and general language editing.
13. **Reference and language audits.** Reference links and bibliographic details were checked against the cited sources, and the final manuscript was audited for consistent scholarly tense and third-person presentation.

Reviewer 1

Comment 1.1 — Numerical-score comparator

Reviewer comment

The manuscript does not describe how the numerical urgency score was presented to participants. Specifically, it is unclear whether participants were informed of the scale range or maximum possible score. If participants were shown only raw values (e.g., "14") without contextual information (e.g., 14/21 or an explained scale), the numerical-score condition may have been disadvantaged relative to the color-coded display. Consequently, the conclusion that color coding outperforms numerical presentation may reflect differences in interpretability rather than an inherent advantage of color coding. The authors should clarify the participant-facing interface and justify why the numerical presentation constituted a fair comparator.

Response

I agree that the original manuscript did not describe the participant-facing numerical condition sufficiently and that the numerical and coloured-bar conditions did not provide identical amounts of immediate interpretive context.

In the numerical-score condition, participants saw wording such as “Urgency score: 14”; a denominator or universal maximum was not repeated on each Part B screen. However, all participants had received common score calibration and had encountered numerical scores and graphical bars in Part A.

The revised manuscript therefore treats the comparison as one between the complete interfaces tested, not as proof that colour is inherently superior to numbers. It emphasises the more closely matched ordinary-bar-versus-coloured-bar contrast, while acknowledging that conventional traffic-light meanings may also have contributed.

Changes made

- Specified the numerical-score wording, prior calibration and absence of a repeated denominator or maximum.
- Added the limitation that the formats provided different amounts of immediate interpretive context.
- Emphasised the ordinary-bar-versus-coloured-bar comparison and added Supplementary Figure S1.

Location

Revised manuscript, Sections 2.4 and 2.5, Table 2, Sections 4.2 and 4.5; Supplementary Figure S1; complete Part B stimuli in the OSF repository.

Comment 1.2 — Construction and validation of the urgency weights

Reviewer comment

Experimental urgency scores lack validation. The urgency scores are entirely author-generated. For example, fever = 3, headache increases score, vomiting increases score, but there is no rationale explaining why these weights are appropriate.

Response

I agree that the original manuscript did not explain the construction and purpose of the experimental urgency weights sufficiently. I assigned the weights before data collection; they were not fitted to participant responses, selected after inspecting the results, copied from a clinical score or calibrated against a clinical outcome. Lower values were generally assigned to relatively common or nonspecific cues and higher values to more concerning warning signs; the values ranged from 3 to 8.

The revised Methods places the design in context: PRO-CTCAE structures patient-reported symptom items, CTCAE grades adverse-event severity, NEWS2 aggregates weighted physiological abnormalities and the NICE fever system uses ordered traffic-light categories. These are methodological precedents, but none supplies a general additive urgency scale for the lay-described symptoms and controlled cumulative trajectories used here.

The weights served only as an internal coordinate for constructing matched trajectories.

Supplementary Tables S1 and S2 report the full mapping and trajectories, and the principal coloured-bar result remained positive in a score-independent analysis using stage and exact scenario version.

Changes made

- Expanded Section 2.3 to explain the timing, rationale and limited purpose of the author-assigned weights.
- Added and carefully distinguished the PRO-CTCAE, CTCAE, NEWS2 and NICE methodological precedents.
- Reported the complete mapping, trajectories and score-independent sensitivity analysis.

Location

Revised manuscript, Section 2.3 and references 23-26; Sections 2.9, 3.5, 4.5 and 4.6; Supplementary Tables S1, S2 and S6; supporting source files and code in the OSF repository.

Comment 1.3 — External clinical meaning of the numerical thresholds**Reviewer comment**

The manuscript explicitly states that these are experimental quantities rather than validated clinical scores. Although acceptable for a proof-of-concept study, the paper should discuss this more prominently because the numerical thresholds (e.g., intervention at score 14.37) have no external clinical meaning.

Response

I agree. The value 14.37 and all other urgency values are internal coordinates of this experimental task, not clinical cut-offs, disease probabilities, measures of diagnostic severity or recommendations about when care should be sought. This limitation is now stated throughout the manuscript, and 14.37 is described only as the mean experimental urgency score at observed intervention in the coloured-bar condition.

The revision also distinguishes earlier intervention from clinically appropriate intervention and notes that earlier action could either reduce delay or increase unnecessary alarm. The score-independent sensitivity analysis further shows that the principal format association did not depend solely on the author-generated values.

Changes made

- Strengthened the non-clinical interpretation throughout the Abstract, Methods, Results, Discussion and Conclusion.
- Distinguished behavioural timing from clinical quality and highlighted the score-independent sensitivity analysis.

Location

Revised manuscript, Abstract; final two paragraphs of the Introduction; Sections 2.3, 2.4, 2.10, 3.1, 3.3, 3.5, 4.1, 4.4, 4.5 and 5; Tables 4 and 6 and Figure 2 interpretation; Supplementary Table S6.

Comment 1.4 — GEE working correlation and reproducibility**Reviewer comment**

The authors should report the GEE working correlation structure (e.g., exchangeable, independence, or autoregressive) and briefly justify its selection. Given that the repeated observations arise from sequential decision opportunities within scenarios rather than conventional longitudinal measurements, the choice of working correlation deserves explanation. Without this reporting, your work cannot be reproduced.

Response

I agree and now report the complete primary specification: a population-average logistic GEE with participant clustering, an exchangeable working-correlation structure, robust sandwich standard errors and an estimated working correlation of 0.323.

Exchangeability was used as a parsimonious efficiency specification rather than a claim that all true within-participant correlations were identical. The same model was refitted using working independence, with the coloured-bar associations remaining positive but attenuated. The manuscript and OSF package also identify the predictors, reference categories, row and cluster counts, software versions, scripts and outputs.

Changes made

- Reported and justified the exchangeable correlation structure, estimated correlation and robust participant-clustered inference.
- Added the working-independence sensitivity analysis.
- Specified Python 3.13.2 and statsmodels 0.14.6 and archived the complete reproducibility materials.

Location

Revised manuscript, Sections 2.8, 2.9, 2.10, 3.2 and 3.5; Tables 5 and 8; Supplementary Tables S4 and S6; analysis scripts, environment records and outputs in the OSF repository.

Comment 1.5 — Derivation of participants, scenarios and stage-level rows

Reviewer comment

Sample size reporting is confusing. Initially the manuscript reports 277 participants but later reports, 1,083 observations and 3,047 decision rows. These are legitimate derived observations, not independent participants, but the manuscript should either explicitly derive these numbers or briefly explain how these came to be.

Response

I agree and have added the complete derivation: 279 submitted surveys minus two without consent left 277 participants; $277 \times$ four Part B scenarios produced 1,108 potential scenario sequences; excluding 25 ambiguous sequences left 1,083 clean scenarios.

Clean scenarios were expanded only across displayed at-risk stages: waits were coded 0, the first “Seek help now” response was coded 1 and no later rows existed; valid non-interveners contributed five 0-coded rows. This produced 3,047 stage-level rows clustered within the original 277 participants. Supplementary Table S7 reports every ambiguous pattern and reason.

Changes made

- Added the complete arithmetic and distinguished participants, scenarios and stage-level rows.
- Defined the valid sequence and risk-set reconstruction rules.
- Added Table 3 and Supplementary Table S7 for allocation, clean counts and excluded patterns.

Location

Revised manuscript, Sections 2.1, 2.2, 2.6, 2.7, 2.8 and 3.1; Figure 1; Table 3; Supplementary Table S7; reconstruction and audit files in the OSF repository.

Comment 1.6 — Generalisation beyond the tested colour implementation

Reviewer comment

Color effects may not generalize. The conclusion is that colour-coded bars improve intervention. However, the experiment actually tested one particular color scheme, interface and graphical implementation. The discussion occasionally sounds broader than the evidence supports. This may not necessarily generalize to different color palettes, alternative visual encodings or presentation formats.

Response

I agree. The experiment tested one green-yellow-orange-red palette, one set of category boundaries and one graphical implementation. The revised article therefore refers to the tested coloured-bar implementation and reports associations within this vignette task, rather than claiming that colour coding generally improves decisions.

The ordinary-bar-versus-coloured-bar contrast provides narrower evidence because the displays shared the same scale, labels, position marker and arrow; however, familiar traffic-light semantics may still have contributed. The Limitations and Future Research sections now address other palettes, layouts, accessibility, cultural conventions and real-world validation.

Changes made

- Narrowed the title, Abstract, Discussion, Limitations and Conclusion to the tested implementation and vignette task.
- Added limitations concerning traffic-light semantics, other palettes and layouts, accessibility and generalisability.

Location

Revised manuscript, title, Abstract, Sections 2.5, 4.1-4.6 and 5; Supplementary Figure S1.

Comment 1.7 — Error bars in Figure 2

Reviewer comment

Add error bars to Figure 2 to facilitate interpretation

Response

I agree. Figure 2 now shows the mean experimental urgency score at observed intervention for each format with 95% participant-cluster bootstrap confidence intervals. The intervals use 20,000 fixed-seed resamples of participants, preserving repeated scenarios within participants; the Methods and caption identify the interval type and resampling unit.

Changes made

- Added participant-cluster bootstrap confidence intervals and documented 20,000 resamples with seed 20260720.
- Updated the caption and archived the figure-generation code and machine-readable output in OSF.

Location

Revised manuscript, Sections 2.7, 2.8 and 3.1; Figure 2 and its caption; figure-generation script and outputs in the OSF repository.

Comment 1.8 — Definition and reporting of the Holm family

Reviewer comment

Holm adjustment is mentioned. However, the manuscript should specify exactly which contrasts, family definition and adjusted p-values, rather than simply stating Holm correction was applied.

Response

I agree. The revised Methods defines one prespecified seven-contrast Holm family: urgency score, urgency label, visual bar, coloured bar and multimodal versus symptoms only; urgency score versus urgency label; and visual bar versus coloured bar.

Table 7 reports every contrast in its prespecified direction with the estimate, standard error, odds ratio, confidence interval, raw p-value and Holm-adjusted p-value. The multimodal-versus-coloured-bar comparison was not in the planned family and is reported separately as exploratory.

Changes made

- Defined the complete Holm family and added Table 7 and Supplementary Table S3 with all raw and adjusted results.
- Separated the exploratory multimodal-versus-coloured-bar comparison from the planned family.

Location

Revised manuscript, Sections 2.8, 3.4 and 3.6; Table 7; Supplementary Table S3; contrast-output files in the OSF repository.

Comment 1.9 — Colour-bar implementation and screenshots

Reviewer comment

The manuscript does not adequately describe the implementation of the colour-coded urgency bar. The Methods should explicitly state the color palette used, the mapping between urgency scores (or stages) and colors, and the thresholds at which colors changed. If the interface used conventional green–yellow–red signaling, this should be stated explicitly, as learned semantic associations with these colors may partly explain the observed behavioral effect. Including an example image or screenshot of each information format would greatly improve reproducibility and interpretation.

Please include screenshots of all 6 display formats in a supplementary file or as an appendix. Currently, the reader is asked to infer what “visual bar,” “coloured bar,” and “multimodal” looked like, making it difficult to interpret or replicate the experiment.

Response

I agree. The Methods now states the category boundaries explicitly: Low = 0–4, Moderate = 5–12, High = 13–20 and Critical = 21 or higher, mapped to green, yellow, orange and red. The ordinary and coloured bars shared the same scale, labels, position marker and arrow; the coloured version added this mapping, and the multimodal display combined symptoms, score, label and coloured bar. The Discussion acknowledges that learned traffic-light meanings may have contributed. Supplementary Figure S1 presents all six formats at the same underlying stage, and the complete original Part B stimuli are available in OSF.

Changes made

- Added the exact thresholds, labels, palette and shared display elements to the Methods and Table 2.
- Added Supplementary Figure S1 and the complete participant-facing stimuli to OSF.
- Acknowledged learned traffic-light semantics in the Discussion and Limitations.

Location

Revised manuscript, Sections 2.4 and 2.5, Table 2, Sections 4.2, 4.3 and 4.5; Supplementary Figure S1; complete participant-facing Part B materials in the OSF repository.

Reviewer 2

Comment 2.1 — Reference links

Reviewer comment

Please verify that all links to the references point to the correct source.

Response

The final reference list was checked against publisher, journal and official-source records, and each included DOI or URL was verified against the corresponding citation. The anonymous OSF link was also tested in a logged-out private browser after the final release archive was uploaded.

Changes made

- Completed the final bibliographic and DOI/URL audit.
- Verified the anonymous OSF review link and current audited archive.

Location

Revised manuscript, References and Data Availability Statement; article-specific anonymous OSF repository.

Comment 2.2 — Disclosure of assistance

Reviewer comment

The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section.

Response

I agree. The revised manuscript includes a dedicated “Acknowledgments and AI Assistance” section stating that Ellie Fassman provided guidance as a project mentor and reviewed the results and overall structure of the paper. It also discloses that OpenAI ChatGPT was used solely for spelling, grammar and general language editing during manuscript preparation.

The disclosure states that I reviewed and approved all final text and that all claims, interpretations and editorial decisions are my responsibility.

Changes made

- Added a dedicated disclosure identifying Ellie Fassman’s project-mentor guidance and review of the results and overall paper structure.
- Disclosed the limited use of OpenAI ChatGPT for spelling, grammar and general language editing, together with my responsibility for all claims, interpretations and editorial decisions.

Location

Revised manuscript, “Acknowledgments and AI Assistance” section.

Comment 2.3 — Recent, relevant and credible references

Reviewer comment

Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid “padding.” Use credible sources (peer-reviewed journals, reputable books, official reports).

Response

I agree and reviewed the literature for recency, relevance and source quality. The revised Introduction combines foundational work with recent systematic reviews of symptom-checker accuracy and user experience and a 2025 scoping review of clinical-deterioration interface design. The Methods also cites peer-reviewed or official sources for PRO-CTCAE, CTCAE, NEWS2 and the NICE fever traffic-light system, while stating explicitly that these precedents do not validate the study’s particular weights. Each retained or added reference supports a specific claim.

Changes made

- Balanced recent reviews with foundational theoretical and statistical sources.
- Added directly relevant peer-reviewed and official references without presenting them as validation of the experimental weights.

Location

Revised manuscript, Introduction, Section 2.3 and References 1-26.

Comment 2.4 — Explicit assumptions and their consequences

Reviewer comment

All assumptions (including implicit assumptions) must be explicitly stated and clearly justified. The authors should explain why each assumption is reasonable and discuss its impact on the results and conclusions.

Response

I agree. A dedicated “Analytical assumptions and scope” subsection now states the risk-set representation, fixed five-stage endpoint, administrative censoring, continuous-score linearity assumption, exchangeable working correlation, participant clustering, prior calibration and hypothetical nature of the outcome.

The impact of key assumptions is examined through working-independence, categorical-urgency, score-independent, conservative-ambiguity and demographic-adjusted analyses. The Discussion and Limitations explain that the findings concern the tested interfaces and five-stage vignette task and do not establish clinical appropriateness, causal mechanisms or real-world generalisability.

Changes made

- Added Section 2.10 and clarified the principal design, modelling and scope assumptions.

- Added five sensitivity analyses probing major specifications and data-handling choices.
- Expanded the Limitations section to explain the consequences for interpretation and generalisability.

Location

Revised manuscript, Sections 2.2, 2.6-2.10, 3.5 and 4.4-4.5; Table 8; Supplementary Tables S6 and S7; relevant OSF analysis and audit files.

Comment 2.5 — Avoiding overreach

Reviewer comment

Avoid overreaching conclusions that extend beyond what is supported by the data and analysis.

Response

I agree and have narrowed the manuscript throughout. It now describes associations rather than causal effects and refers specifically to the tested green-yellow-orange-red bar within this vignette task. It does not claim that colour coding generally improves decisions, that earlier intervention was clinically preferable or that the results predict actual care-seeking.

The ordinary-bar comparison is presented as narrower evidence about the tested mapping while acknowledging traffic-light semantics, and the multimodal-versus-coloured-bar result is labelled exploratory. The Abstract, Discussion, Limitations and Conclusion now state the principal unresolved questions and limits to generalisation.

Changes made

- Replaced broad or causal language with bounded statements about the tested implementation and observed thresholds.
- Expanded the limitations and narrowed the Conclusion to the evidence supported by the task.

Location

Revised manuscript, title, Abstract, final paragraphs of the Introduction, Sections 3.6, 4.1-4.6 and 5.

Comment 2.6 — Tense and third-person audit

Reviewer comment

Verify that you have used past perfect tense and third person throughout the manuscript wherever applicable.

Response

I interpreted this as a request for consistent scholarly tense and third-person presentation and completed a final language audit. Completed methods and results use the past tense; present tense is retained for established knowledge and current interpretation; past perfect is used only where temporal sequencing requires it. First-person narrative has been removed from the article body.

Changes made

- Completed the final tense and third-person audit and removed first-person narrative from the article body.

Location

Revised manuscript throughout.

I thank the Editor and both Reviewers for comments that substantially improved the manuscript's methodological transparency, reproducibility and restraint of interpretation.

Thank you for your detailed revision and comprehensive response to the reviewers. The manuscript has improved substantially in terms of transparency, reproducibility, and reporting. However, after careful evaluation of your responses and the revised manuscript, I conclude that a fundamental methodological concern remains unresolved.

The central premise of the manuscript is that the colour-coded urgency display is superior to the numerical urgency display. However, the two interfaces do not provide equivalent information. In the numerical condition, participants were shown values such as "Urgency score: 14" without a

denominator or maximum score on the decision screen, whereas the colour-coded display inherently communicated relative position within the urgency scale, category boundaries, and proximity to the highest urgency level. Consequently, the comparison is not between equivalent presentations that differ only by colour, but between a context-poor numerical display and a context-rich graphical display.

Your response argues that this difference is mitigated because all participants received common calibration during Part A. Unfortunately, this does not adequately resolve the concern. The revised manuscript does not establish two critical points that would be necessary to support this argument: That participants were explicitly informed of the maximum possible score (or full numerical range) during Part A, thereby allowing them to interpret a value such as “14” relative to the complete urgency scale; and

That Part B was administered immediately after Part A, such that this calibration would reasonably be expected to remain active during the decision-making task.

Without documentation of both conditions, the assumption that participants retained and applied the calibration throughout Part B is unsupported. Calibration presented earlier is not equivalent to providing the same contextual information at the moment of decision. Participants viewing only “Urgency score: 14” may reasonably interpret that value very differently depending on whether the maximum possible score is 20, 50, or 100. In contrast, the colour-coded bar continuously provides contextual information on every decision screen.

Simply reframing the manuscript as an evaluation of “interfaces” rather than “colour” does not eliminate this methodological confound. The underlying experimental comparison remains unchanged. Because the numerical and colour-coded conditions differed in the amount of immediately available contextual information, the study cannot isolate whether the observed effect arises from colour, graphical representation, relative-position information, category boundaries, or the combination of these features. As a result, the principal experimental comparison—and therefore the central premise of the manuscript—cannot be interpreted unambiguously.

Accordingly, I am unable to recommend acceptance of the manuscript in its current form. Unless it can be demonstrated that participants were explicitly informed of the maximum score during Part A and that Part B followed immediately thereafter so that the calibration formed an integral part of the experimental condition, the comparison between the numerical and colour-coded displays remains fundamentally confounded. This is a methodological limitation of the study design rather than a reporting issue and cannot be resolved simply by narrowing the interpretation of the findings.

Dear Dr. Apte,

Thank you for considering my revised manuscript. I would be grateful if you could arrange one additional review, because the rejection rests on a concern about one comparison, while the manuscript’s other core comparison was not addressed.

Following the reviews, I addressed all fifteen reviewer comments in full and made the requested revisions. The one remaining concern is that the numerical score displayed less immediate context than the coloured bar. I accept this limitation and am glad to remove the numerical-versus-coloured comparison entirely.

The manuscript does not depend on that comparison. Its other central finding is the direct comparison between the coloured bar and the ordinary non-coloured bar. **This was already one of the manuscript’s two core claims in the original submission, not a new analysis introduced after the rejection.** These two displays had the same scale, the same Low/Moderate/High/Critical categories, the same position marker and the same layout; neither displayed any numerical score. They differed only in whether the bar used colour. The coloured bar was associated with higher odds of intervention than the non-coloured bar, $OR = 1.70$, Holm-adjusted $p = .0027$.

Because neither of these two conditions is the numerical condition, the specific contextual asymmetry described in the decision does not apply to this matched comparison. The participant-facing stimulus files in the OSF repository show the two displays directly.

The concern in the decision therefore affects one removable comparison, while the manuscript's other core coloured-bar-versus-non-
I respect your decision either way, and I am grateful for the review process throughout.
Yours sincerely,

Dear author,
I see what you are saying. I have sent it for a second review.
Best,
Shireesh Apte

The second reviewer agreed with the author's interpretation of the first review. I read through the manuscript and also agree that the first review was probably based on a miscommunication /misconception by the first reviewer – who, also, in retrospect agrees with the author's interpretation. Accordingly, I have recommended 'acceptance'.