nature portfolio

Peer Review File

Large-Scale Community Study Reveals Information Sampling Drives Fairness Decisions

Corresponding Author: Ms Sarah Vahed

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Ms Vahed,

Thank you for your patience during the peer-review process. Your manuscript titled "Large-Scale Community Study Reveals Information Sampling Drives Fairness Decisions" has now been seen by 3 reviewers, and I include their comments at the end of this message. They find your work of interest but raised some important points. We are interested in the possibility of publishing your study in Communications Psychology, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response to the reviewers. Please highlight all changes in the manuscript text file.

Editorially, we ask that you revise your analyses to address the methodological concerns raised by all reviewers. Additionally, please enhance the conceptual clarity of the manuscript by providing stronger justifications for your analytical approach and clearly articulating the study's contributions, central research questions, and key take-home messages. Importantly, please remove data from participants under the age of 16.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

I am attaching an Editorial Requests Table that details critical reporting requirements for the revised manuscript. Please attend to each item and ensure your manuscript is fully compliant. If your revised manuscript is not aligned with these requests on major issues, such as those concerning statistics, it may be returned to you for further revisions without re-review.

Please submit the following items:

- Revised manuscript
- Point-by-point response to the referees' comments
- Cover letter (as a separate document)
- Nature Research Reporting Summary
- Completed Editorial Request Table (attached).

via this link: Link Redacted .

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Additional guidance is available in our style and formatting guide Communications Psychology formatting guide.

We hope to receive your revised paper within 8 weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Yafeng Pan

Yafeng Pan, PhD
Editorial Board Member
Communications Psychology
orcid.org/0000-0002-5633-8313

REVIEWER EXPERTISE:
Reviewer #1: Decision-making
Reviewer #2: Decision-making
Reviewer #3: Behavioural Economics

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

Thank you for the opportunity to review this submission, which I found both interesting and well-written. In this paper, Vahed and Sanfey have studied behavior in the Ultimatum game outside the usual laboratory setting, in this case measuring responder behavior with a large sample from a public space. They show that participants not only seek out information about proposers' previous offers, but that this information causes them to be less accepting of unfair offers, especially if the proposer was previously generous. Additionally, they show that acceptance (but not sampling) rates increase with age, and sampling (but not acceptance) rates increase with time of day. These are new findings and will be of interest to those in the fields of psychology and/or behavioral economics, and they were both ethically collected and properly placed in the context of previous findings.

The authors' inferences are reasonable and their analyses thorough. I do think the manuscript would benefit from revision, especially regarding the methods and justification of some of the claims. Critically, in its current state I don't think there is enough material (in either the main text or supplementals) for an outside experimenter to reproduce the experiment.

Best regards,
Damon Tomlin
Wesleyan University

Methodology:

1. The instructions in the Methods section are quite terse - is this all that participants saw? From the text, I could not tell whether

a) participants thought that proposers were at nearby kiosks (pixelated images notwithstanding)

b) participants found the context believable. Based on the answer to a), where did participants believe the proposed offers (and prior offers) occur? Given that hyperfair offers are both empirically rare and, I believe, harder for naive participants to justify, is there evidence that participants thought/did not think these were genuine offers?

c) participants believed they would actually be paid (or had a chance to be paid), or whether this was purely hypothetical. I suspect it was the latter, unless the use of the term "coins" is a translation from a more specific monetary denomination.

2. My biggest methodological issue is that the partners are apparently presented in the same order for all participants. This opens the study up to anchoring and fatigue effects. Is there a compelling reason for this consistent ordering that outweighs the risk of these effects? On the topic of ordering, it also wasn't clear to me whether participants a) saw samples from ALL the partners, then saw the offers from them, or b) saw samples from Partner 1, then an offer from partner 1, then samples from Partner 2, etc.. Figure 1A makes it clear that samples preceded offers, but not how they were interleaved.

Issues with analyses/conclusions

1. Given my methodological concern above, I think "choice order" needs to be included as a factor in the relevant analyses. It may be already, but if so this needs to be clearer. For example, the description of the logistic regression says "A mixed-effects logistic regression model was then fit to examine the interaction effects of sampling behavior (Did not sample vs Sampled) and offer (Game Partner 1 – 4) on the likelihood of accepting offers." Given how the results are described, it seems that these were run as four independent regressions (one for each partner). However, given that the partners were run in a consistent order, I would run an additional regression that takes the order into account.

2. In a similar vein, given that the likelihood of non-sampling nearly doubles from partner 1 to partner 4, are fatigue effects being accounted for in the analyses that depend on the sampling/non-sampling distinction?

3. The authors periodically point out the diversity of the sample. While the diversity in terms of age, especially relative to the usual university population, is evident, it is less clear how diverse the sample is otherwise. Those unfamiliar with this museum and its location can only guess as to the degree of linguistic/ethnic/SES diversity present. If these data were collected from this sample they should be in the supplementals, especially since the authors mention selection bias. If not, data about the museum's clientele in general would be helpful. Otherwise, I would qualify the claims of diversity to highlight age.

4. It appears that even hyperfair offers had a non-trivial rejection rate. Is there evidence (or do the authors have suspicions) that these are the result of a) cultural norms regarding inequality, b) impulsive choosing, c) lack of belief in the experiment's veracity, etc.? Reaction time data and/or consistency in rejection of hyperfair offers would be informative.

5. The authors conclude that "participants who engaged with this information responded distinctly to fairness deviations in their subsequent choices, specifically showing elevated rejection rates to selfish offers when the game partner had previously been shown to be generous." As I understand it, this distinct response to deviations pertained only to deviations toward selfishness, not generosity. Is this merely a ceiling effect or is there a more nuanced explanation for why one occurred and not the other?

6. I had hoped for a bit more on why age would be correlated with prosociality and exploration with time of day. Even if evolutionary arguments are offered in existing literature, I think they could be briefly recapitulated here.

Small details:

1. The caption for Figure 2C refers to "vertical lines," I believe referring to the graphed data. To avoid confusion with the gray gridlines, I would refer to them by color, data labels, or something equally precise.

2. For figure 3B, the third data series from the top is not consistently color coded with figure 2A or the legend.

3. For figure 3 in general, I don't know that the icons on the left are necessary. If they are to remain, I would put them either above, inside, or below their respective graphs, as they refer to the x-axis, rather than the y-axis that each is next to.

4. There's a typo on page 12 - the word "yhud."

Reviewer #2 (Remarks to the Author):

Report on COMMSPSYCHOL-25-0345

The paper reports on an experiment using a non-incentivized ultimatum game (UG) and conducted among museum goes over the course of 13 months. The study replicates the classic findings in the literature in terms of fairness (i.e., an overall willingness to reject unfair offers), but also that many participants are interested in knowing the history of the proposer's decisions and are more likely to reject unfair offers from otherwise fair-minded proposers.

The paper is well written and I enjoyed reading it. Having said that, I have important reservations regarding the contribution of the study and the take-home message it delivers. Currently, many elements of the paper look a bit ad hoc and sometimes appear to be more of a fishing expedition.

1. The ultimatum game has been studied for over 40 years now, and it is unclear to me how the current experiment adds to this huge literature. Many important studies are not cited which makes it hard to put the current article in a proper perspective. In particular, I do not fully grasp how is an non-incentivized experiment using kiosks and run with the population of museum goers more naturalistic and ecologically valid than, for example:

- experiments using ecologically valid (i.e., high) stakes, such as:

Learning in high stakes ultimatum games: An experiment in the Slovak Republic
R Slonim, AE Roth - Econometrica, 1998

Stakes matter in ultimatum games
S Andersen, S Ertaç, U Gneezy, M Hoffman, JA List - American Economic Review, 2011

- experiments comparing various (student-based vs. representative) populations, such as:

Do lab experiments misrepresent social preferences? The case of self-selected student samples
A Falk, S Meier, C Zehnder - Journal of the European Economic Association, 2013

Experimental subjects are not different
F Exadaktylos, AM Espín, P Branas-Garza - Scientific reports, 2013

Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game
F Cochard, J Le Gallo, N Georgantzis, JC Tisserand - Journal of Behavioral and Experimental Economics, 2021

Experiments directly comparing lab vs. field, such as:

On the external validity of social preference games: a systematic lab-field study
MM Galizzi, D Navarro-Martínez
Management Science 65 (3), 976-1002

At the more granular level, the life-cycle variation in fairness have also been document by previous studies, see:

Dictator games: a meta study
C Engel - Experimental Economics, 2011.

Furthermore, the authors should carefully explain the virtues and the limitations of their methodology relative to the existing ones. For instance, what is the advantage of setting up an experiment in a museum over an online experiment with a similar N using a representative sample (and also allowing for using monetary incentives)? I can see an important disadvantage: since the experiment was not supervised, people could participate multiple times, answer collectively, etc.

2. The result on information sampling and its behavioral consequences seems new. Has it been pre-registered? What is the underlying theory and what implications (theoretical? policy?) does it yield? I think this really needs to be clarified.

3. Also, what is the actual motivation and the expected implications of the time of day analyses?

Reviewer #3 (Remarks to the Author):

The truth is that I find the paper very interesting, and the idea of conducting the (uncontrolled) experiments in the museum seems brilliant. That said, I found myself a bit lost with the paper because I don't quite understand what the main question is. I think it would be important to clearly define the focus and what is the central research question of this paper.

Below, I include a series of comments that might help improve and better center the manuscript:

1. Assuming that congruence is the central theme of the paper, I think it would be helpful to define congruence in more detail. I wasn't able to find a clear definition of it in the paper.

2. The binary definition of sampling seems a bit odd. Although the authors acknowledge this as a limitation, it might be worth exploring this issue further, especially regarding what implications it has for congruence.

3. Everything is hypothetical. The decisions have no real consequences for the subjects, yet it seems reasonable to discuss this point. I'm not against hypothetical experiments, and in fact, there is literature suggesting this is not a major issue (see paper 1).

4. The strategy space is binary (which I understand is reasonable), but this has implications for the results, since participants can only accept or reject the offers they receive (hot), and therefore we don't know the minimum acceptable offer, which would be desirable. It might be a good idea to explore the MAO with the existing data, if possible.

5. There is a substantial body of literature on experiments conducted outside controlled lab settings. Papers 2 and 3 are examples, but there are many more.

References

1. https://www.ub.edu/sustainwell-eu-project/wp-content/uploads/2024/12/LOYOLA_Paying-1-or-nothing-in-dictator-games_Unexpected-differences.pdf

2. https://www.nature.com/articles/srep01213

3. https://doi.org/10.1016/j.socec.2022.101910

Version 1:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Ms Vahed,

Your manuscript titled "Large-Scale Community Study Reveals Information Sampling Drives Fairness Decisions" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed here https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commspsychol/open-access

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

* CODE AVAILABILITY: All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

* DATA AVAILABILITY:
All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

Please use the following link to submit the above items:
Link Redacted
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Jennifer Bellingtier

Jennifer Bellingtier, PhD
Senior Editor
Communications Psychology

Yafeng Pan, PhD
Editorial Board Member
Communications Psychology
orcid.org/0000-0002-5633-8313

REVIEWER EXPERTISE:
Reviewer #1: Decision Making, Social Cognition, Neuroimaging, Neuroeconomics
Reviewer #2: Economic Behavior, Decision Making
Reviewer #3: Economic Behavior, Decision Making

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Thank you for the opportunity to review the revised version of the manuscript by Vahed and Sanfey. The additions and revisions that have been made had adequately addressed my concerns and I am happy to recommend the new version of the manuscript.

Best regards,
Damon Tomlin
Wesleyan University

Reviewer #2 (Remarks to the Author):

I would like to start by thanking the authors for a serious and professionally executed revision. The new draft is better embedded in the existing literature, and gives a fair perspective on the contribution, strenghts and limitations of the study.

I'm still unsure whether my comment 3 (related to the time of day analyses) has been fully addressed. It remains unclear to me how the authors interpret these findings, e.g. in lines 405-409 they say:

"For sampling, the chi-square test revealed a significant 403 association between time period (before or after midday) and sampling choices (χ^2 = 6. 25, df = 1, p = 404 .01), with greater sampling observed later in the day. Since our analyses identified a significant effect of time of day on sampling behaviour, we further examined whether age and time of day were related."

So is it association or (causal) effect? The term "association" is safe to use, yet a bit frustrating, since the next question is: what is the underlying driver? For instance, are the subsamples of participants showing up at 10AM and those arriving at 4PM different? I believe the dataset allows to say something about the variation in individual characteristics (e.g. age which seems to matter) of the crowd of museum goers from 9AM till 5PM. Then, do the reported time-of-day differences hold once we control for these individual characteristics? Making causal claims ("effect") is more interesting (it echoes a body of experimental research on circadian rythm and social preferences) yet trickier to defend due to potential selection effects (time slot is not randomized across participants, but remains their choice), especially without any evidence on the participant sample variation during the day.

Minor comment:
Appendix, line 126: typo -- "uutput"

Reviewer #3 (Remarks to the Author):

I am pretty happy with the new version.

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

Thank you for the opportunity to review this submission, which I found both interesting and well-written. In this paper, Vahed and Sanfey have studied behavior in the Ultimatum game outside the usual laboratory setting, in this case measuring responder behavior with a large sample from a public space. They show that participants not only seek out information about proposers' previous offers, but that this information causes them to be less accepting of unfair offers, especially if the proposer was previously generous. Additionally, they show that acceptance (but not sampling) rates increase with age, and sampling (but not acceptance) rates increase with time of day. These are new findings and will be of interest to those in the fields of psychology and/or behavioral economics, and they were both ethically collected and properly placed in the context of previous findings.

The authors' inferences are reasonable and their analyses thorough. I do think the manuscript would benefit from revision, especially regarding the methods and justification of some of the claims. Critically, in its current state I don't think there is enough material (in either the main text or supplementals) for an outside experimenter to reproduce the experiment.

Best regards,
Damon Tomlin
Wesleyan University

We thank the reviewer for their thorough and helpful evaluation of our manuscript. In addition to addressing the point-by-point comments as detailed more fully below, we have incorporated the important suggestion to clarify the details of the task. Specifically, we have now substantially expanded the methodological descriptions (Methods section, specifically under the 'Procedure and Stimuli' section), including more detailed information on participant recruitment as well as the task instructions and structure. Importantly, we have provided the full task materials, including visuals of each experimental screen (including instructions and debriefing), in the Supplementary Materials (See Supplementary Note 1). Together, we believe these additions address the reviewer's suggestion to ensure that the study design is fully reproduceable.

Methodology:

1. The instructions in the Methods section are quite terse - is this all that participants saw? From the text, I could not tell whether

a) participants thought that proposers were at nearby kiosks (pixelated images notwithstanding)

b) participants found the context believable. Based on the answer to a), where did participants believe the proposed offers (and prior offers) occur? Given that hyperfair offers are both empirically rare and, I believe, harder for naive participants to justify, is there evidence that participants thought/did not think these were genuine offers?

c) participants believed they would actually be paid (or had a chance to be paid), or whether this was purely hypothetical. I suspect it was the latter, unless the use of the term "coins" is a translation from a more specific monetary denomination.

We thank the reviewer for raising these points of clarification. Given that our study was situated in the waiting area of a public museum space, we deliberately designed the task (including instructions) to be brief, anonymous, and hypothetical as we purposely sought large-scale public participation in a task that could be completed within a 5 minute window (the mean task completion time was 3 minutes and 23 seconds; SD = 1 minute, 22 seconds). Participants were not informed that offers came from real individuals, nor were they led to believe the task involved real pay-outs. The use of "coins" was not linked to any specific currency. These design choices were specifically made to balance ecological validity with logistical feasibility of data collection in a public setting. We have now clarified these aspects, including a brief rationale for our design choices, in the revised Methods section, under the section on 'Procedure and Stimuli'.

2. My biggest methodological issue is that the partners are apparently presented in the same order for all participants. This opens the study up to anchoring and fatigue effects. Is there a compelling reason for this consistent ordering that outweighs the risk of these effects? On the topic of ordering, it also wasn't clear to me whether participants a) saw samples from ALL the partners, then saw the offers from them, or b) saw samples from Partner 1, then an offer from partner 1, then samples from Partner 2, etc.. Figure 1A makes it clear that samples preceded offers, but not how they were interleaved.

We thank the reviewer for these thoughtful questions. In response to the query regarding trial structure, we have now clarified in the manuscript that, as correctly inferred by the reviewer from Figure 1A, participants viewed trials in a Partner-by-Partner sequence. That is, each trial began with the choice to sample past offers of Partner 1, followed by Partner 1's offer, then the choice to sample past offers of Partner 2, followed by Partner 2's offer, and so on. This detail is included in the Methods section under the 'Procedure and Stimuli' section.

Regarding the fixed trial order, we confirm that all participants encountered the game partners in the same sequence, which could raise concerns about anchoring or fatigue effects. However, we believe such effects are unlikely to have substantially influenced behaviour given the extreme brevity of the task (approximately 3 minutes). We can also point to our finding regarding the consistency in mean acceptance rates for the two identical hyperfair offers, one which was presented first (Game Partner 1 $M_{acceptance}$ = 92.91%) and the other presented last (Game Partner 4 $M_{acceptance}$ = 91.74%). This symmetry in decisions across these orders provides evidence against obvious anchoring or fatigue effects. Finally, we believe these effects have also been addressed in our analysis approach as detailed more fully to Point 1 below.

Issues with analyses/conclusions

1. Given my methodological concern above, I think "choice order" needs to be included as a factor in the relevant analyses. It may be already, but if so this needs to be clearer.

For example, the description of the logistic regression says "A mixed-effects logistic regression model was then fit to examine the interaction effects of sampling behavior (Did not sample vs Sampled) and offer (Game Partner 1 – 4) on the likelihood of accepting offers." Given how the results are described, it seems that these were run as four independent regressions (one for each partner). However, given that the partners were run in a consistent order, I would run an additional regression that takes the order into account.

We thank the reviewer for this comment regarding our analyses, and can confirm that the reviewer is correct that this was indeed catered for in our analyses. As partners were presented in a fixed sequence (Game Partner 1 through Game Partner 4), partner in our logistic regression also reflects presentation order (that is partner, and trial number carry the same information), and thus this factor was indeed included in our model as a fixed effect, along with its interaction with sampling choice. Our model was estimated as a single mixed-effects logistic regression with all four partners simultaneously included, not as separate regressions per partner. We have clarified in the revised manuscript that the inclusion of the Game Partner factor also accounts for choice order. To ensure full transparency, we have also added the complete regression and post-hoc outputs in the Supplementary Materials (See Table S2 and S3).

2. In a similar vein, given that the likelihood of non-sampling nearly doubles from partner 1 to partner 4, are fatigue effects being accounted for in the analyses that depend on the sampling/non-sampling distinction?

We thank the reviewer for this observation and are happy to clarify. As noted, the proportion of non-sampling trials increased across the four Game Partner trials, which could potentially be explained by a fatigue or time-on-task effect. To cater for this we have indeed accounted for differences in sampling and its impact on responses to unfair offers. Specifically, through our iterative downsampling analyses we randomly matched individuals who sampled versus those who did not sample, and repeated the acceptance analyses for Offers 2 and 3 (selfish offers). Across 1,000 iterations, the results were consistent with the original GLMM estimates. Thus, while sampling frequency declined across partners, our analyses did indeed account for these sampling differences and our conclusions are robust to this imbalance. To ensure that this is made clear in our manuscript, we now amended our Methods section (under 'Analyses') with a rationale for our analysis choice and further explain how it helps to account for possible fatigue or time-on-task effects.

3. The authors periodically point out the diversity of the sample. While the diversity in terms of age, especially relative to the usual university population, is evident, it is less clear how diverse the sample is otherwise. Those unfamiliar with this museum and its location can only guess as to the degree of linguistic/ethnic/SES diversity present. If these data were collected from this sample they should be in the supplementals, especially since the authors mention selection bias. If not, data about the museum's clientele in general would be helpful. Otherwise, I would qualify the claims of diversity to highlight age.

We appreciate the reviewer's suggestion to clarify the nature of our sample. As correctly identified by the reviewer, our primary diversity metric is indeed age, which

stands in contrast to typical university samples, which we now emphasise in our revised manuscript. Following the reviewer's suggestion, we have also contacted the museum to obtain public summary information about typical visitor demographics and included this in the Supplementary Materials (See Supplementary Note 2).

4. It appears that even hyperfair offers had a non-trivial rejection rate. Is there evidence (or do the authors have suspicions) that these are the result of a) cultural norms regarding inequality, b) impulsive choosing, c) lack of belief in the experiment's veracity, etc.? Reaction time data and/or consistency in rejection of hyperfair offers would be informative.

We thank the reviewer for this observation. While the hyperfair offers primarily served as attentional controls, as suggested by the reviewer, we have now conducted additional analyses with the goal of better understanding participants who consistently rejected hyperfair offers.

In our sample, we observed an average rejection rate of 7.69% for the hyperfair offers (Offers 1 and 4). We further found that 122 participants (2.61%) of the total sample rejected both of the hyperfair offers. Among these, 79 (1.7%) also rejected both selfish offers, suggesting that a portion of hyperfair rejections may indeed reflect a general "reject-all" strategy, or perhaps a misunderstanding of the task. These exploratory findings have now been included in the Supplementary Materials (Section Figure S2A-C). We acknowledge that the absence of reaction time data limits our ability to distinguish between impulsive responding and more reasoned decisions, and we have noted this as a limitation and area for further research, in the Discussion section of our revised manuscript.

5. The authors conclude that "participants who engaged with this information responded distinctly to fairness deviations in their subsequent choices, *specifically* showing elevated rejection rates to selfish offers when the game partner had previously been shown to be generous." As I understand it, this distinct response to deviations pertained only to deviations toward selfishness, not generosity. Is this merely a ceiling effect or is there a more nuanced explanation for why one occurred and not the other?

We thank the reviewer for pointing out this unclear language. As the reviewer suggests, elevated rejection rates were found when offers deviated negatively (i.e. became more selfish) from earlier behaviour, but not when they became more generous. We have removed the word "specifically" from the relevant sentence to avoid overstatement, as of course we do not directly examine deviations towards generosity. As we clarify, we were particularly interested here in participants' responses to unfairness, with the hyperfair offers primarily serving as attentional controls.

6. I had hoped for a bit more on why age would be correlated with prosociality and exploration with time of day. Even if evolutionary arguments are offered in existing literature, I think they could be briefly recapitulated here.

We appreciate the reviewer's encouragement to elaborate on the potential mechanisms behind the observed age and time-of-day effects, including the possibility

of interesting evolutionary arguments for understanding these effects. We have now added a section to the Discussion, referencing existing evidence on age-related changes in prosociality across lifespan, as well as time-of-day variation in exploratory behaviour, including circadian frameworks.

Small details:

1. The caption for Figure 2C refers to "vertical lines," I believe referring to the graphed data. To avoid confusion with the gray gridlines, I would refer to them by color, data labels, or something equally precise.

2. For figure 3B, the third data series from the top is not consistently color coded with figure 2A or the legend.

3. For figure 3 in general, I don't know that the icons on the left are necessary. If they are to remain, I would put them either above, inside, or below their respective graphs, as they refer to the x-axis, rather than the y-axis that each is next to.

4. There's a typo on page 12 - the word "yhud."

We thank the reviewer for these helpful changes. We have implemented all of them, namely:

- *Figure 2C caption now refers to the coloured data lines to avoid confusion with gridlines.*
- *Figure 3B's third series is now colour-coded consistently with earlier figures and the legend.*
- *Figure 3 icons have been repositioned above the graphs and next to the figure titles to better align with the figure layout.*
- *The typo on page 12 has been corrected.*

Reviewer #2 (Remarks to the Author):

Report on COMMSPSYCHOL-25-0345

The paper reports on an experiment using a non-incentivized ultimatum game (UG) and conducted among museum goes over the course of 13 months. The study replicates the classic findings in the literature in terms of fairness (i.e., an overall willingness to reject unfair offers), but also that many participants are interested in knowing the history of the proposer's decisions and are more likely to reject unfair offers from otherwise fair-minded proposers.

The paper is well written and I enjoyed reading it. Having said that, I have important reservations regarding the contribution of the study and the take-home message it delivers. Currently, many elements of the paper look a bit ad hoc and sometimes appear to be more of a fishing expedition.

1. The ultimatum game has been studied for over 40 years now, and it is unclear to me how the current experiment adds to this huge literature. Many important studies are not cited which makes it hard to put the current article in a proper perspective. In

particular, I do not fully grasp how is an non-incentivized experiment using kiosks and run with the population of museum goers more naturalistic and ecologically valid than, for example:

- experiments using ecologically valid (i.e., high) stakes, such as:

Learning in high stakes ultimatum games: An experiment in the Slovak Republic
R Slonim, AE Roth - Econometrica, 1998

Stakes matter in ultimatum games
S Andersen, S Ertaç, U Gneezy, M Hoffman, JA List - American Economic Review, 2011

- experiments comparing various (student-based vs. representative) populations, such as:

Do lab experiments misrepresent social preferences? The case of self-selected student samples
A Falk, S Meier, C Zehnder - Journal of the European Economic Association, 2013

Experimental subjects are not different
F Exadaktylos, AM Espín, P Branas-Garza - Scientific reports, 2013

Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game
F Cochard, J Le Gallo, N Georgantzis, JC Tisserand - Journal of Behavioral and Experimental Economics, 2021

Experiments directly comparing lab vs. field, such as:

On the external validity of social preference games: a systematic lab-field study
MM Galizzi, D Navarro-Martínez
Management Science 65 (3), 976-1002

At the more granular level, the life-cycle variation in fairness have also been document by previous studies, see:

Dictator games: a meta study
C Engel - Experimental Economics, 2011.

Furthermore, the authors should carefully explain the virtues and the limitations of their methodology relative to the existing ones. For instance, what is the advantage of setting up an experiment in a museum over an online experiment with a similar N using a representative sample (and also allowing for using monetary incentives)? I can see an important disadvantage: since the experiment was not supervised, people could participate multiple times, answer collectively, etc.

We thank the reviewer for this very thoughtful and constructive feedback and critique. We fully concur with the comment that the UG has been extensively studied, and that any new contribution must be carefully situated within this large and diverse literature.

In response, we have substantially revised parts of the Introduction and Discussion to engage more directly with key prior work, including many of the studies highlighted by the reviewer (e.g., Slonim & Roth, 1998; Andersen et al., 2011; Falk et al., 2013; Engel, 2011). Importantly, we have substantially revised our section of the Discussion detailing the virtues and limitations our methodological approach.

Through these revisions, we aim to clarify that the intention of our study is not to position our design as necessarily more ecologically valid than previous high-stakes, representative-samples, or supervised experiments. Rather, we see our approach as complementary, extending the range of settings in which economic decision-making can be observed. Specifically, we see our design as contributing to a growing class of field-embedded behavioural paradigms that allow for scalable assessment of decision-making in minimally scripted and more natural environments than conventional, tightly controlled experimental environments.

Unlike laboratory or online settings, our study was designed to be conducted in a public, unsupervised space, without monetary incentives, researcher presence, or elaborate instructions or tutorials. While this introduces certain limitations, including the possibility of non-unique participation (as the reviewer rightly notes and which we now address in the manuscript), it also, in our view, reduces demand effects and artificiality, particularly as the setting is less anchored in formal experimental norms (i.e. cubicles or experimenters that induce an experimental mindset). We thus view our approach as offering a distinct contribution: capturing more naturalistic patterns of decision-making, at significant scale, in a more public context where participants are not primed by typical experimental expectations.

Importantly, core UG effects (including fairness sensitivity and context-dependent rejection) emerged robustly despite the relaxed constraints of our setup. This supports the utility of our method and its potential to complement more controlled paradigms and designs. Moreover, and as we highlight in response to Point 3 below, one of the key advantages of our field-based approach is the ability to capture behavioural data across diverse temporal and environmental conditions variables (such as time of day effects) that are typically controlled or unmeasurable in laboratory settings.

Ultimately, the value of our study lies not in claiming a more ecologically valid alternative, but in offering a novel, intermediate space between tightly controlled laboratory paradigm and large-scale online survey. We now emphasize this point more clearly in the revised manuscript, including directly comparing our approach to lab, field, and online designs. Where appropriate, we now flag limitations of our approach and describe steps taken to mitigate these concerns (e.g. anonymity, brevity, and a simple UI structure).

2. The result on information sampling and its behavioral consequences seems new. Has it been pre-registered? What is the underlying theory and what implications (theoretical? policy?) does it yield? I think this really needs to be clarified.

We thank the reviewer for these important points regarding the theory and implications of our findings. The analysis was not pre-registered, and was designed to test whether individuals actively seek out non-instrumental social information and whether such information shapes fairness decisions.

While previous work has shown that contextual information can influence fairness decisions, our findings demonstrate that individuals proactively sample such information, even when it has no bearing on outcomes. Unfair offers were more likely to be rejected when they came from proposers who had previously behaved fairly, indicating that people respond to a proposer's behavioural incongruence.

Theoretically, we propose that these findings contribute to work on fairness expectations and highlight the non-instrumental value of social information in guiding norm enforcement. From a policy perspective, the results may inform interventions targeting fairness communication or reputational sensitivity, particularly in settings where perceived consistency may influence decision-making. We now include these points more clearly in the Discussion section to articulate both the value and scope of these contributions.

3. Also, what is the actual motivation and the expected implications of the time of day analyses?

We thank the reviewer for this question and are happy to clarify the motivation behind the time-of-day analyses. As noted in our response to Point 1 above, one key advantage of our field-based design is the ability to examine temporal and environmental factors (such as time-of-day) that are often held constant or remain unmeasurable in traditional laboratory studies, but which may nonetheless impact behaviour.

As our data were collected continuously across the museum's full daily opening hours over a 13-month period, we were thus uniquely positioned through this public venue to explore whether patterns in behaviour (e.g. information sampling, offer acceptance) varied systematically across the day. This exploratory analysis was motivated by existing, though extremely limited, work suggesting that diurnal rhythms in attention, curiosity and social motivation can shape social decision-making.

We now clarify in the manuscript that this was an exploratory, hypothesis-generating analysis intended to illustrate the kind of behavioural variability that can emerge in a more naturalistic, unsupervised setting. In terms of implications, we propose that time-of-day effects on social behaviour may be relevant for studies aiming to understand fluctuations in decision-making or temporal variation in social reasoning. More broadly, such findings could inform the design and timing of interventions or policies in domains where contextual information and fairness decisions are time-sensitive.

Reviewer #3 (Remarks to the Author):

The truth is that I find the paper very interesting, and the idea of conducting the (uncontrolled) experiments in the museum seems brilliant. That said, I found myself a bit lost with the paper because I don't quite understand what the main question is. I think it would be important to clearly define the focus and what is the central research question of this paper.

Below, I include a series of comments that might help improve and better center the manuscript:

1. Assuming that congruence is the central theme of the paper, I think it would be helpful to define congruence in more detail. I wasn't able to find a clear definition of it in the paper.

We thank the reviewer for their constructive and helpful review of our manuscript. We thank the reviewer for the compliment above, and also for pointing out the need for clearer framing of the central research questions of our study. In response, we have now revised the Introduction and Discussion sections by explicitly stating the two main research questions guiding the paper: (1) Do people seek out non-instrumental social information, as captured through information sampling? and (2) If so, does the (in)congruence of a game partner's behaviour impact social decision-making?

In the revised Introduction and Methods sections, we now more clearly define 'congruence' as the alignment between a proposer's past behaviour (i.e. their prior offers to others) and their current behaviour (i.e. the offer made to the participant). Incongruent trials involve a deviation from prior behaviour, for example, when a proposer previously made generous offers to others but now presents a selfish one. We clarify that one of our key research questions focuses on how participants respond to such incongruent behaviour, and whether these deviations result in increased rejection rates to selfish offers as a form of social punishment.

2. The binary definition of sampling seems a bit odd. Although the authors acknowledge this as a limitation, it might be worth exploring this issue further, especially regarding what implications it has for congruence.

We agree with the reviewer that the binary operationalisation of sampling behaviour (Sampled vs. Did not sample) simplifies a potentially richer behavioural variable. In this regard, we highlight our further analyses in our Results section, which now contains further details, where we treated sampling as the actual number of prior offers sampled (as opposed to a binary classification), and its impact on subsequent decision-making. Interestingly, the amount of sampling did not reveal consistent, significant effects. More specifically, sampling two to four boxes (i.e. past offers) did not significantly alter acceptance rates as compared to sampling just one. Therefore, for simplicity we opted use the binary classification of sampling in the manuscript. We agree that this area warrants further research with designs that allow for capturing more nuanced sampling patterns and more complex proposer profiles. This point is specifically highlighted, and expanded upon, in our limitations section of the Discussion.

3. Everything is hypothetical. The decisions have no real consequences for the subjects, yet it seems reasonable to discuss this point. I'm not against hypothetical experiments, and in fact, there is literature suggesting this is not a major issue (see paper 1).

We appreciate this important point raised by the reviewer and the helpful reference to supporting literature. We now discuss the hypothetical nature of the task in the Discussion, referencing the very helpful paper suggested by the reviewer (Brañas-Garza et al. (2024)). As we note there, and also in response to Reviewer 1, Point 1, this design choice was specifically made to balance ecological validity with the practical feasibility of running a large-scale study in a public setting. We have added

a brief rationale for the use of a hypothetical task in our context in the revised Methods section.

4. The strategy space is binary (which I understand is reasonable), but this has implications for the results, since participants can only accept or reject the offers they receive (hot), and therefore we don't know the minimum acceptable offer, which would be desirable. It might be a good idea to explore the MAO with the existing data, if possible.

We acknowledge that the binary accept/reject format in respect to highly generous and highly selfish offers limits the inferences we can draw about participants' minimum acceptable offers (MAO). While our current dataset does not allow for direct estimation of MAO, we agree this would be a valuable direction for future work. We now include this within our limitation section and suggest follow-up studies to estimate MAO distributions in lab-in-the-field settings.

5. There is a substantial body of literature on experiments conducted outside controlled lab settings. Papers 2 and 3 are examples, but there are many more.

References

1. https://www.ub.edu/sustainwell-eu-project/wp-content/uploads/2024/12/LOYOLA_Paying-1-or-nothing-in-dictator-games_Unexpected-differences.pdf

2. https://www.nature.com/articles/srep01213

3. https://doi.org/10.1016/j.socec.2022.101910

We are immensely grateful to the reviewer for these highly relevant references which help us better position our approach in the context of existing literature. We have now incorporated them into the revised manuscript. Specifically, we refer to Brañas-Garza et al. (2024) in our mention of prior hypothetical designs, and their validity in interpreting findings of (incentivised) economic games. We cite Exadaktylos et al. (2013) when referencing literature probing the utility (and similarities) between heterogenous subject pools compared to student samples, and refer to Prisse & Jorrat (2022) when situating our study within the broader literature on economic games which have been conducted in controlled versus (comparatively) uncontrolled environments.

As noted in our response to Reviewer 2, Point 1, and now with these helpful additions, we have revised sections of the Introduction and Discussion to engage more directly with this body of work and to better contextualize our study in relation to lab, field, and online approaches. These revisions clarify how our design contributes to a growing class of field-embedded behavioural paradigms that support scalable, minimally

scripted assessment of social decision-making outside conventional laboratory setups.