THE ROYAL SOCIETY PUBLISHING

PROCEEDINGS B

A preference to learn from successful rather than common behaviours in human social dilemmas

Maxwell N. Burton-Chellew and Victoire D'Amico

Article citation details
Proc. R. Soc. B **288**: 20211590.
http://dx.doi.org/10.1098/rspb.2021.1590

Review timeline

Original submission:	1 March 2021
1st revised submission:	14 July 2021
2nd revised submission:	7 October 2021
3rd revised submission:	19 November 2021
Final acceptance:	29 November 2021

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2021-0512.R0 (Original submission)

Review form: Reviewer 1 (Pat Barclay)

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Reports © 2021 The Reviewers; Decision Letters © 2021 The Reviewers and Editors; Responses © 2021 The Reviewers, Editors and Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is an interesting manuscript that tests one of the major assumptions of gene-culture coevolutionary models of cooperation. Those models claim that people will copy common behaviours like cooperation, even when such behaviours result in a net negative payoff, but then the behaviours persist because of differential success of groups with high cooperation. A key assumption is that people will actually copy cooperation even when that cooperation is costly. This manuscript tests that key assumption, and shows that people have a much greater tendency to copy successful behaviours than common behaviours. The authors also show that this copying of successful behaviours will undermine cooperation, given that group cooperation is individually costly. As such, this manuscript shows that gene-culture coevolutionary models of cooperation rest on an incorrect assumption: that cooperation is readily copied once it's common, despite the cost. This is strong manuscript that makes a useful contribution to the literature on cooperation.

I have some comments to improve the manuscript, but these should be interpreted as minor revisions. None of them affect the publishability of the manuscript.

First, sometimes the writing is not clear about whether participants had access to the social norm in all conditions. It's written clearly in some places, but in other places it's easy for readers to forgot this, and incorrectly assume that there is a one condition with zero information (i.e., not even the social norm) or that participants have access to success info but not the overall average. The authors should read through again with this in mind, to make sure it's clear that participants always have access to the norm, and in some conditions they have additional information as well.

Second, Figure 4 shows that people do clearly copy social norms, at least when they don't have info on successful behaviors. As such, the strategy seems to be copy success if you know who's successful, otherwise copy the norm. A proponent of cultural group selection might respond that this is evidence in favour of their theory, given that people often don't know which strategies are successful. If the authors wish to dispute gene-culture co-evolutionary theories of cooperation, then they should discuss this possibility: that people copy successful others when it's clear who is successful (or which strategies are successful), and copy common behaviours when it's not clear who or what is successful. When success is clear or the costs of cooperation are high, then the authors' conclusions hold. When success is opaque and the costs of cooperation are low, then there is still room for copying common behaviours.

Third, the authors should include effect sizes on their comparisons.

Very Minor Comments:

Figure 2 captions says “empty grey circles”, but they’re square/rectangles in my pdf - relabel

Figure 3 shows the full data distribution, which is great. Many readers aren’t yet used to comparing full distributions, and are more familiar with error bars – is there a way to add error bars in addition to what’s already there? For example, “pirate plots” include this, and are easy to implement in R. This might not be possible if you also have the dashed lines with mean social info shown; if not, it’s not necessary, because this is just a suggestion. But if it is possible, it would help.

Figure 2 & 3 captions describe a comparison of Shown+Success versus Shown+Free-Riders. However, the wording is unclear about whether this a comparison of Shown Social Norm versus Show+Free-Riders (i.e., instead of Shown+Success vs. Shown+Free-Riders). This could be written more clearly so that readers know for sure which comparison is going into those stats

Review form: Reviewer 2

Recommendation
Reject – article is not of sufficient interest (we will consider a transfer to another journal)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Thank you for the opportunity to review this paper, and please accept my apologies for the delay in providing my report.

This paper examines human social learning in a cooperative setting. The authors test assumptions of popular models of cultural evolution of cooperation which have not received due empirical scrutiny. They present a set of economic experiments involving social dilemmas to examine whether people's cooperativeness is based on following a 'social norm' (common behaviour in a group facing the same decision) or 'success' (the behaviour of individuals achieving the highest payoffs). On aggregate, experimental participants were more interested in observing success than social norms, and successful behaviours were also more likely to be copied. It is argued that the results provide evidence for 'success-psychology' and against 'norm-psychology'.

This paper poses important questions, and I am sympathetic to the authors' aims to critically test the assumptions about learning strategies underlying models of gene-culture coevolution. Overall, the paper is easy to follow and the topic will be of interest to a wide readership. However, I think the experimental design and associated analyses limits the authors' ability to draw strong conclusions. I list my concerns below.

MAJOR POINTS

1. The presented experiments were designed to cleanly distinguish whether individuals prefer to follow the social norm or be successful (cf. lines 89-90). To this end, two treatments are presented, one showing the social norm, and another treatment also showing the social norm, but also the behaviour of the highest earners. Why did the authors choose this cumulative design? For a 'clean' experimental examination it seems necessary to have a treatment in which only the behaviour of the highest earners is shown.

2. The abstract suggests that copying successful others is the driving force behind the demise of cooperation. However, the role of 'copying' is rather unclear. In a public goods game (PGG) with an action set of [0,20], exactly copying the mean contributions of others is a special case. Ample research with this paradigm shows that many people's response patterns reflect 'imperfect conditional cooperation' (e.g., Fischbacher & Gaechter 2010 AER), meaning that they do not exactly copy the mean contributions of others, but systematically contribute bit less than that, presumably explaining the decline of cooperation when interactions are repeated. The current paper summarises the literature on this point by saying that in PGGs, people tend to match others' behaviour 'approximately', suggesting that exact copying is actually not to be expected to occur at high frequencies. And indeed, the results show that exact matching is quite rare. Because of the low copying frequency, it seems implausible that copying – in this narrow sense as it is analysed in this paper (lines 217-237) - is the driving force behind the demise of cooperation, as suggested in the abstract.

3. The operationalisation of 'social norm' in the experiment is problematic. The literature on social norms is huge and it seems that there isn't much agreement on basic definitions. However, something that most scholars do seem to agree on is that social norms are rules that apply to certain specific situations ("in situation X, do Y"). In the current study, however, the shown social norm is derived from a strategically different situation than the participants in are facing. In the main experiment, the situation is a public goods game without punishment, and in the previous experiment, the situation is a public goods game with punishment. This difference leads to three problems in the manuscript: (i) it is unclear whether or not it is actually a relevant social norm that was shown to the participants. (ii) The possibility of punishment changes whether or not cooperation would be a 'successful' behaviour, which is one of the central aspects

of this study. (iii) In multiple places the authors indicate that no deception has been used in this study, but I think this setup is in a grey area.

4. The analysis of the PGG data is unclear. The GLMM reported on lines 141-143 is a binomial model fitted to values varying between 0 and 20. Was the dependent variable (contribution) first normalised to fall in the range [0, 1]? This requires explanation. In addition, it is not clear what the reported test refers to here. Is it the interaction effect only? What estimates did the main effects have? I think the model results should be reported in full (e.g., in the supplement) so the reader can make up their own mind about your data and this analysis. A similar comment applies to the GLM right after this (lines 145-149). In addition, in this analysis it is unclear why you chose this link function and how breaking down by round accounts for within-participant dependencies. Overall, it seems that your PGG data allows for replacing all piece-wise analyses - including the ones on lines 199-208 - by fitting a single regression model to all data.

5. The decline of cooperation over time (Fig. 2) is interesting. This commonly observed pattern is often attributed to limitations to reciprocity (imperfect conditional cooperation and the downward adjustment of beliefs about partners' cooperativeness; see for example the Fischbacher and Gaechter paper cited above). However, the current results suggest that this effect also emerges in the absence of feedback from within the group. Some interpretation of the observed effect would enhance the paper.

6. The experiments in which participants had to choose which social information to observe lack a clear motivation. The manuscript would benefit from making explicit why it is interesting to compare a forced choice with a costly one. It is also unclear from the main text whether or not the participants the same participants to the main experiment. I had a similar feeling about comparing the strategy method before and after the repeated public goods game. What do we learn from this in light of the theories on norm-psychology vs success-psychology?

7. The overall conclusion that most individuals preferred to copy examples of successful behaviour (line 309-310) is somewhat misleading. Indeed, 55% of participants copied that successful behaviour at least once (out of five rounds where it was possible), but the overall rate of copying this behaviour was only 9%.

8. I note that the study was not pre-registered. I realise that not much can be done about that anymore, but I wonder what was the reason for not doing this.

MINOR POINTS

- I believe Proceedings B allows for putting the Methods right after the Introduction. Although I think in terms of structure, the manuscript reads fine as is, doing this might clarify some of the experimental details before seeing the results (e.g., that the highest payoffs from the previous experiment were calculated before any punishment took place).

- The examples in the introduction seem to be far removed from the principles they intend to illustrate. The connection between genetic evolution and 'tipping a waiter' seems rather indirect.
- The model results at the bottom of caption 2 duplicate the main text.
- The current study aims to disentangle whether people follow of social norms or successful others (lines 89-102). The first cited reason relates to 'previous studies' but no references are given so it's unclear which work the authors seek to contrast their study with. Some relevant papers here include Frey & Meier (2004), Fowler and Christakis (2010), and Nook et al (2016; reference below.

TYPOS / SMALL STUFF

The references to the treatments are inconsistent throughout the manuscript. The short-hand 'Shown' for 'Shown social norm' are not very descriptive. In the text sometimes the 'Shown' is omitted, which is somewhat confusing.
line 34: whom --> who
line 54: the reference of the parenthesised 'social preference' could be tightened.
line 54: preference --> preferences
line 94-95: in this context I am not sure how to interpret the quotation marks around 'real world'
line 96: nom--> norm
line 98: allowed --> allowing
lines 101-102: not really sure if 'social interactions' as such is an appropriate generic umbrella term for 'reciprocity, signalling, revenge etc'. Maybe these phenomena manifest in social interactions, but they are not the same thing.
lines 105-106: 'costly cooperative game' --> 'game involving (individually) costly cooperation'
line 111: "This mean that the participants knew, in theory, how to be successful." apart from the typo (mean --> means), not sure what you mean by 'in theory'
line 156: add 'MU' to Shown + Success = 2.9 MU (15%)
line 188: why is Success with a capital S?
line 193: allow --> allows

References

Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. American economic review, 100(1), 541-56.
Fowler, J. H. & Christakis, N. A. (2010) Cooperative behavior cascades in human social networks. Proceedings of the National Academy of Sciences 107, 5334–5338, https://doi.org/10.1073/pnas.0913149107.
Frey, B. S. & Meier, S. (2004) Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment. The American Economic Review 94, 1717–1722, https://doi.org/10.1257/0002828043052187.
Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P. & Zaki, J. (2016) Prosocial Conformity: Prosocial Norms Generalize Across Behavior and Empathy. Personality and Social Psychology Bulletin 42, 1045–1062, https://doi.org/10.1177/0146167216649932.

Review form: Reviewer 3

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Marginal

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Marginal

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
No

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
Review for "Humans prefer to copy success rather than the social norm in a cooperative game"

I read this paper "Humans prefer to copy success rather than the social norm in a cooperative game" with much interest. I feel like I should disclose that, prior to being asked to review this paper, I had followed some discussion on social media around its pre-print. That discourse informs my review and I echo some of the points raised here.

This paper presents an experiment where Swiss college students play an economic game under laboratory conditions and are given information about the behavior and success of players in another laboratory experiment. The paper examines how much players contribute to a standard public goods game under different treatments where they have different types of information about the players' success from the previous experiment. The authors seek to use their results to distinguish between what they call "success psychology" from an existing theoretical paradigm called "norm psychology." However, the experiment fails to do this final part due to misunderstandings of the theory behind and implications of norm psychology (and related theories) in the literature.

Overall, I do not think there is enough correspondence between the experiments and the theory the experiments are said to address for this paper to be published in its current form without at least a major revision. The experimental results could, perhaps be useful if the author narrows the scope of the discussion to estimating the relative contributions of success and conformist learning to play in a public goods game in a specific population. It could be that there has already been work on this that I am unaware of, but McElreath et al (2008) – who performed a similar study with a different payoff structure - could be a guide for how a paper like this would work.

Misunderstandings in the paper.

The paper has multiple misunderstandings of the norm psychology theory. The most important of these misunderstandings, in rough order of importance, for this paper are that:

(A) The paper tests whether players will use conformist learning to follow norms in the laboratory experiment. However, norm psychology is premised, by its proponents, to operate over years and decades of a person's social development. It is not premised to operate, especially in enculturated adults, over the course of a brief experiment. For this reason, experimental tests have relied on examining differences in experimental play between societies (where individuals

may have internalized different norms) or in the same society in children at different developmental ages (to examine the process by which norms are internalized during development). The premise, therefore is that norms have already formed or are in the process of forming *outside* of the experiment and then are applied in an experimental context. Therefore, failure to find norm formation in a short economic game is not a failure of norm psychology. See Henrich and Muthukrishna (2021) page 212 and Apicella and Silk (2019, cited by author) page R450 for summaries of cross-cultural and developmental studies. Richerson and Henich (2012) explain this in more depth.

(B) The broader theory of how norm psychology operates with cooperative traits is premised on the interaction between conformity and other institutions, such as punishment, rewards, and reciprocity. Therefore, *even if* cooperative norms were to develop a laboratory experiment, removing punishment, rewards, and reciprocity from consideration, as is done in this experiment, is not a proper test of norm psychology. Henrich and Boyd (2001) discuss how conformist transmission and punishment can work together to stabilize norms in cooperative dilemmas. Chudek and Henrich (2011, cited by the author) also make this point repeatedly.

In this paper the author showed participants data from a previous experiment where cooperative norms are stabilized by punishment. However, the author showed participants data of payoffs that did *not* include the costs of being punished (lines 431-437)! This is not an accurate reflection of how institutions for collective action are supposed to work with success-based learning. According to the theories under question, social norms are maintained specifically because violating the norm is costly. Therefore success-based learning enforces the norm when norm violators are punished. Conformist learning just helps people learn about norms when the connect to payoffs are unclear. In the experiment the author destroys this key premise of the theory by severing the link between success and norm enforcement via punishment. The experiment would align with theory if individuals were show the actual payoffs from the previous experiment instead of excluding the effects of punishment.

In short, it would be more accurate to say that, according to theory, what the author calls a "success psychology" is a potential component of a larger "norm psychology."

(C) The author sets up a dichotomy between theories of cooperation based on genetic evolution vs those based on cultural evolution and says that these theories are in conflict. However, the theories under consideration are actually those based on genetic transmission alone and those based on genes *plus* cultural transmission (sometimes called "dual inheritance" or "gene-culture coevolution"). In the genes plus culture theories of cooperation, culture evolution creates institutions that align genetic success and self-interest with cooperative behavior. Successful individuals are those that tend to follow cultural norms. It is incorrect, therefore, to say that these theories pit norm adoption against genetic or material self-interest. Chudek and Henrich (2011), which the author cites, makes this point repeatedly throughout. As do other articles the author cites, such as Henrich 2003.

This is from Chudek and Henrich (2011):

"The interaction between culture and genes is continuous. The more genes respond by building and honing the above-described norm-psychology, the more they power up the cultural processes that generate and sustain local phenotypic assortment, sanction deviations within groups and select for more cooperative norms. This creates a culture–gene coevolutionary ratchet for both the importance of social norms and the intensity of prosociality. As cultural group selection increasingly guarantees that learners find themselves in social groups organized by norms that incentivize prosocial or cooperative behavior, within-group (and between-group) genetic selection processes will favor genes that build prosocial, norm-adhering phenotypes. This evolutionary trajectory – from cultural learning, to norm-psychology, to cultural group selection for prosocial norms to psychological adaptation to a world dominated by prosocial norms – may help explain some of the puzzling prosocial experimental results that have been dubbed 'strong

reciprocity'."
Previous experimental results

This paper does not reference or engage with previous experiments on success vs conformist transmission, with the most important of these, McElreath et al (2008), published in Proc B some years ago. McElreath et al used an inter-generational experiment and found that participants in their experiments used a mix of success based and conformist learning, but used success-based learning more. A difference is that McElreath et al used a multi-armed bandit problem instead of a public goods game, so this paper has a different set-up in that it uses a cooperative game. The author should discuss why the cooperative game is a better model for the question at hand.

McElreath et al also used a statistical modeling approach that used an explicit model of population dynamics to estimate the mix of conformist and social learning the participants employed. This paper would be much strengthened by using a similar approach. The author should at least engage with previous results.

McElreath et al also reference additional "multi-generational" experiments where participants observe payoffs and strategies of participants in previous games. This is salient because the author of the paper under review implies that having participants observe strategies and payoffs from previous players is an innovation. However, the author should put this design in the context of previous experimental results looking at multi-generational social learning experiments.

More narrowly focus the theoretical scope of this paper

I think this paper could be more accurately written as a simple examination of to what extent Swiss college students tend to use success-based learning and conformist learning in a public goods experiment. We already have well established theories about how conformist learning should be used less when there is better information about the relative success of different behaviors and this study seem to confirm this theory. Models of conformist transmission, which pre-date and are more general than the norm psychology theories, operate under the premise that the conformist learning occurs when the connection between behaviors and success is unclear (please see Henrich and McElreath (2007), section 38.2.3, page 562 and models referenced therein). Therefore, it is not surprising that adding success information decreases conformist learning. I think the author's attempts to use this experiment to test the more encompassing theory of norm psychology falls flat because they do not adequately account for the assumptions of those theories and the author misstates the premises of those theories in many cases. However, applying these results more narrowly to conformist vs success-based learning does not require as many assumptions. However the author would need to be clear that the results might not generalize outside of the study population or outside of the experimental context.

I also think this would be a better paper if the author determined the amount of conformist vs success-based learning the participants engaged in. If participants are engaging in a mix of both, it would be better if the author explicitly modeled this possibility. Perhaps the methods in McElreath et al (2008) would be of use.

Specific references:

The author cites references 11-17 as describing the theory of cooperation that they are testing. I think three of these are as close to canonical as these things get: Henrch (2004), Chudek and Henrich (2011), and Richerson et al (2016) and I am familiar with two others (Apicella and Silk 2019 and Handley and Mathew 2020). I would add a few more as part of the cannon.

Henrich and Boyd (2001) on conformist transmission and cooperation.
Richerson and Henirch (2012) on cultural evolution and collective action problems.

Henrich and Muthukrishna (2021) on cultural evolution and human cooperation.

Minor comments:

Title: The title "Humans prefer…" is overly broad. This study was not conducted with a random sample of humans. It was conducted with a sample of "mostly students enrolled at either UNIL or Swiss Federal Polytechnic School." It is, by now, well established that individuals from Western Educated Industrialized Rich and Democratic groups play and interpret experimental games differently than people from other societies. See papers and books on WEIRD societies and economic games by Henrich and collaborators. The author should use a title that is more reflective of the population with which the research was conducted.
Line 51: Instead of "cannot be explained by genetical evolution" it would be more accurate to say: "cannot be explained by genetical evolution alone" or better "cannot be explained by genetical evolution without another transmission mechanism." The author should update this to better reflect the theory.
Line 58: "evolved culturally, through behavioral, rather than genetical, copying of traits" I don't understand what "behavioral" means in this context. Both cultural and genetic traits can influence behavior and can therefore be considered "behavioral." A better description of the theory would be something like "have evolved cultural transmission systems that co-evolved with genetic selection to stabilize norms and that that between group processes tend to select for the more cooperative norms." The author should update this to better reflect the theory.

Line 59: "against our material or genetic interests." See misunderstanding (C) above.. Norm psychology is premised on the assumption that people evolve to adopt norms because it *is* in their material and genetic interests to do so. This is clear from the theory as described in the papers the author cites and elsewhere. The author should update this to better reflect the theory. One place this is stated concisely is in Boyd (2017, pg 187):

"Cultural group selection models assume that behavior within groups is motivated by individual self-interest. Norms are maintained by rewards and punishments that make it beneficial to follow the norms. If individuals did not benefit from conforming to norms, then the cultural group selection hypothesis would be falsified."
Lines 104-105: "We experimentally tested if individuals prefer to copy either the social (norm psychology hypothesis) or successful behaviors (success-psychology hypothesis) as described above, norm psychology is premised to occur over years and decades of development and align norm compliance with payoffs. Therefore, there will typically be no conflict between success-based and norm-based learning. The paper sets up a strange dichotomy here. It would be more accurate to say that this "success-psychology" is part of a larger "norm psychology," but I find this dichotomy an ill fit for the theory.

Lines 117-118 and 187: At first I found these two descriptions of the experimental set-up confusing: "We also ensured individuals observed a stable social norm by allowing individuals in the previous experiment to punish each other, which stabilized cooperation." and "In the standard public goods game, the highest earning individual are those that contribute the least. Therefore, when we showed individuals examples of successful behaviors, we were also showing them the behavior of those individuals who contributed the least." If punishment stabilized cooperation in the original experiment, presumably free-riders are punished enough that their payoffs were lower than cooperators. However, it is not until line 431 that we find out that the experimenter did not show the actual payoffs to the participants. Instead, the experimenter removed the costs of being punished from the player payoffs. As described above, this violates one of the premises of the theory under question, but in any case the author should explain this earlier to avoid confusion earlier in the paper.

Line 315. "Even if humans…" Again, this is an overly broad statement from research conducted on one WEIRD sample of humans. The author should re-write this sentence in a way that does

not imply that the experimental results apply to humans generally.

References:

Apicella, C. L., & Silk, J. B. (2019). The evolution of human cooperation. Current Biology, 29(11), R447-R450.

Boyd, R. (2017). A different kind of animal: how culture transformed our species (Vol. 46). Princeton University Press.

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. Trends in cognitive sciences, 15(5), 218-226.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of theoretical biology, 208(1), 79-89.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. Annual Review of Psychology, 72, 207-240.

Henrich, J., & Richerson, P., (2012). Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems. Cliodynamics, 3(1) 38-80.

McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1509), 3515-3528.

Decision letter (RSPB-2021-0512.R0)

04-May-2021

Dear Dr Burton-Chellew:

I am writing to inform you that your manuscript RSPB-2021-0512 entitled "Humans prefer to copy success rather than the social norm in a cooperative game" has, in its current form, been rejected for publication in Proceedings B. I have now received comments from three reviewers and the Associate Editor and as you will see, the reviews diverge. In general, I think that it is an important topic to test empirically, so I am offering you the opportunity to substantially revise your manuscript to address these concerns and resubmit it. I will not repeat the reviewers' and AE's comments here, but do highlight a few points. First, one key issue is the degree to which you are testing gene-culture co-evolution per se, or aspects of it. Many of these concerns might be addressed with a more nuanced consideration of what your particular study can and can't address. Second, I agree with the AE that you need to better define what you mean by "norm" and the degree to which the norm in your study is indeed one. Definitions of norm vary, so more precision would be very helpful in answering many of the concerns raised. Finally, as all of the reviewers mention in different ways, the it is critical to interpreting your results to understand the details of the methods, which are not always clear. I realize that there are space limitations, but please work to present them more clearly. I agree with the AE that moving the methods to the front of the paper would likely help. Of course, please be certain to carefully address each of the reviewers' comments, and please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

1) A 'response to referees' document including details of how you have responded to the comments, and the adjustments you have made.
2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
3) Line numbers in your main document.
4) Data - please see our policies on data sharing to ensure that you are complying (https://royalsociety.org/journals/authors/author-guidelines/#data).

To upload a resubmitted manuscript, log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
This paper aims to test adults' decisions to follow a social norm by copying group-level choices to cooperate, versus copying the most successful strategy. The paper aims to test ideas from gene-culture coevolution by directly pitting the cooperative norm against successful behavior to see if cooperative norms versus self-interest wins out when evidence is provided that such a strategy pays off. The reviewers and myself agree that this is an important question, and I think that providing experimental tests of ideas from gene-culture coevolution is important given the high impact of this idea on the field of human cooperation and social decision making. However, there are several important concerns raised about the theoretical framework and interpretation of the results that should be addressed.

The reviewers have provided thoughtful comments coming from a range of perspectives and here I emphasize additional big-picture points that I think should be addressed and also highlight some additional concerns. One concern I have is the extent to which the 'norm' presented in this study can actually be considered a norm. I understand the logic for presenting information about the "social norm" using data from another group (e.g., in another group where punishment was possible, such that people did cooperate at higher rates). However, I think addressing whether this manipulation actually represents good example of a social norm would be important. For example, while the paper lays out the logic for not using the actual behavior of the group as the model, the flip side is that a normative rule within one's own group likely has greater (psychological) weight that a putative norm from another group. In addition, based on the supplement the data from the past group was not presented as a "rule" but rather as some info about how other people acted on average in a game. Does some group average data actually have the weight of a social norm to participants? While lots of people doing the same thing may be one sign that a given behavioral pattern is normative, it does not seem sufficient to establish the

behavioral pattern as a norm. Thus it would be helpful to see clearer evidence that participants in this game actually considered the group average data to represent something 'normative' in the sense of something that 'ought' to be done, which is the way it is used by many social psychologists as well as proponents of gene-culture coevolution. This may require additional data collection looking at (for example) how participants actually understood and interpreted the group average data, or by directly manipulating the presentation of this data to make it appear more or less normative in different conditions. R3 also highlights a large body of work looking at the psychology of norms that may be relevant here.

Second, it seems important to assess the ways in which the full scope of the data here does or does not support gene-culture coevolution, as one benefit of this set of experiments is that it reveals the richness of people's responses across several contexts and thus provides nuanced set of data to evaluate this theory. First, I note the points by R1 noting contexts where copying of norms does appear alignment with that view, and in a similar vein a discussion of the short-term nature of the experiment versus the long-term view of norm acquisition in real life as raised by R3. Second, the points by R2 about the importance of exact copying (and the lack of exact copying here) seems relevant for considering what kind of evidence actually supports this theory or not. I will say that while I do think that ideas from gene-culture coevolution can be fruitfully studied in the short term and in lab experiments, the paper would be strengthened by making clear the benefits versus limitations of a short-term lab experiment compared to the scope of the theories being tested.

Finally, I agree that rearranging the order of the paper (e.g., methods before results) would make the paper more comprehensible to readers. I would further note that more concrete details about the specific way the game was played in the main text, which are currently described in abstract terms in the main text with the full script only in the supplement, does seem relevant here.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
This is an interesting manuscript that tests one of the major assumptions of gene-culture coevolutionary models of cooperation. Those models claim that people will copy common behaviours like cooperation, even when such behaviours result in a net negative payoff, but then the behaviours persist because of differential success of groups with high cooperation. A key assumption is that people will actually copy cooperation even when that cooperation is costly. This manuscript tests that key assumption, and shows that people have a much greater tendency to copy successful behaviours than common behaviours. The authors also show that this copying of successful behaviours will undermine cooperation, given that group cooperation is individually costly. As such, this manuscript shows that gene-culture coevolutionary models of cooperation rest on an incorrect assumption: that cooperation is readily copied once it's common, despite the cost. This is strong manuscript that makes a useful contribution to the literature on cooperation.

I have some comments to improve the manuscript, but these should be interpreted as minor revisions. None of them affect the publishability of the manuscript.

First, sometimes the writing is not clear about whether participants had access to the social norm in all conditions. It's written clearly in some places, but in other places it's easy for readers to forgot this, and incorrectly assume that there is a one condition with zero information (i.e., not even the social norm) or that participants have access to success info but not the overall average. The authors should read through again with this in mind, to make sure it's clear that participants always have access to the norm, and in some conditions they have additional information as well.

Second, Figure 4 shows that people do clearly copy social norms, at least when they don't have info on successful behaviors. As such, the strategy seems to be copy success if you know who's

successful, otherwise copy the norm. A proponent of cultural group selection might respond that this is evidence in favour of their theory, given that people often don't know which strategies are successful. If the authors wish to dispute gene-culture co-evolutionary theories of cooperation, then they should discuss this possibility: that people copy successful others when it's clear who is successful (or which strategies are successful), and copy common behaviours when it's not clear who or what is successful. When success is clear or the costs of cooperation are high, then the authors' conclusions hold. When success is opaque and the costs of cooperation are low, then there is still room for copying common behaviours.

Third, the authors should include effect sizes on their comparisons.

Very Minor Comments:

Figure 2 captions says "empty grey circles", but they're square/rectangles in my pdf - relabel

Figure 3 shows the full data distribution, which is great. Many readers aren't yet used to comparing full distributions, and are more familiar with error bars – is there a way to add error bars in addition to what's already there? For example, "pirate plots" include this, and are easy to implement in R. This might not be possible if you also have the dashed lines with mean social info shown; if not, it's not necessary, because this is just a suggestion. But if it is possible, it would help.

Figure 2 & 3 captions describe a comparison of Shown+Success versus Shown+Free-Riders. However, the wording is unclear about whether this a comparison of Shown Social Norm versus Show+Free-Riders (i.e., instead of Shown+Success vs. Shown+Free-Riders). This could be written more clearly so that readers know for sure which comparison is going into those stats

Referee: 2
Comments to the Author(s)
Thank you for the opportunity to review this paper, and please accept my apologies for the delay in providing my report.

This paper examines human social learning in a cooperative setting. The authors test assumptions of popular models of cultural evolution of cooperation which have not received due empirical scrutiny. They present a set of economic experiments involving social dilemmas to examine whether people's cooperativeness is based on following a 'social norm' (common behaviour in a group facing the same decision) or 'success' (the behaviour of individuals achieving the highest payoffs). On aggregate, experimental participants were more interested in observing success than social norms, and successful behaviours were also more likely to be copied. It is argued that the results provide evidence for 'success-psychology' and against 'norm-psychology'.

This paper poses important questions, and I am sympathetic to the authors' aims to critically test the assumptions about learning strategies underlying models of gene-culture coevolution. Overall, the paper is easy to follow and the topic will be of interest to a wide readership. However, I think the experimental design and associated analyses limits the authors' ability to draw strong conclusions. I list my concerns below.

MAJOR POINTS

1. The presented experiments were designed to cleanly distinguish whether individuals prefer to follow the social norm or be successful (cf. lines 89-90). To this end, two treatments are presented, one showing the social norm, and another treatment also showing the social norm, but also the behaviour of the highest earners. Why did the authors choose this cumulative design? For a 'clean' experimental examination it seems necessary to have a treatment in which only the behaviour of the highest earners is shown.

2. The abstract suggests that copying successful others is the driving force behind the demise of cooperation. However, the role of 'copying' is rather unclear. In a public goods game (PGG) with an action set of [0,20], exactly copying the mean contributions of others is a special case. Ample research with this paradigm shows that many people's response patterns reflect 'imperfect conditional cooperation' (e.g., Fischbacher & Gaechter 2010 AER), meaning that they do not exactly copy the mean contributions of others, but systematically contribute bit less than that, presumably explaining the decline of cooperation when interactions are repeated. The current paper summarises the literature on this point by saying that in PGGs, people tend to match others' behaviour 'approximately', suggesting that exact copying is actually not to be expected to occur at high frequencies. And indeed, the results show that exact matching is quite rare. Because of the low copying frequency, it seems implausible that copying – in this narrow sense as it is analysed in this paper (lines 217-237) - is the driving force behind the demise of cooperation, as suggested in the abstract.

3. The operationalisation of 'social norm' in the experiment is problematic. The literature on social norms is huge and it seems that there isn't much agreement on basic definitions. However, something that most scholars do seem to agree on is that social norms are rules that apply to certain specific situations ("in situation X, do Y"). In the current study, however, the shown social norm is derived from a strategically different situation than the participants in are facing. In the main experiment, the situation is a public goods game without punishment, and in the previous experiment, the situation is a public goods game with punishment. This difference leads to three problems in the manuscript: (i) it is unclear whether or not it is actually a relevant social norm that was shown to the participants. (ii) The possibility of punishment changes whether or not cooperation would be a 'successful' behaviour, which is one of the central aspects of this study. (iii) In multiple places the authors indicate that no deception has been used in this study, but I think this setup is in a grey area.

4. The analysis of the PGG data is unclear. The GLMM reported on lines 141-143 is a binomial model fitted to values varying between 0 and 20. Was the dependent variable (contribution) first normalised to fall in the range [0, 1]? This requires explanation. In addition, it is not clear what the reported test refers to here. Is it the interaction effect only? What estimates did the main effects have? I think the model results should be reported in full (e.g., in the supplement) so the reader can make up their own mind about your data and this analysis. A similar comment applies to the GLM right after this (lines 145-149). In addition, in this analysis it is unclear why you chose this link function and how breaking down by round accounts for within-participant dependencies. Overall, it seems that your PGG data allows for replacing all piece-wise analyses - including the ones on lines 199-208 - by fitting a single regression model to all data.

5. The decline of cooperation over time (Fig. 2) is interesting. This commonly observed pattern is often attributed to limitations to reciprocity (imperfect conditional cooperation and the downward adjustment of beliefs about partners' cooperativeness; see for example the Fischbacher and Gaechter paper cited above). However, the current results suggest that this effect also emerges in the absence of feedback from within the group. Some interpretation of the observed effect would enhance the paper.

6. The experiments in which participants had to choose which social information to observe lack a clear motivation. The manuscript would benefit from making explicit why it is interesting to compare a forced choice with a costly one. It is also unclear from the main text whether or not the participants the same participants to the main experiment. I had a similar feeling about comparing the strategy method before and after the repeated public goods game. What do we learn from this in light of the theories on norm-psychology vs success-psychology?

7. The overall conclusion that most individuals preferred to copy examples of successful behaviour (line 309-310) is somewhat misleading. Indeed, 55% of participants copied that

successful behaviour at least once (out of five rounds where it was possible), but the overall rate of copying this behaviour was only 9%.

8. I note that the study was not pre-registered. I realise that not much can be done about that anymore, but I wonder what was the reason for not doing this.

MINOR POINTS

• I believe Proceedings B allows for putting the Methods right after the Introduction. Although I think in terms of structure, the manuscript reads fine as is, doing this might clarify some of the experimental details before seeing the results (e.g., that the highest payoffs from the previous experiment were calculated before any punishment took place).
• The examples in the introduction seem to be far removed from the principles they intend to illustrate. The connection between genetic evolution and 'tipping a waiter' seems rather indirect.
• The model results at the bottom of caption 2 duplicate the main text.
• The current study aims to disentangle whether people follow of social norms or successful others (lines 89-102). The first cited reason relates to 'previous studies' but no references are given so it's unclear which work the authors seek to contrast their study with. Some relevant papers here include Frey & Meier (2004), Fowler and Christakis (2010), and Nook et al (2016; reference below.

TYPOS / SMALL STUFF

The references to the treatments are inconsistent throughout the manuscript. The short-hand 'Shown' for 'Shown social norm' are not very descriptive. In the text sometimes the 'Shown' is omitted, which is somewhat confusing.
line 34: whom --> who
line 54: the reference of the parenthesised 'social preference' could be tightened.
line 54: preference --> preferences
line 94-95: in this context I am not sure how to interpret the quotation marks around 'real world'
line 96: nom--> norm
line 98: allowed --> allowing
lines 101-102: not really sure if 'social interactions' as such is an appropriate generic umbrella term for 'reciprocity, signalling, revenge etc'. Maybe these phenomena manifest in social interactions, but they are not the same thing.
lines 105-106: 'costly cooperative game' --> 'game involving (individually) costly cooperation'
line 111: "This mean that the participants knew, in theory, how to be successful." apart from the typo (mean --> means), not sure what you mean by 'in theory'
line 156: add 'MU' to Shown + Success = 2.9 MU (15%)
line 188: why is Success with a capital S?
line 193: allow --> allows

References

Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. American economic review, 100(1), 541-56.
Fowler, J. H. & Christakis, N. A. (2010) Cooperative behavior cascades in human social networks. Proceedings of the National Academy of Sciences 107, 5334–5338, https://doi.org/10.1073/pnas.0913149107.
Frey, B. S. & Meier, S. (2004) Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment. The American Economic Review 94, 1717–1722, https://doi.org/10.1257/0002828043052187.
Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P. & Zaki, J. (2016) Prosocial Conformity: Prosocial Norms Generalize Across Behavior and Empathy. Personality and Social Psychology Bulletin 42, 1045–1062, https://doi.org/10.1177/0146167216649932.

Referee: 3
Comments to the Author(s)
Review for "Humans prefer to copy success rather than the social norm in a cooperative game"

I read this paper "Humans prefer to copy success rather than the social norm in a cooperative game" with much interest. I feel like I should disclose that, prior to being asked to review this paper, I had followed some discussion on social media around its pre-print. That discourse informs my review and I echo some of the points raised here.

This paper presents an experiment where Swiss college students play an economic game under laboratory conditions and are given information about the behavior and success of players in another laboratory experiment. The paper examines how much players contribute to a standard public goods game under different treatments where they have different types of information about the players' success from the previous experiment. The authors seek to use their results to distinguish between what they call "success psychology" from an existing theoretical paradigm called "norm psychology." However, the experiment fails to do this final part due to misunderstandings of the theory behind and implications of norm psychology (and related theories) in the literature.

Overall, I do not think there is enough correspondence between the experiments and the theory the experiments are said to address for this paper to be published in its current form without at least a major revision. The experimental results could, perhaps be useful if the author narrows the scope of the discussion to estimating the relative contributions of success and conformist learning to play in a public goods game in a specific population. It could be that there has already been work on this that I am unaware of, but McElreath et al (2008) – who performed a similar study with a different payoff structure - could be a guide for how a paper like this would work.

Misunderstandings in the paper.

The paper has multiple misunderstandings of the norm psychology theory. The most important of these misunderstandings, in rough order of importance, for this paper are that:

(A) The paper tests whether players will use conformist learning to follow norms in the laboratory experiment. However, norm psychology is premised, by its proponents, to operate over years and decades of a person's social development. It is not premised to operate, especially in enculturated adults, over the course of a brief experiment. For this reason, experimental tests have relied on examining differences in experimental play between societies (where individuals may have internalized different norms) or in the same society in children at different developmental ages (to examine the process by which norms are internalized during development). The premise, therefore is that norms have already formed or are in the process of forming *outside* of the experiment and then are applied in an experimental context. Therefore, failure to find norm formation in a short economic game is not a failure of norm psychology. See Henrich and Muthukrishna (2021) page 212 and Apicella and Silk (2019, cited by author) page R450 for summaries of cross-cultural and developmental studies. Richerson and Henich (2012) explain this in more depth.

(B) The broader theory of how norm psychology operates with cooperative traits is premised on the interaction between conformity and other institutions, such as punishment, rewards, and reciprocity. Therefore, *even if* cooperative norms were to develop a laboratory experiment, removing punishment, rewards, and reciprocity from consideration, as is done in this experiment, is not a proper test of norm psychology. Henrich and Boyd (2001) discuss how conformist transmission and punishment can work together to stabilize norms in cooperative dilemmas. Chudek and Henrich (2011, cited by the author) also make this point repeatedly.

In this paper the author showed participants data from a previous experiment where cooperative norms are stabilized by punishment. However, the author showed participants data of payoffs that did *not* include the costs of being punished (lines 431-437)! This is not an accurate reflection of how institutions for collective action are supposed to work with success-based learning. According to the theories under question, social norms are maintained specifically because violating the norm is costly. Therefore success-based learning enforces the norm when norm violators are punished. Conformist learning just helps people learn about norms when the connect to payoffs are unclear. In the experiment the author destroys this key premise of the theory by severing the link between success and norm enforcement via punishment. The experiment would align with theory if individuals were show the actual payoffs from the previous experiment instead of excluding the effects of punishment.

In short, it would be more accurate to say that, according to theory, what the author calls a "success psychology" is a potential component of a larger "norm psychology."

(C) The author sets up a dichotomy between theories of cooperation based on genetic evolution vs those based on cultural evolution and says that these theories are in conflict. However, the theories under consideration are actually those based on genetic transmission alone and those based on genes *plus* cultural transmission (sometimes called "dual inheritance" or "gene-culture coevolution"). In the genes plus culture theories of cooperation, culture evolution creates institutions that align genetic success and self-interest with cooperative behavior. Successful individuals are those that tend to follow cultural norms. It is incorrect, therefore, to say that these theories pit norm adoption against genetic or material self-interest. Chudek and Henrich (2011), which the author cites, makes this point repeatedly throughout. As do other articles the author cites, such as Henrich 2003.

This is from Chudek and Henrich (2011):

"The interaction between culture and genes is continuous. The more genes respond by building and honing the above-described norm-psychology, the more they power up the cultural processes that generate and sustain local phenotypic assortment, sanction deviations within groups and select for more cooperative norms. This creates a culture–gene coevolutionary ratchet for both the importance of social norms and the intensity of prosociality. As cultural group selection increasingly guarantees that learners find themselves in social groups organized by norms that incentivize prosocial or cooperative behavior, within-group (and between-group) genetic selection processes will favor genes that build prosocial, norm-adhering phenotypes. This evolutionary trajectory – from cultural learning, to norm-psychology, to cultural group selection for prosocial norms to psychological adaptation to a world dominated by prosocial norms – may help explain some of the puzzling prosocial experimental results that have been dubbed 'strong reciprocity'."

Previous experimental results

This paper does not reference or engage with previous experiments on success vs conformist transmission, with the most important of these, McElreath et al (2008), published in Proc B some years ago. McElreath et al used an inter-generational experiment and found that participants in their experiments used a mix of success based and conformist learning, but used success-based learning more. A difference is that McElreath et al used a multi-armed bandit problem instead of a public goods game, so this paper has a different set-up in that it uses a cooperative game. The author should discuss why the cooperative game is a better model for the question at hand.

McElreath et al also used a statistical modeling approach that used an explicit model of population dynamics to estimate the mix of conformist and social learning the participants employed. This paper would be much strengthened by using a similar approach. The author should at least engage with previous results.

McElreath et al also reference additional "multi-generational" experiments where participants observe payoffs and strategies of participants in previous games. This is salient because the author of the paper under review implies that having participants observe strategies and payoffs from previous players is an innovation. However, the author should put this design in the context of previous experimental results looking at multi-generational social learning experiments.

More narrowly focus the theoretical scope of this paper

I think this paper could be more accurately written as a simple examination of to what extent Swiss college students tend to use success-based learning and conformist learning in a public goods experiment. We already have well established theories about how conformist learning should be used less when there is better information about the relative success of different behaviors and this study seem to confirm this theory. Models of conformist transmission, which pre-date and are more general than the norm psychology theories, operate under the premise that the conformist learning occurs when the connection between behaviors and success is unclear (please see Henrich and McElreath (2007), section 38.2.3, page 562 and models referenced therein). Therefore, it is not surprising that adding success information decreases conformist learning. I think the author's attempts to use this experiment to test the more encompassing theory of norm psychology falls flat because they do not adequately account for the assumptions of those theories and the author misstates the premises of those theories in many cases. However, applying these results more narrowly to conformist vs success-based learning does not require as many assumptions. However the author would need to be clear that the results might not generalize outside of the study population or outside of the experimental context.

I also think this would be a better paper if the author determined the amount of conformist vs success-based learning the participants engaged in. If participants are engaging in a mix of both, it would be better if the author explicitly modeled this possibility. Perhaps the methods in McElreath et al (2008) would be of use.
Specific references:

The author cites references 11-17 as describing the theory of cooperation that they are testing. I think three of these are as close to canonical as these things get: Henrch (2004), Chudek and Henrich (2011), and Richerson et al (2016) and I am familiar with two others (Apicella and Silk 2019 and Handley and Mathew 2020). I would add a few more as part of the cannon.

Henrich and Boyd (2001) on conformist transmission and cooperation.
Richerson and Henirch (2012) on cultural evolution and collective action problems.
Henrich and Muthukrishna (2021) on cultural evolution and human cooperation.

Minor comments:

Title: The title "Humans prefer…" is overly broad. This study was not conducted with a random sample of humans. It was conducted with a sample of "mostly students enrolled at either UNIL or Swiss Federal Polytechnic School." It is, by now, well established that individuals from Western Educated Industrialized Rich and Democratic groups play and interpret experimental games differently than people from other societies. See papers and books on WEIRD societies and economic games by Henrich and collaborators. The author should use a title that is more reflective of the population with which the research was conducted.
Line 51: Instead of "cannot be explained by genetical evolution" it would be more accurate to say: "cannot be explained by genetical evolution alone" or better "cannot be explained by genetical evolution without another transmission mechanism." The author should update this to better reflect the theory.
Line 58: "evolved culturally, through behavioral, rather than genetical, copying of traits" I don't understand what "behavioral" means in this context. Both cultural and genetic traits can influence behavior and can therefore be considered "behavioral." A better description of the

theory would be something like "have evolved cultural transmission systems that co-evolved with genetic selection to stabilize norms and that that between group processes tend to select for the more cooperative norms." The author should update this to better reflect the theory.

Line 59: "against our material or genetic interests." See misunderstanding (C) above.. Norm psychology is premised on the assumption that people evolve to adopt norms because it *is* in their material and genetic interests to do so. This is clear from the theory as described in the papers the author cites and elsewhere. The author should update this to better reflect the theory. One place this is stated concisely is in Boyd (2017, pg 187):

"Cultural group selection models assume that behavior within groups is motivated by individual self-interest. Norms are maintained by rewards and punishments that make it beneficial to follow the norms. If individuals did not benefit from conforming to norms, then the cultural group selection hypothesis would be falsified."
Lines 104-105: "We experimentally tested if individuals prefer to copy either the social (norm psychology hypothesis) or successful behaviors (success-psychology hypothesis) as described above, norm psychology is premised to occur over years and decades of development and align norm compliance with payoffs. Therefore, there will typically be no conflict between success-based and norm-based learning. The paper sets up a strange dichotomy here. It would be more accurate to say that this "success-psychology" is part of a larger "norm psychology," but I find this dichotomy an ill fit for the theory.

Lines 117-118 and 187: At first I found these two descriptions of the experimental set-up confusing: "We also ensured individuals observed a stable social norm by allowing individuals in the previous experiment to punish each other, which stabilized cooperation." and "In the standard public goods game, the highest earning individual are those that contribute the least. Therefore, when we showed individuals examples of successful behaviors, we were also showing them the behavior of those individuals who contributed the least." If punishment stabilized cooperation in the original experiment, presumably free-riders are punished enough that their payoffs were lower than cooperators. However, it is not until line 431 that we find out that the experimenter did not show the actual payoffs to the participants. Instead, the experimenter removed the costs of being punished from the player payoffs. As described above, this violates one of the premises of the theory under question, but in any case the author should explain this earlier to avoid confusion earlier in the paper.

Line 315. "Even if humans…" Again, this is an overly broad statement from research conducted on one WEIRD sample of humans. The author should re-write this sentence in a way that does not imply that the experimental results apply to humans generally.

References:

Apicella, C. L., & Silk, J. B. (2019). The evolution of human cooperation. Current Biology, 29(11), R447-R450.
Boyd, R. (2017). A different kind of animal: how culture transformed our species (Vol. 46). Princeton University Press.

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. Trends in cognitive sciences, 15(5), 218-226.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of theoretical biology, 208(1), 79-89.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. Annual Review of Psychology, 72, 207-240.

Henrich, J., & Richerson, P., (2012). Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems. Cliodynamics, 3(1) 38-80.

McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1509), 3515-3528.

Author's Response to Decision Letter for (RSPB-2021-0512.R0)

See Appendix A.

RSPB-2021-1590.R0

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors have taken serious effort in revising this manuscript. The presentation of the research has much improved. Extending the Methods section in the main text has enhanced the paper by motivating the cumulative design and the forced choice tests. I have a few (rather minor) lingering issues I would like to see addressed before I would recommend publication.

The paper seems to use 'copying', 'social learning', 'following others', and 'matching' interchangeably. In the literature on conditional cooperation, 'copying' (defined as exactly matching behaviour of others) is not as common as the authors make it sound. In particular, the statement that 'most [people] will match their cooperation to local levels' (l. 57-58) and that individuals 'prefer to match the common patterns' (l. 82, 83) is not supported by the cited papers. If anything, the cited papers (most notably Fischbacher et al 2001 but also Fischbacher and Gaechter 2010) show 'imperfect conditional cooperation', in which people undercut rather than match average contributions. Later on, the authors state that 'matching' might be imperfect (l. 135) but it would be helpful to make the writing precise when referring to these concepts from the outset.

Fehr & Gaechter (2002) is about punishment and not about conditional cooperation as studied in the current paper, so this reference should be removed from the citations in l. 83.

I appreciated that the revision discusses the differences between the previous and the main experiment in terms of punishment. However, I did not understand the argument around this on lines 143-148. In particular, I am not sure what you mean with a "correct" definition of success here. This passage also has a few typos.

From reading the experimental instructions cited in the main text I worried that participants might have interpreted success information as stemming from the person who made most money *overall* rather than in that particular round (which would make the success information more about 'prestige'). From the SI I see that in a preceding paragraph, the instructions did make it clear that the success information stemmed from the corresponding round. It would be useful to briefly point out this fact in the main text.

Descriptions of the GLMM analyses have improved and presenting the model results in the SI is very helpful. One minor point: it might be better to avoid the term 'slope' to refer to 'period in the repeated game' because 'slope' is often used to refer to a coefficient.

It would be useful if the authors proof-read the paper once more. The revision has a number of typos. The ones I spotted are listed below:
- l 75: 'they'
- l 186: 'individuals'
- l. 189: 'pattern'
- l. 219: 'group'
- l. 220: 'or an'

Review form: Reviewer 4

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
N/A

Is it clear?
N/A

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is an excellent paper, touching on some important issues at the interface of evolution and human social behavior. Because the paper runs contrary to some popular views, it should be taken all the more seriously as a potentially important contribution. (It is a sad but seemingly inevitable fact of the practice of science--given the psychological and sociological contexts in which science is done--that dissent is pushed to the margins.).

The experiment was well-framed in the introducton, with a solid experimental design and appropriate analysis of data. The conclusions are accurate and not overblown. The paper is also exceptionally well written.

My comments are rather few. All of the comments can be addressed quickly in a very light revision.

1. Lines 131-140: I accept the fact that the data they used to show subjects how previous subjects had contributed in previous games were obtained from games in which high levels of cooperation were sustained by peer punishment. I understand the proximate rationale for doing so (they wanted to avoid giving subjects the expectation that declines in cooperation were in some way normative), but their ultimate rationale was a bit unclear. Did they go to all of this trouble in the first place? Couldn't they have devised other ways to show subjects' previous subjects' choices?

Line 285. The passage beginning "Meaning that the..." is a sentence fragment.

Lines 379-395. In the Forced-Choice test, approximately 38% of people chose to copy the common pattern rather than the successful pattern. In the Costly-Choice test, exactly 35% of the subjects who paid for information did so in order to see the common pattern. These percentages, obviously are identical, suggesting that the free-choice vs. costly-choice manipulaton makes no difference in assessing people's informational preferences. For this reason, I would like to see the results of an binomial sign test in which the data from the two data sets were combined, giving a total of 39 subjects who wanted to see the common pattern and 64 who wanted to see the successful pattern.

Review form: Reviewer 5

Recommendation
Reject – article is scientifically unsound

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Marginal

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
N/A

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
Yes

Comments to the Author
See attached file

Decision letter (RSPB-2021-1590.R0)

07-Sep-2021

Dear Dr Burton-Chellew:

Thank you for resubmitting your revised manuscript to Proceedings B. I have now received two reviews as well as the comments of the Associate Editor. As you will see, the reviews are generally quite positive, and based on my own reading of your manuscript as well as the advice of the AE, I think that with another round of revision your manuscript will likely be acceptable for publication. However, you will see that despite their positive assessments, the reviewers have raised some concerns, and I invite you to address them. Both of the reviewers and the AE have written particularly clear and constructive comments, appended below, so I will spare you repeating them here.

We typically do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor Board Member
Comments to Author:
This is a revision of a paper testing adults' decisions to follow a common strategy by copying group-level choices to cooperate, versus copying the most successful strategy. This is an important question and an interesting way to test ideas from gene-culture coevolution. The reviewers have provided thoughtful comments about the new version of the manuscript that

should be addressed, and I provide some additional comments and highlight some specific aspects I view as particularly important.

Some of the issues raised by reviewers are partially addressed in the manuscript but should be better highlighted to make sure these points are clear to readers. For example, my understanding is that the use of a sample where punishment was possible is used as a model here because this is a well-known way to stabilize cooperation in order to even show examples where other people cooperated, but it would be important to take care that the reasoning for doing so is clear in the paper.

It would also be important to clarify use of terminology like 'copying', 'social learning', or 'following others' as noted by R1. This seems especially important in that a lot of the cultural evolutionary work on social learning is focused on imitation actions (often of a more complex behavioral nature than responses in this kind of economic game), so being clear about what the study is testing and its implication is important. The revisions should also take care to accurately characterize claims about (imperfect) conditional cooperation made by other teams.

R3 raises several questions about the potential ambiguity of the information provided by the "common pattern". I think that if participants are told that a particular response is the most common response (rather than given direct access to the full distribution of participant responses) then some of the concerns raised by R3 may not necessarily apply as such. Nonetheless, clarification about what the common pattern is, and how this information was presented to participants, would be important, as well as some discussion of potential limitations of the current design and next steps that could build on this. The authors should also clarify the question about potential use of deception in the task; while deception can be important in psychology studies to address various questions, it is nonetheless something to clarify about the procedure.

Finally, the revision has adjusted the use of social norm through the paper to the "common" pattern, which makes sense to me. But it still describes this common pattern as representing a social norm throughout, thus undermining the responsiveness of this change. For example, the introduction states that following common patterns means confirming to norms ("desire to follow common social behaviors (conform with social norms)"), later the common pattern is described as a "descriptive social norm" (a term which is not defined in the paper), and the methods and results use the terms social norm interchangeably with common pattern (e.g., "35% of individuals matched the social norm at least once"). Following the comments in the prior version of the manuscript, this work does not demonstrate that the common strategy has the force of a norm the way this is typically used by people working in gene-culture coevolution, and while this is not necessary for the paper to be an interesting contribution it is important to make this clear. A discussion of this issue, including a more nuanced discussion of how this is relevant for theories about gene-culture coevolution (which are actually focused on norms as such) would be relevant.

Reviewer(s)' Comments to Author:
Referee: 2
Comments to the Author(s).
The authors have taken serious effort in revising this manuscript. The presentation of the research has much improved. Extending the Methods section in the main text has enhanced the paper by motivating the cumulative design and the forced choice tests. I have a few (rather minor) lingering issues I would like to see addressed before I would recommend publication.

The paper seems to use 'copying', 'social learning', 'following others', and 'matching' interchangeably. In the literature on conditional cooperation, 'copying' (defined as exactly matching behaviour of others) is not as common as the authors make it sound. In particular, the statement that 'most [people] will match their cooperation to local levels' (l. 57-58) and that individuals 'prefer to match the common patterns' (l. 82, 83) is not supported by the cited papers.

If anything, the cited papers (most notably Fischbacher et al 2001 but also Fischbacher and Gaechter 2010) show 'imperfect conditional cooperation', in which people undercut rather than match average contributions. Later on, the authors state that 'matching' might be imperfect (l. 135) but it would be helpful to make the writing precise when referring to these concepts from the outset.

Fehr & Gaechter (2002) is about punishment and not about conditional cooperation as studied in the current paper, so this reference should be removed from the citations in l. 83.

I appreciated that the revision discusses the differences between the previous and the main experiment in terms of punishment. However, I did not understand the argument around this on lines 143-148. In particular, I am not sure what you mean with a "correct" definition of success here. This passage also has a few typos.

From reading the experimental instructions cited in the main text I worried that participants might have interpreted success information as stemming from the person who made most money *overall* rather than in that particular round (which would make the success information more about 'prestige'). From the SI I see that in a preceding paragraph, the instructions did make it clear that the success information stemmed from the corresponding round. It would be useful to briefly point out this fact in the main text.

Descriptions of the GLMM analyses have improved and presenting the model results in the SI is very helpful. One minor point: it might be better to avoid the term 'slope' to refer to 'period in the repeated game' because 'slope' is often used to refer to a coefficient.

It would be useful if the authors proof-read the paper once more. The revision has a number of typos. The ones I spotted are listed below:
- l 75: 'they'
- l 186: 'individuals'
- l. 189: 'pattern'
- l. 219: 'group'
- l. 220: 'or an'

Referee: 4
Comments to the Author(s).
This is an excellent paper, touching on some important issues at the interface of evolution and human social behavior. Because the paper runs contrary to some popular views, it should be taken all the more seriously as a potentially important contribution. (It is a sad but seemingly inevitable fact of the practice of science--given the psychological and sociological contexts in which science is done--that dissent is pushed to the margins.).

The experiment was well-framed in the introducton, with a solid experimental design and appropriate analysis of data. The conclusions are accurate and not overblown. The paper is also exceptionally well written.

My comments are rather few. All of the comments can be addressed quickly in a very light revision.

1. Lines 131-140: I accept the fact that the data they used to show subjects how previous subjects had contributed in previous games were obtained from games in which high levels of cooperation were sustained by peer punishment. I understand the proximate rationale for doing so (they wanted to avoid giving subjects the expectation that declines in cooperation were in some way normative), but their ultimate rationale was a bit unclear. Did they go to all of this trouble in the first place? Couldn't they have devised other ways to show subjects' previous subjects' choices?

Line 285. The passage beginning "Meaning that the..." is a sentence fragment.

Lines 379-395. In the Forced-Choice test, approximately 38% of people chose to copy the common pattern rather than the successful pattern. In the Costly-Choice test, exactly 35% of the subjects who paid for information did so in order to see the common pattern. These percentages, obviously are identical, suggesting that the free-choice vs. costly-choice manipulaton makes no difference in assessing people's informational preferences. For this reason, I would like to see the results of an binomial sign test in which the data from the two data sets were combined, giving a total of 39 subjects who wanted to see the common pattern and 64 who wanted to see the successful pattern.

Referee: 5
Comments to the Author(s).
See attached file

Author's Response to Decision Letter for (RSPB-2021-1590.R0)

See Appendix B.

Decision letter (RSPB-2021-1590.R1)

16-Nov-2021

Dear Dr Burton-Chellew

I am very pleased to inform you that your manuscript RSPB-2021-1590.R1 entitled "A preference to learn from successful rather than common behaviours in human social dilemmas" has been accepted for publication in Proceedings B.

The Associate Editor has recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the comments in a revision.

I addition, based upon the statements of one of the reviewers regarding the possibility of unauthorized deception in the current study, on 20 October 2021 the Editor for this paper requested clarification from the University of Lausanne board of ethics. The ethics board has verified that the work was fully approved and was completed as per the approved protocols. The ethics board sent us the following statement:
“After checking the ethics submissions of the different projects submitted by Dr. Burton-Chellew, we have concluded that in none of them the problem pointed out by the reviewer was visible.”

Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days. If you do not think you will be able to meet this date please let us know.

To revise your manuscript, log into https://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been

appended to denote a revision. You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referee(s) and upload a file "Response to Referees". You can use this to document any changes you make to the original manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Before uploading your revised files please make sure that you have:

1) A text file of the manuscript (doc, txt, rtf or tex), including the references, tables (including captions) and figure captions. Please remove any tracked changes from the text before submission. PDF files are not an accepted format for the "Main Document".

2) A separate electronic file of each figure (tiff, EPS or print-quality PDF preferred). The format should be produced directly from original creation package, or original software format. PowerPoint files are not accepted.

3) Electronic supplementary material: this should be contained in a separate file and where possible, all ESM should be combined into a single file. All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

4) A media summary: a short non-technical summary (up to 100 words) of the key findings/importance of your manuscript.

5) Data accessibility section and data citation
It is a condition of publication that data supporting your paper are made available either in the electronic supplementary material or through an appropriate repository.

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should be fully cited. To ensure archived data are available to readers, authors should include a 'data accessibility' section immediately after the acknowledgements section. This should list the database and accession number for all data from the article that has been made publicly available, for instance:
• DNA sequences: Genbank accessions F234391-F234402
• Phylogenetic data: TreeBASE accession number S9123
• Final DNA sequence assembly uploaded as online supplemental material
• Climate data and MaxEnt input files: Dryad doi:10.5521/dryad.12311
NB. From April 1 2013, peer reviewed articles based on research funded wholly or partly by RCUK must include, if applicable, a statement on how the underlying research materials – such as data, samples or models – can be accessed. This statement should be included in the data accessibility section.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link

http://datadryad.org/submit?journalID=RSPB&manu=(Document not available) which will take you to your unique entry in the Dryad repository. If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link. Please see https://royalsociety.org/journals/ethics-policies/data-sharing-mining/ for more details.

6) For more information on our Licence to Publish, Open Access, Cover images and Media summaries, please visit https://royalsociety.org/journals/authors/author-guidelines/.

Once again, thank you for submitting your manuscript to Proceedings B and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Sincerely,
Dr Sarah Brosnan
Editor, Proceedings B
mailto:proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
This is a revision of a paper testing adults' decisions to follow a common strategy by copying group-level choices to cooperate, versus copying the most successful strategy. This is a responsive revision that has addressed the key points in the last round of review. My main comment is that the current wording of the text on page 4 implies that there may have been some minor deception in the study. This needs to be appropriately reworded to reflect that there was not deception and to clarify that the study as it was run had appropriate ethics approvals.

Decision letter (RSPB-2021-1590.R2)

30-Nov-2021

Dear Dr Burton-Chellew

I am pleased to inform you that your manuscript entitled "A preference to learn from successful rather than common behaviours in human social dilemmas" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Data Accessibility section
Please remember to make any data sets live prior to publication, and update any links as needed when you receive a proof to check. It is good practice to also add data sets to your reference list.

Open Access

You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
The authors have addressed the few remaining points.

Appendix A

Maxwell N. Burton-Chellew & Victoire D'Amico
Department of Ecology & Evolution
& Department of Economics
University of Lausanne
CH-1015 Lausanne
Switzerland
maxwell.burton@unil.ch

Dear Editors,

Thank you for inviting us to resubmit our paper, originally entitled, "**Humans prefer to copy success rather than the social norm in a cooperative game**" (Manuscript ID RSPB-2021-0512), as a research article for Proceeding of the Royal Society B series. The new title is now, "*A preference to follow successful rather than common behaviours in human social dilemmas.*"

We have substantially revised our manuscript considering the reviewer's comments and editorial requests. Primarily we have shifted our emphasis from a desire to follow descriptive social norms to a preference to follow either common or successful behaviours in social dilemmas, and we have clarified our Methods section and brought it forward to the front-end of the manuscript.

Please find our detailed response to reviewers below.

Best wishes,

Max and Victoire.

04-May-2021
Dear Dr Burton-Chellew:

I am writing to inform you that your manuscript RSPB-2021-0512 entitled "Humans prefer to copy success rather than the social norm in a cooperative game" has, in its current form, been rejected for publication in Proceedings B. I have now received comments from three reviewers and the Associate Editor and as you will see, the reviews diverge. In general, I think that it is an important topic to test empirically, so I am offering you the opportunity to substantially revise your manuscript to address these concerns and resubmit it. I will not repeat the reviewers' and AE's comments here, but do highlight a few points.

First, one key issue is the degree to which you are testing gene-culture co-evolution per se, or aspects of it. Many of these concerns might be addressed with a more nuanced consideration of what your particular study can and can't address.

Second, I agree with the AE that you need to better define what you mean by "norm" and the degree to which the norm in your study is indeed one. Definitions of norm vary, so more precision would be very helpful in answering many of the concerns raised.

Finally, as all of the reviewers mention in different ways, it is critical to interpreting your results to understand the details of the methods, which are not always clear. I realize

that there are space limitations, but please work to present them more clearly. I agree with the AE that moving the methods to the front of the paper would likely help. Of course, please be certain to carefully address each of the reviewers' comments, and please note that this is not a provisional acceptance.

Dr Sarah Brosnan
Editor, Proceedings B

Associate Editor
Board Member: 1
Comments to Author:
This paper aims to test adults' decisions to follow a social norm by copying group-level choices to cooperate, versus copying the most successful strategy. The paper aims to test ideas from gene-culture coevolution by directly pitting the cooperative norm against successful behavior to see if cooperative norms versus self-interest wins out when evidence is provided that such a strategy pays off. The reviewers and myself agree that this is an important question, and I think that providing experimental tests of ideas from gene-culture coevolution is important given the high impact of this idea on the field of human cooperation and social decision making. However, there are several important concerns raised about the theoretical framework and interpretation of the results that should be addressed.

The reviewers have provided thoughtful comments coming from a range of perspectives and here I emphasize additional big-picture points that I think should be addressed and also highlight some additional concerns.

One concern I have is the extent to which the 'norm' presented in this study can actually be considered a norm. I understand the logic for presenting information about the "social norm" using data from another group (e.g., in another group where punishment was possible, such that people did cooperate at higher rates). However, I think addressing whether this manipulation actually represents good example of a social norm would be important. For example, while the paper lays out the logic for not using the actual behavior of the group as the model, the flip side is that a normative rule within one's own group likely has greater (psychological) weight that a putative norm from another group.

RESPONSE: we provided participants with information on the 'typical' behaviour from a sample of 20 participants from the same participant pool (University of Lausanne participant pool). This provides a descriptive social norm (as recently defined by Hertz, 2021, in this very journal, (Hertz 2021)), and is a more reliable sample than just 3 or 4 members of one's own randomly formed transient group in an experiment. Nevertheless, in response to Reviewer 3's requests, we have changed our terminology from The Social Norm to the The Common Pattern and our introduction links to theoretical predictions that people prefer to not deviate from the common pattern.

In addition, based on the supplement the data from the past group was not presented as a "rule" but rather as some info about how other people acted on average in a game. Does some group average data actually have the weight of a social norm to participants? While

lots of people doing the same thing may be one sign that a given behavioral pattern is normative, it does not seem sufficient to establish the behavioral pattern as a norm. Thus it would be helpful to see clearer evidence that participants in this game actually considered the group average data to represent something 'normative' in the sense of something that 'ought' to be done, which is the way it is used by many social psychologists as well as proponents of gene-culture coevolution.

This may require additional data collection looking at (for example) how participants actually understood and interpreted the group average data, or by directly manipulating the presentation of this data to make it appear more or less normative in different conditions. R3 also highlights a large body of work looking at the psychology of norms that may be relevant here.

RESPONSE: we will not be collecting additional data for this study. Our design enables us to test if individuals are motivated to follow either the common pattern or a pattern of success in social dilemmas. How people psychologically evaluate whether common behaviours are just common or ought to be done is interesting but not relevant for the comparisons we are making with prior literature, and beyond the scope of this studies' aims. We have added the following to our Discussion, "Our results suggest that those previous studies**, which used similar participant pools (students at Swiss universities),** but did not include examples of success, may have over-estimated rates of norm conformity and altruistic cooperation. **Although we caution that it is possible that psychology may differ when observing in-group and out-group members (De Cremer & Van Vugt, 1999).**

Second, it seems important to assess the ways in which the full scope of the data here does or does not support gene-culture coevolution, as one benefit of this set of experiments is that it reveals the richness of people's responses across several contexts and thus provides nuanced set of data to evaluate this theory.

RESPONSE: Our data contradict claims that we have evolved through gene-culture co-evolution to follow the common pattern in social dilemmas. The results are not particularly nuanced. When shown both the common pattern and examples of success, behaviour imitates success, and this is not because people are conditioning their cooperation on free riders. When allowed to choose which information to see, the success pattern is the most popular choice. These findings occur despite our participants having received the standard instructions from experiments previously used to conclude that human cooperation is special, requiring special evolutionary forces, and that people desire to match the common behaviour.

First, I note the points by R1 noting contexts where copying of norms does appear alignment with that view, and in a similar vein a discussion of the short-term nature of the experiment versus the long-term view of norm acquisition in real life as raised by R3.

Second, the points by R2 about the importance of exact copying (and the lack of exact copying here) seems relevant for considering what kind of evidence actually supports this theory or not.

I will say that while I do think that ideas from gene-culture coevolution can be fruitfully studied in the short term and in lab experiments, the paper would be strengthened by making clear the benefits versus limitations of a short-term lab experiment compared to the scope of the theories being tested.

- RESPONSE: a single empirical paper is not really the place to discuss the pros and cons of laboratory experiments. Gene-culture evolutionary papers have relied on many laboratory studies, the evidence for the central phenomenon of special human cooperation is nearly all derived from laboratory studies. Nevertheless, we have made our introduction clearer about what is predicted,

"If people desire to copy common behaviours this can help fuel a self-reinforcing gene-culture coevolutionary "*ratchet for both the importance of social norms and the intensity of prosociality*". For example, it is argued that humans have evolved a psychological preference for "***avoiding behaviors that deviate from the common pattern***" and that such preferences are "*probably either the products of purely cultural evolution (driven by cultural group selection), or coevolved products of genes responding to the novel social environments created by cultural group selection*."

- This makes clear that gene-culture co-evolutionary theory predicts a desire to match the common pattern in our experiment.

And we now discuss how short-term laboratory tests related to long term norm formation. Specifically, in our new Discussion,

"*One may argue that our participants were not interested in social information because they had already internalized, over many years, the local social norm for social dilemmas corresponding to the anonymous public goods game. However, this would not explain why our participants chose to learn about success, and why they contributed less when they saw examples of success*."

As much as we would like to discuss these issues in more detail, we do not think there is room for more elaborate discussions in a non-review article.

Finally, I agree that rearranging the order of the paper (e.g., methods before results) would make the paper more comprehensible to readers. I would further note that more concrete details about the specific way the game was played in the main text, which are currently described in abstract terms in the main text with the full script only in the supplement, does seem relevant here.

RESPONSE: we have moved the methods and incorporated copy of the key instructions in the main text, although this does stretch the article length.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)

This is an interesting manuscript that tests one of the major assumptions of gene-culture coevolutionary models of cooperation. Those models claim that people will copy common behaviours like cooperation, even when such behaviours result in a net negative payoff, but then the behaviours persist because of differential success of groups with high cooperation. A key assumption is that people will actually copy cooperation even when that cooperation is costly. This manuscript tests that key assumption, and shows that people have a much

greater tendency to copy successful behaviours than common behaviours. The authors also show that this copying of successful behaviours will undermine cooperation, given that group cooperation is individually costly. As such, this manuscript shows that gene-culture coevolutionary models of cooperation rest on an incorrect assumption: that cooperation is readily copied once it's common, despite the cost. This is strong manuscript that makes a useful contribution to the literature on cooperation.

- RESPONSE: We thank you for your valuable time as a reviewer! Thank you for your kind comments!

I have some comments to improve the manuscript, but these should be interpreted as minor revisions. None of them affect the publishability of the manuscript.

First, sometimes the writing is not clear about whether participants had access to the social norm in all conditions. It's written clearly in some places, but in other places it's easy for readers to forgot this, and incorrectly assume that there is a one condition with zero information (i.e., not even the social norm) or that participants have access to success info but not the overall average. The authors should read through again with this in mind, to make sure it's clear that participants always have access to the norm, and in some conditions they have additional information as well.

- RESPONSE: We apologize for our previous lack of clarity. We have remedied the resubmission to hopefully make this point clear at all times.

Second, Figure 4 shows that people do clearly copy social norms, at least when they don't have info on successful behaviors. As such, the strategy seems to be copy success if you know who's successful, otherwise copy the norm. A proponent of cultural group selection might respond that this is evidence in favour of their theory, given that people often don't know which strategies are successful.

- RESPONSE: This is a fair point, which we already covered in the first submission, and still do in our results, "*When we showed individuals just the ~~social norm~~ Common pattern by itself, they matched it 9% of the time, significantly more than expected by chance, and thus consistent with ~~the Norm-Psychology hypothesis~~* a desire to not deviate from the Common pattern").

If the authors wish to dispute gene-culture co-evolutionary theories of cooperation, then they should discuss this possibility: that people copy successful others when it's clear who is successful (or which strategies are successful), and copy common behaviours when it's not clear who or what is successful. When success is clear or the costs of cooperation are high, then the authors' conclusions hold. When success is opaque and the costs of cooperation are low, then there is still room for copying common behaviours.

- RESPONSE: We agree that uncertain individuals may copy common behaviours, that is why our original and revised introduction writes, "*However, many experiments have only provided information on which behaviours are common, and not the relative success of different behaviours. Consequently, individuals in these experiments could not copy successful behaviours, and may have only copied the common behaviour because they were unsure how to maximize income*."

- We had already discussed the importance of whether people could see examples of success or not in our concluding sentence, "And policy makers should not overly-rely on

people's desire to follow social norms, unless they can prevent people seeing examples of success." Nevertheless, we have now added the following sentence (in bold) to our discussion when summarizing our results. "**However, it may be that individuals do rely on copying common behaviours when examples of success are rare or opaque, which could favour the gene-cultural co-evolution of cooperation in some circumstances, such as when costs are low."**

Third, the authors should include effect sizes on their comparisons.

- RESPONSE: We have now included regression coefficients and effect sizes such as Cohen's *d* in our reporting of comparisons.

Very Minor Comments:

Figure 2 captions says "empty grey circles", but they're square/rectangles in my pdf – relabel

- RESPONSE: Thank you for pointing this out. Unfortunately the pdf you originally received from PRSB was wrong, the legend is correct at the time of submission, but the figure became distorted. We have now tried uploading the ms as a pdf to remedy this problem.

Figure 3 shows the full data distribution, which is great. Many readers aren't yet used to comparing full distributions, and are more familiar with error bars – is there a way to add error bars in addition to what's already there? For example, "pirate plots" include this, and are easy to implement in R. This might not be possible if you also have the dashed lines with mean social info shown; if not, it's not necessary, because this is just a suggestion. But if it is possible, it would help.

- RESPONSE: We were not aware of Pirate plots, thanks! We had a look but still prefer the ggplot framework. As for error bars, it's a bit tricky because we used the Kruskal-Wallis test to compare medians which is non-parametric and does not have confidence intervals. However we have added the individual data points, and their mean with a bootstrapped confidence interval. Please note the means plus confidence intervals are different to the statistical tests we used, which compared medians, so we have retained the median on top.

Figure 2 & 3 captions describe a comparison of Shown+Success versus Shown+Free-Riders. However, the wording is unclear about whether this a comparison of Shown Social Norm versus Show+Free-Riders (i.e., instead of Shown+Success vs. Shown+Free-Riders). This could be written more clearly so that readers know for sure which comparison is going into those stats

- RESPONSE: We apologize for being unclear. We have remedied the figure legend to make it clear which treatments are being compared when.

Referee: 2

Comments to the Author(s)

Thank you for the opportunity to review this paper, and please accept my apologies for the delay in providing my report.

- RESPONSE: Thank you for your valuable time as a reviewer!

This paper examines human social learning in a cooperative setting. The authors test assumptions of popular models of cultural evolution of cooperation which have not received due empirical scrutiny. They present a set of economic experiments involving social dilemmas to examine whether people's cooperativeness is based on following a 'social norm' (common behaviour in a group facing the same decision) or 'success' (the behaviour of individuals achieving the highest payoffs). On aggregate, experimental participants were more interested in observing success than social norms, and successful behaviours were also more likely to be copied. It is argued that the results provide evidence for 'success-psychology' and against 'norm-psychology'.

This paper poses important questions, and I am sympathetic to the authors' aims to critically test the assumptions about learning strategies underlying models of gene-culture coevolution. Overall, the paper is easy to follow and the topic will be of interest to a wide readership.

- RESPONSE: Thank you, we agree!

However, I think the experimental design and associated analyses limits the authors' ability to draw strong conclusions. I list my concerns below.

MAJOR POINTS

1. The presented experiments were designed to cleanly distinguish whether individuals prefer to follow the social norm or be successful (cf. lines 89-90). To this end, two treatments are presented, one showing the social norm, and another treatment also showing the social norm, but also the behaviour of the highest earners. Why did the authors choose this cumulative design? For a 'clean' experimental examination it seems necessary to have a treatment in which only the behaviour of the highest earners is shown.

- RESPONSE: To measure a preference we only need one treatment, showing both forms of information (as we did, Shown Common+Success treatment) or providing a choice (as we also did). That alone would be very informative, and these two treatments show that behaviour imitates success over time and that people choose to see success.

However our cumulative design also provides two controls: 1) the Shown Common pattern treatment (previously Shown Social norm treatment) which provides a baseline control measure for what happens when no success is shown, and allows us to measure the size of the reduction in cooperation that occurs when individuals can learn from success; and 2) the Shown Common+Free Riders treatment which allows us to measure how much of the reduction is due to non-success related factors. With an infinite budget we could have included more treatments but we think our design allows us to adequately address the research question.

We now make this clearer at various places in our new Methods, *"…treatments allowed us to infer from changes in behaviour if individuals were learning from success*", and, "*If individuals have a preference for learning from success, then contributions will decline compared to the baseline treatment (Shown Common pattern), which contained no*

information on relative success.", and, "*Our cumulative information design allows us to directly test if individuals prefer to copy successful or common behaviours in the Shown Common+Success treatment. The other two dynamic treatments can be thought of as controls, which allow us to know the baseline behaviour in the Shown Common pattern treatment, and to control for all the changes due to factors other than learning from success in the Shown Common+Free Riders treatment*.", and in our new Results,

"We can use these relative differences to conservatively estimate that approximately one-third of the reduction when shown the Success pattern (-3.2 MU) was due to effects such as responding to a competing coordination point or to the knowledge of relatively low contributors in the population ('Free Riders') (-1.1 MU / -3.2 MU = 0.34). Meaning that the remaining two-thirds approximately was due to learning about success (2.1/3.2 = 0.66)."

2. The abstract suggests that copying successful others is the driving force behind the demise of cooperation. However, the role of 'copying' is rather unclear. In a public goods game (PGG) with an action set of [0,20], exactly copying the mean contributions of others is a special case. Ample research with this paradigm shows that many people's response patterns reflect 'imperfect conditional cooperation' (e.g., Fischbacher & Gaechter 2010 AER), meaning that they do not exactly copy the mean contributions of others, but systematically contribute bit less than that, presumably explaining the decline of cooperation when interactions are repeated.

- RESPONSE: - We now clarify how we test for copying, "*Copying can be defined in various ways. To provide a strict and objective test to facilitate treatment comparisons we recorded how often participants' contributions exactly matched the previous social information we had shown them. Of course, social learning and cultural transmission can also occur in less exact ways, especially in a repeated scenario with non-constant examples*."

The current paper summarises the literature on this point by saying that in PGGs, people tend to match others' behaviour 'approximately', suggesting that exact copying is actually not to be expected to occur at high frequencies. And indeed, the results show that exact matching is quite rare. Because of the low copying frequency, it seems implausible that copying – in this narrow sense as it is analysed in this paper (lines 217-237) - is the driving force behind the demise of cooperation, as suggested in the abstract.

- RESPONSE: We have changed our title to, "*A preference to follow successful rather than common behaviours in human social dilemmas.*" and our abstract to the more general sense, "…we find that individuals are primarily motivated to ~~copy~~ *learn* ***from*** successful rather than common behaviours. Consequently, social learning disfavours costly cooperation…".

We control for imperfect conditional cooperation by only showing external social information, which is relatively stable and cannot be affected by responses in this experiment. This design prevents a potential process of undercutting ('imperfect conditional cooperation') from driving down the group average. For example, imagine all players do as you propose, and see X (say 10 MU) and then contribute X-y (eg 7 MU). Then they will see 10 again, and not 7, so they will just stay consistently below the social information they see, and no decline will happen. We already explained this in our introduction, and do now in the Methods ("*This meant that if contributions declined in our experiment, this could not be attributed to individuals attempting to match, even imperfectly, the common pattern*"), and

our introduction still says, "*This prevented any within group interactions and importantly, prevented social learning from changing the social information itself.*"

- Our results clearly how that learning from success (copying in the general sense) is the driving force behind the demise of cooperation in our study, and not imperfect conditional cooperation.

3. The operationalisation of 'social norm' in the experiment is problematic. The literature on social norms is huge and it seems that there isn't much agreement on basic definitions. However, something that most scholars do seem to agree on is that social norms are rules that apply to certain specific situations ("in situation X, do Y"). In the current study, however, the shown social norm is derived from a strategically different situation than the participants in are facing. In the main experiment, the situation is a public goods game without punishment, and in the previous experiment, the situation is a public goods game with punishment. This difference leads to three problems in the manuscript: (i) it is unclear whether or not it is actually a relevant social norm that was shown to the participants.

- RESPONSE: By showing the common behaviour we are reliably showing the descriptive social norm (what people do), but yes, it is not an injunctive social norm (what ought to be done). Our methods now make this clearer, "*In all three dynamic treatments we allowed individuals to infer the most common behaviour (ergo the descriptive social norm (Hertz 2021)) by showing them the overall average contribution of all 20 individuals to the public good (Common pattern, Figure 1).*"

As our new introduction explains, Henrich, 2004, argues that, "humans have evolved psychological preferences for "*avoiding behaviors that deviate from the common pattern*", therefore it is not clear that it needs to be an injunctive social norm. We test if individuals prefer to copy the common pattern or success.

The average cooperation level in the punishment experiment is vey comparable to the normal level in the opening rounds of games without punishment, so even an expert would not be able to know if punishment was a feature or not just by looking at the initial level of cooperation.

The players did not know the other experiment had punishment, so it was relevant sample of behaviour for the same decision mechanism (to contribute or not)

(ii) The possibility of punishment changes whether or not cooperation would be a 'successful' behaviour, which is one of the central aspects of this study.

- RESPONSE: Yes, in a sense, but this does not matter here, we just needed to show a stable example. We have explained this in our new Methods section, "Previous experiment"

It is not relevant that punishment affects subsequent payoffs, both of those who receive punishment and those who choose to pay to punish others, this is not part of the cognition here and we are interested in behavioural mechanisms, not the dynamics of full gene-culture coevolution models. The definition of success that we used was correct for the situation the current participants faced. Changing the calculation of success would not change the fact that our participant's showed more desire for information on success.

(iii) In multiple places the authors indicate that no deception has been used in this study, but I think this setup is in a grey area.

- RESPONSE: We disagree, the information we gave participants was relevant to how to be successful in the decision phase of a public good game, which was the set-up they faced. Our information could not harm them. It would have been a problem if the current participants faced punishment and we told them how to be successful in an experiment with punishment but without including the consequences of punishment.

4. The analysis of the PGG data is unclear. The GLMM reported on lines 141-143 is a binomial model fitted to values varying between 0 and 20. Was the dependent variable (contribution) first normalised to fall in the range [0, 1]? This requires explanation.

- RESPONSE: In this case the data are modelled as a series of trials (20), that can be successful or no (1 or 0), common practice for generalized linear binomial models with more than one trial (e.g. sex allocation research).

In addition, it is not clear what the reported test refers to here. Is it the interaction effect only? What estimates did the main effects have? I think the model results should be reported in full (e.g., in the supplement) so the reader can make up their own mind about your data and this analysis.

- RESPONSE: Yes that is correct, the interaction between treatment and period of the game, i.e. the different rates of decline (as we wrote, Contribution ~ Treatment*Period). We have now added six supplementary tables (1-6) fully detailing the results of each model.

A similar comment applies to the GLM right after this (lines 145-149). In addition, in this analysis it is unclear why you chose this link function and how breaking down by round accounts for within-participant dependencies. Overall, it seems that your PGG data allows for replacing all piece-wise analyses - including the ones on lines 199-208 - by fitting a single regression model to all data.

- RESPONSE: Our apologies for being unclear. That is exactly what we did (a single model), but we also did these extra per round models just so that we could answer that specific question (was it higher in each round or not). We also took each individuals aggregate data over all the rounds and compared the medians, therefore we have robustly tested the results in multiple ways, all of which support the same qualitative interpretation. We have now added six supplementary tables (1-6) fully detailing the results of each model.

5. The decline of cooperation over time (Fig. 2) is interesting. This commonly observed pattern is often attributed to limitations to reciprocity (imperfect conditional cooperation and the downward adjustment of beliefs about partners' cooperativeness; see for example the Fischbacher and Gaechter paper cited above). However, the current results suggest that this effect also emerges in the absence of feedback from within the group. Some interpretation of the observed effect would enhance the paper.

- RESPONSE: Yes it is interesting, and a central feature of the manuscript! As we have already explained, our design means that there is no way the decline here is caused by imperfect conditional cooperators. However, what is clear is that when examples of success are included, the behaviour changes. We have now made this clearer in our results and discussion, thank you, e.g. from our new Discussion, "***Our experimental design used stable social information from a previous experiment. This means the decline in contributions cannot be explained by a process of 'imperfect conditional cooperation' whereby individuals repeatedly undercut the previous group average. The obvious conclusion is***

that participants were interested in learning from success because they were either unsure or confused about the costs and benefits of the situation."

6. The experiments in which participants had to choose which social information to observe lack a clear motivation. The manuscript would benefit from making explicit why it is interesting to compare a forced choice with a costly one.

- RESPONSE: We apologize, the more detailed motivation was cut for space issues. The two approaches are complementary and inspired by both behavioural economics and studies of animal behaviour/welfare. We have now added the following text to the Methods to make the motivation clearer.

"*To directly measure the preferences and desire of individuals for social information we also randomly assigned some groups , in the same sessions as above, to a choice treatment…*

The Forced Choice Test measured the relative preferences for the two forms of social information, but risked obscuring preferences because even indivdiuals who have no desire for social information were still forced to make a choice (analogous to compulsory voting with no option to abstain). To complement this approach, we also used the Costly-Choice treatment, where individuals had to pay 1 monetary unit to make their choice, or they could choose to not pay and not receive any information (to abstain). This allowed us to measure the desire for the two forms of social information more generally."

It is also unclear from the main text whether or not the participants the same participants to the main experiment.

- RESPONSE: The choice treatments were conducted in the same sessions as the non-choice treatments, but different participants (participants were randomly assigned to one treatment only). Our methods now make this clear.

I had a similar feeling about comparing the strategy method before and after the repeated public goods game. What do we learn from this in light of the theories on norm-psychology vs success-psychology?

- RESPONSE: Results from the strategy method are often taken as evidence of a uniquely human altruistic desire for fairness (Conditional Cooperation) that can only be explained by evolutionary theories that include cultural evolution. Conditional cooperators should be highly motivated to learn about the Common pattern. Our use of the strategy method before the repeated game allowed us to test this prediction. We found it to be false. We also found that learning about success reduced conditional cooperation and the preference for fair outcomes. As the strategy method controls for beliefs about what one's groupmates might do, this reduction cannot rationally be attributed to examples of success making individuals more pessimistic about their groupmates. If social learning can affect preferences, this could have various implications for cultural evolution.

- In light of your comments we have cut this section from the manuscript but are open to editorial advice on this.

7. The overall conclusion that most individuals preferred to copy examples of successful behaviour (line 309-310) is somewhat misleading. Indeed, 55% of participants copied that

successful behaviour at least once (out of five rounds where it was possible), but the overall rate of copying this behaviour was only 9%.

- RESPONSE: We have changed it to 'overall, our participants… preferred…'. Also, the overall exact copying of this behaviour was 15%, not 9%, and yes 55% copied it at least once. We do not expect exact copying in multiple rounds, because the examples vary slightly but do show a consistent pattern. The key thing is the relative comparisons on the rates of contribution and/or copying. The median aggregate contributions were very close to the examples of success (Figure 3). The aim is to experimentally compare relative differences. In the Shown Common+Success treatment, 55% of participants matched exactly with an example of success at least once, which is more than the 8% that matched with the social norm at least once, therefore there is a strong overall preference for success over the common pattern. And likewise in the choice tests.

8. I note that the study was not pre-registered. I realise that not much can be done about that anymore, but I wonder what was the reason for not doing this.

- RESPONSE: Pre-registration is not a requirement for publication.

MINOR POINTS

• I believe Proceedings B allows for putting the Methods right after the Introduction. Although I think in terms of structure, the manuscript reads fine as is, doing this might clarify some of the experimental details before seeing the results (e.g., that the highest payoffs from the previous experiment were calculated before any punishment took place).

- RESPONSE: we have moved the methods to directly after the introduction.

• The examples in the introduction seem to be far removed from the principles they intend to illustrate. The connection between genetic evolution and 'tipping a waiter' seems rather indirect.

- RESPONSE: We have added more examples and corresponding references. "For example, individuals may tip a waiter in a restaurant they'll never visit again, **honestly report their taxable income, spend time recycling rubbish, pay more for environmentally friendly products,** donate food to food banks and blood to blood banks, even though they'll never learn who benefited, **and in extreme cases even risk their lives to heroically save strangers [refs].**"

- Our examples aimed to satisfy the demands that 1) most people would agree they actually do happen outside the laboratory, and 2) are not easy to explain with a 'standard' genetical evolution account.

- Many relevant papers on gene-culture coevolution provide similar examples, for example, Chudek & Henrich, 2011 TiCS give the following examples "*voting, giving blood, food sharing, not extorting each other, policing and territorial defense*"; and Henrich and Muthukrishna in Annual Reviews Psychology write, "*People in some populations readily give blood anonymously to strangers, recycle, help the poor, report crime, and volunteer for war." (henrich, 2011; Henrich, 2021)*

• The model results at the bottom of caption 2 duplicate the main text.

- Removed.

• The current study aims to disentangle whether people follow of social norms or successful others (lines 89-102). The first cited reason relates to 'previous studies' but no references are given so it's unclear which work the authors seek to contrast their study with. Some relevant papers here include Frey & Meier (2004), Fowler and Christakis (2010), and Nook et al (2016; reference below.

- References now added thank you.

TYPOS / SMALL STUFF

The references to the treatments are inconsistent throughout the manuscript. The short-hand 'Shown' for 'Shown social norm' are not very descriptive. In the text sometimes the 'Shown' is omitted, which is somewhat confusing.

- we apologize for being inconsistent, we have remedied this.

line 34: whom --> who

- thank you, remedied

line 54: the reference of the parenthesised 'social preference' could be tightened.

- thank you, remedied

line 54: preference --> preferences

- thank you, remedied

line 94-95: in this context I am not sure how to interpret the quotation marks around 'real world'

- RESPONSE: we meant outside the lab, i.e. the real situations that laboratory experiments attempt to model. We have removed the quotation marks and written, "Outside the laboratory, **in real world scenarios that economic games aim to model and understand**, individuals may be able to observe both common and successful behaviours."

line 96: nom--> norm

- thank you, remedied

line 98: allowed --> allowing

- thank you, remedied

lines 101-102: not really sure if 'social interactions' as such is an appropriate generic umbrella term for 'reciprocity, signalling, revenge etc'. Maybe these phenomena manifest in social interactions, but they are not the same thing.

- we do not understand this comment, but have changed to 'within-group interactions' anyway.

lines 105-106: 'costly cooperative game' --> 'game involving (individually) costly cooperation'

- removed

line 111: "This mean that the participants knew, in theory, how to be successful." apart from the typo (mean --> means), not sure what you mean by 'in theory'

- We have changed to, "This means that the participants knew how to be successful if they understood the instructions, as has often been assumed in prior studies"

line 156: add 'MU' to Shown + Success = 2.9 MU (15%)

- thank you, remedied

line 188: why is Success with a capital S?

- this was a mistake, apologies for being inconsistent

line 193: allow --> allows

- thank you, remedied

Referee: 3

Comments to the Author(s)
Review for "Humans prefer to copy success rather than the social norm in a cooperative game"

I read this paper "Humans prefer to copy success rather than the social norm in a cooperative game" with much interest. I feel like I should disclose that, prior to being asked to review this paper, I had followed some discussion on social media around its pre-print. That discourse informs my review and I echo some of the points raised here.

- Thank you for your valuable time as a reviewer! We have summarized our response at the end of this review.

This paper presents an experiment where Swiss college students play an economic game under laboratory conditions and are given information about the behavior and success of players in another laboratory experiment. The paper examines how much players contribute to a standard public goods game under different treatments where they have different types of information about the players' success from the previous experiment. The authors seek to use their results to distinguish between what they call "success psychology" from an existing theoretical paradigm called "norm psychology." However, the experiment fails to do this final part due to misunderstandings of the theory behind and implications of norm psychology (and related theories) in the literature.

Overall, I do not think there is enough correspondence between the experiments and the theory the experiments are said to address for this paper to be published in its current form without at least a major revision. The experimental results could, perhaps be useful if the author narrows the scope of the discussion to estimating the relative contributions of success and conformist learning to play in a public goods game in a specific population.

It could be that there has already been work on this that I am unaware of, but McElreath et al (2008) – who performed a similar study with a different payoff structure - could be a guide for how a paper like this would work.

Misunderstandings in the paper.
The paper has multiple misunderstandings of the norm psychology theory. The most important of these misunderstandings, in rough order of importance, for this paper are that:

(A) The paper tests whether players will use conformist learning to follow norms in the laboratory experiment. However, norm psychology is premised, by its proponents, to operate over years and decades of a person's social development. It is not premised to operate, especially in enculturated adults, over the course of a brief experiment. For this

reason, experimental tests have relied on examining differences in experimental play between societies (where individuals may have internalized different norms) or in the same society in children at different developmental ages (to examine the process by which norms are internalized during development). The premise, therefore is that norms have already formed or are in the process of forming *outside* of the experiment and then are applied in an experimental context. Therefore, failure to find norm formation in a short economic game is not a failure of norm psychology.

See Henrich and Muthukrishna (2021) page 212 and Apicella and Silk (2019, cited by author) page R450 for summaries of cross-cultural and developmental studies. Richerson and Henich (2012) explain this in more depth.

(B) The broader theory of how norm psychology operates with cooperative traits is premised on the interaction between conformity and other institutions, such as punishment, rewards, and reciprocity. Therefore, *even if* cooperative norms were to develop a laboratory experiment, removing punishment, rewards, and reciprocity from consideration, as is done in this experiment, is not a proper test of norm psychology. Henrich and Boyd (2001) discuss how conformist transmission and punishment can work together to stabilize norms in cooperative dilemmas. Chudek and Henrich (2011, cited by the author) also make this point repeatedly.

In this paper the author showed participants data from a previous experiment where cooperative norms are stabilized by punishment. However, the author showed participants data of payoffs that did *not* include the costs of being punished (lines 431-437)! This is not an accurate reflection of how institutions for collective action are supposed to work with success-based learning

According to the theories under question, social norms are maintained specifically because violating the norm is costly.

Therefore success-based learning enforces the norm when norm violators are punished. Conformist learning just helps people learn about norms when the connect to payoffs are unclear. In the experiment the author destroys this key premise of the theory by severing the link between success and norm enforcement via punishment. The experiment would align with theory if individuals were show the actual payoffs from the previous experiment instead of excluding the effects of punishment.

In short, it would be more accurate to say that, according to theory, what the author calls a “success psychology” is a potential component of a larger “norm psychology.”

(C) The author sets up a dichotomy between theories of cooperation based on genetic evolution vs those based on cultural evolution and says that these theories are in conflict. However, the theories under consideration are actually those based on genetic transmission alone and those based on genes *plus* cultural transmission (sometimes called “dual inheritance” or “gene-culture coevolution”). In the genes plus culture theories of cooperation, culture evolution creates institutions that align genetic success and self-interest with cooperative behavior. Successful individuals are those that tend to follow

cultural norms. It is incorrect, therefore, to say that these theories pit norm adoption against genetic or material self-interest. Chudek and Henrich (2011), which the author cites, makes this point repeatedly throughout. As do other articles the author cites, such as Henrich 2003.

This is from Chudek and Henrich (2011):
"The interaction between culture and genes is continuous. The more genes respond by building and honing the above-described norm-psychology, the more they power up the cultural processes that generate and sustain local phenotypic assortment, sanction deviations within groups and select for more cooperative norms. This creates a culture–gene coevolutionary ratchet for both the importance of social norms and the intensity of prosociality. As cultural group selection increasingly guarantees that learners find themselves in social groups organized by norms that incentivize prosocial or cooperative behavior, within-group (and between-group) genetic selection processes will favor genes that build prosocial, norm-adhering phenotypes. This evolutionary trajectory – from cultural learning, to norm-psychology, to cultural group selection for prosocial norms to psychological adaptation to a world dominated by prosocial norms – may help explain some of the puzzling prosocial experimental results that have been dubbed 'strong reciprocity'."

Previous experimental results

This paper does not reference or engage with previous experiments on success vs conformist transmission, with the most important of these, McElreath et al (2008), published in Proc B some years ago.

McElreath et al used an inter-generational experiment and found that participants in their experiments used a mix of success based and conformist learning, but used success-based learning more. A difference is that McElreath et al used a multi-armed bandit problem instead of a public goods game, so this paper has a different set-up in that it uses a cooperative game. The author should discuss why the cooperative game is a better model for the question at hand.

McElreath et al also used a statistical modeling approach that used an explicit model of population dynamics to estimate the mix of conformist and social learning the participants employed. This paper would be much strengthened by using a similar approach. The author should at least engage with previous results.

McElreath et al also reference additional "multi-generational" experiments where participants observe payoffs and strategies of participants in previous games. This is salient because the author of the paper under review implies that having participants observe strategies and payoffs from previous players is an innovation. However, the author should put this design in the context of previous experimental results looking at multi-generational social learning experiments.

- We have added two references on intergenerational advice in games, thank you (Schotter, & Sopher, Social Learning and Coordination Conventions in Intergenerational

Games, 2003; Chaudhuri et al. Social Learning and Norms in a Public Goods Experiment with Inter-Generational Advice, 2006).

More narrowly focus the theoretical scope of this paper. I think this paper could be more accurately written as a simple examination of to what extent Swiss college students tend to use success-based learning and conformist learning in a public goods experiment.

We already have well established theories about how conformist learning should be used less when there is better information about the relative success of different behaviors and this study seem to confirm this theory. Models of conformist transmission, which pre-date and are more general than the norm psychology theories, operate under the premise that the conformist learning occurs when the connection between behaviors and success is unclear (please see Henrich and McElreath (2007), section 38.2.3, page 562 and models referenced therein). Therefore, it is not surprising that adding success information decreases conformist learning.

I think the author's attempts to use this experiment to test the more encompassing theory of norm psychology falls flat because they do not adequately account for the assumptions of those theories and the author misstates the premises of those theories in many cases. However, applying these results more narrowly to conformist vs success-based learning does not require as many assumptions. However the author would need to be clear that the results might not generalize outside of the study population or outside of the experimental context.

I also think this would be a better paper if the author determined the amount of conformist vs success-based learning the participants engaged in. If participants are engaging in a mix of both, it would be better if the author explicitly modeled this possibility. Perhaps the methods in McElreath et al (2008) would be of use.

Specific references:

The author cites references 11-17 as describing the theory of cooperation that they are testing. I think three of these are as close to canonical as these things get: Henrch (2004), Chudek and Henrich (2011), and Richerson et al (2016) and I am familiar with two others (Apicella and Silk 2019 and Handley and Mathew 2020). I would add a few more as part of the cannon.
Henrich and Boyd (2001) on conformist transmission and cooperation.
Richerson and Henirch (2012) on cultural evolution and collective action problems.
Henrich and Muthukrishna (2021) on cultural evolution and human cooperation.

- We added these references thank you.

Minor comments:

Title: The title "Humans prefer…" is overly broad. This study was not conducted with a random sample of humans. It was conducted with a sample of "mostly students enrolled at

either UNIL or Swiss Federal Polytechnic School." It is, by now, well established that individuals from Western Educated Industrialized Rich and Democratic groups play and interpret experimental games differently than people from other societies. See papers and books on WEIRD societies and economic games by Henrich and collaborators. The author should use a title that is more reflective of the population with which the research was conducted.

Line 51: Instead of "cannot be explained by genetical evolution" it would be more accurate to say: "cannot be explained by genetical evolution alone" or better "cannot be explained by genetical evolution without another transmission mechanism." The author should update this to better reflect the theory.

Line 58: "evolved culturally, through behavioral, rather than genetical, copying of traits" I don't understand what "behavioral" means in this context. Both cultural and genetic traits can influence behavior and can therefore be considered "behavioral."

A better description of the theory would be something like "have evolved cultural transmission systems that co-evolved with genetic selection to stabilize norms and that that between group processes tend to select for the more cooperative norms." The author should update this to better reflect the theory.

Line 59: "against our material or genetic interests." See misunderstanding (C) above.. Norm psychology is premised on the assumption that people evolve to adopt norms because it *is* in their material and genetic interests to do so. This is clear from the theory as described in the papers the author cites and elsewhere. The author should update this to better reflect the theory. One place this is stated concisely is in Boyd (2017, pg 187): "Cultural group selection models assume that behavior within groups is motivated by individual self-interest. Norms are maintained by rewards and punishments that make it beneficial to follow the norms. If individuals did not benefit from conforming to norms, then the cultural group selection hypothesis would be falsified."

Lines 104-105: "We experimentally tested if individuals prefer to copy either the social (norm psychology hypothesis) or successful behaviors (success-psychology hypothesis) as described above, norm psychology is premised to occur over years and decades of development and align norm compliance with payoffs. Therefore, there will typically be no conflict between success-based and norm-based learning. The paper sets up a strange dichotomy here. It would be more accurate to say that this "success-psychology" is part of a larger "norm psychology," but I find this dichotomy an ill fit for the theory.

Lines 117-118 and 187: At first I found these two descriptions of the experimental set-up confusing: "We also ensured individuals observed a stable social norm by allowing individuals in the previous experiment to punish each other, which stabilized cooperation." and "In the standard public goods game, the highest earning individual are those that contribute the least. Therefore, when we showed individuals examples of successful behaviors, we were also showing them the behavior of those individuals who contributed the least." If punishment stabilized cooperation in the original experiment, presumably free-riders are punished enough that their payoffs were lower than cooperators. However, it is

not until line 431 that we find out that the experimenter did not show the actual payoffs to the participants. Instead, the experimenter removed the costs of being punished from the player payoffs. As described above, this violates one of the premises of the theory under question, but in any case the author should explain this earlier to avoid confusion earlier in the paper.

Line 315. "Even if humans…" Again, this is an overly broad statement from research conducted on one WEIRD sample of humans. The author should re-write this sentence in a way that does not imply that the experimental results apply to humans generally.

References:

Apicella, C. L., & Silk, J. B. (2019). The evolution of human cooperation. Current Biology, 29(11), R447-R450.
Boyd, R. (2017). A different kind of animal: how culture transformed our species (Vol. 46). Princeton University Press.
Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. Trends in cognitive sciences, 15(5), 218-226.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of theoretical biology, 208(1), 79-89.
Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. Annual Review of Psychology, 72, 207-240.
Henrich, J., & Richerson, P., (2012). Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems. Cliodynamics, 3(1) 38-80.
McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1509), 3515-3528.

RESPONSE: We summarize our response to Reviewer 3 here rather than providing repetitive responses above. We summarize their review into 7 points:

Point 1) The reviewer states that, "*The experimental results could, perhaps be useful if the author narrows the scope of the discussion to estimating the relative contributions of success and conformist learning to play in a public goods game in a specific population*."

Summary response: To this end we have changed our focus to copying/learning about common behaviours rather than the social norm per se.

We still refer to our idea that this sample of common behaviour in 20 fellow Swiss-based students is indicative of the descriptive social norm, so that the reader can make up their own mind, but we accept that our operationalization of the social norm was not as strong as it could have been.

Some examples of the changes we have made, we have changed the title to, *"A preference to follow successful rather than common behaviours in human social dilemmas."*

Our introduction now reviews a proposed desire for "*avoiding behaviors that deviate from the common pattern*" [Henrich, 2004], and the accompanying evidence, "Experiments using social dilemmas, such as the public-goods game, have provided evidence consistent with a human desire to conform with common social behaviours and thus '*not deviate from the common pattern'*

We have changed the name of our treatment from Shown Social Norm to Shown Common Pattern, and our methods now read, "In all three dynamic treatments we allowed individuals to infer the most common behaviour (ergo the descriptive social norm [Hertz, 2021]) by showing them the overall average contribution of all 20 individuals to the public good (Common pattern, Figure 1))."

Point 2) The reviewer requests that we take more care to caution against over generalizations from our study sample.

Summary response: we have made several changes to the text cautioning against over generalizations:

Specifically, the reviewer says that, "The title should not say Humans"

We understand their point of view, but in biology journals it is customary to have species names in the title. Most scientists understand that studies are based on samples of the study species. Our new title now refers to "a preference" in "*human social dilemmas*…" indicating it's not universal. We are happy to take the Editor's direction on this.

We have also updated our discussion in many places, to clarify that our results may not generalize but that we used a similar participant pool used by previous studies, for example (**new in bold**),

"Our results suggest that those previous studies, which **used similar participant pools (students at Swiss universities), but** did not include examples of success, may have…", and, "**One could use a range of different instructions to test if our results generalize to different set-ups or different cultures. Here we simply replicated 'standard' instructions and used a common participant pool, to enable comparisons with many prior key studies**."

We have also modified another statement the reviewer objected to, "*Even if humans do have ~~a norm psychology~~ desire to follow the 'common pattern', it **seems this desire** was less motivating, **in our participants,** than their desire to follow '~~success psychology'~~"*

Point 3) The reviewer, repeatedly, discusses our omission of one particular study, e.g. "*This paper does not reference or engage with previous experiments on success vs conformist transmission, with the most important of these, McElreath et al (2008), published in Proc B some years ago*."); "McElreath et al used a multi-armed bandit problem instead of a public goods game, so this paper has a different set-up in that it uses a cooperative game. The author should discuss why the cooperative game is a better model for the question at hand."

Summary response: a cooperative game is necessary to study cooperation.

Our ultimate question is how human cooperation evolves, therefore a cooperative game is a better model for the question at hand, we do not think this needs to be justified. We are disputing the assumption that because people often conform in non-social contexts, this can explain cooperation. McElreath et al 2008 studies how people use social learning to decide whether to plot wheat or potato crops, we do not think this is relevant here. One

could provide literally thousands of experiments supporting the idea of conformist transmission but if they are in non-cooperative contexts the results are irrelevant here.

We have added the following text to our methods, "*Other studies have investigated social learning in non-cooperative games (McElreath, 2008). However, because we are interested in the evolution of cooperation, we thought it more relevant to study a cooperative game."*

The reviewer also suggests we adopt the statistical modelling approaches of McElreath et al. 2008 to measure if individuals use a mixture of conformity and payoff based learning. This is not a sensible idea, because in our experiment they could only 'copy' (or choose) at most one form of information. Nevertheless, we have now examined how many individuals matched with both the common behaviour and the successful behaviour at least once across the five rounds (obviously they could not do this at the same time because the values were different). We found that in the Shown Common+Success treatment, only 3 participants matched both at least once, compared to 67 that only matched success, and 7 that only matched the social norm. We have now updated our results.

Point 4) The reviewer does not appreciate the significance of our results. "*Models of conformist transmission, which pre-date and are more general than the norm psychology theories, operate under the premise that the conformist learning occurs when the connection between behaviors and success is unclear. Therefore, it is not surprising that adding success information decreases conformist learning*."

Summary response: Yes, our results are surprising!

Our experiment is not one where participants have to figure out the payoffs of a unknown situation (e.g. wheat versus potato crops). Instead, our participants had the same, publicly available, standard instructions that have been used in the experiments that concluded human cooperation is unique, motivated by altruistic pro-social norms and cannot be explained by standard genetical evolution alone. These previous studies assumed individuals understood the game and already knew how to be successful. Therefore, showing examples of success is redundant if this is true. However, we find individuals decrease their cooperation when they can learn from success, despite having had full access to the instructions and 78% of them behaving like Conditional Cooperators in the preceding strategy method.

If the reviewer wishes to insist that the results are not surprising, then they must also accept that the previously documented cooperation is not so special after all, in which case we do not need cultural evolution models to explain behaviours such as 'strong reciprocity' or the conditional cooperation norm (Fehr and Schurtenberger 2018).

We have added the following to our discussion to make the significance of our results more clear, "***The obvious conclusion is that participants were interested in learning from success because they were either unsure or confused about the costs and benefits of the situation. While the use of success-based learning in contexts where the costs are unclear is consistent with many theories, including some gene-culture co-evolutionary theories, this conclusion would invalidate the claim that public goods games show evidence of uniquely human cooperation demanding unique evolutionary explanations***."

Point 5) The reviewer lists as their primary concern that we test "*whether players will use conformist learning to follow norms in the laboratory experiment. However, norm psychology is premised, by its proponents, to operate over years and decades of a person's*

social development…**The premise, therefore is that norms have already formed or are in the process of forming *outside* of the experiment and then are applied in an experimental context."**

Summary Response: We address these issues in our new Discussion.

The reviewer does not provide any specific critique in this point or engage with our results, so it is unclear what their actual complaint is. It may well be true that social norms take years to develop but we are not testing for norm formation. We are testing for a preference to learn about and/or copy either common or successful social behaviours. Even if the reviewer is right, and people do not pay any attention to the social information in the short-term laboratory experiment, this critique does not explain why they cared about the examples of success (or why individuals match the group average in previous experiments with no information on success). These internalized norms do not seem very resilient.

Arguing people bring outside behaviours into the lab just means we no longer know what game they think they are playing. There is no guarantee the context free, technical, mathematical instructions will activate the corresponding internalized norm (or biological adaptation) which is argued to have evolved to save the costs of computation (as Henrich & Muthukrishna 2021 write,"*As part of this norm psychology, evidence suggests that humans have evolved to (at least partially) internalize norms as context-specific motivations or* ***frugal heuristics for navigating daily life****. This internalization may have evolved for several reasons, including* ***to minimize cognitive effort and/or to mitigate the decision-making challenges****…" (Henrich and Muthukrishna 2021))*. For instance, in Henrich et al. 2005, BBS, they report that participants often though the experiment was like local situations, such as 'the Harambee game' (Henrich, Boyd et al. 2005). However, the payoffs of such local external situations, with reputations and repeated interactions, will rarely map perfectly onto the payoffs of the experimental game. Therefore it is not clear what cross-cultural studies are sampling, there is no guarantee that it is different social norms for the same dilemma.

We address these issues in our new Discussion, "*One may argue that our participants were not interested in social information because they had already internalized, over many years, the local social norm for social dilemmas corresponding to the anonymous public goods game. However, this would not explain why our participants chose to learn about success, and why they contributed less when they saw examples of success*."

Point 6) The reviewer Is confused about the aims of our experiment. "In this paper the author showed participants data from a previous experiment where cooperative norms are stabilized by punishment. However, the author showed participants data of payoffs that did *not* include the costs of being punished (lines 431-437)! This is not an accurate reflection of how institutions for collective action are supposed to work with success-based learning"

Summary response: This is not a valid critique. The issue is not relevant here and does not explain why participants had a preference for learning from examples of success.

We agree institutions are potentially important in the dynamics of cultural evolution but we were testing the relative preference for learning from either common or successful behaviours. Changing the prior calculation of success, to incorporate the costs of both those who receive punishment and those who pay to punish others, could be interesting to see how institutions interact with success learning, but would not affect our participant's desire to learn from success.

Our participants were concerned with one decision mechanism, and we gave them social information about that. The fact that punishment could subsequently alter payoffs in a different environment with punishers and punishees is irrelevant here.

We have now made our methods clearer, " *Although punishment is often part of gene-culture coevolutionary theories, it is important to realize that our use of punishment is not a key feature of this experiment. We were investigating a behavioural preference for learning from either common or successful behaviours, therefore the definition of success is not crucial. We used a defintion that was correct for the siituation our current participants faced. Changing the defition of success may have different outcomes for gene-culture coevolutionary processes but we were not aiming to simulate cultural evolutionary processes. An experiment invesitgating how individuals learn from payoffs when the costs and benefits of punishment are incorporated would require different assumptions and would address different questions, such as how institutions interact with success-based learning*."

Point 7) The reviewer says we are wrong to pit norm adoption against genetic or material self-interest (what they refer to as misunderstanding C). The reviewer quotes Chudek & Henrich 2011;

"The interaction between culture and genes is continuous. The more genes respond by building and honing the above-described norm-psychology, the more they power up the cultural processes that generate and sustain local phenotypic assortment, sanction deviations within groups and select for more cooperative norms. This creates a culture–gene coevolutionary ratchet for both the importance of social norms and the intensity of prosociality. As cultural group selection increasingly guarantees that learners find themselves in social groups organized by norms that incentivize prosocial or cooperative behavior, within-group (and between-group) genetic selection processes will favor genes that build prosocial, norm-adhering phenotypes. This evolutionary trajectory – from cultural learning, to norm-psychology, to cultural group selection for prosocial norms to psychological adaptation to a world dominated by prosocial norms – **may help explain some of the puzzling prosocial experimental results that have been dubbed 'strong reciprocity'."**

Summary response: We have made several changes to the text.

The confusion perhaps arises from proponents of gene-culture co-evolution theories frequently claiming to explain human altruism and the patterns of human cooperation observed in the economic games e.g. the 'strong reciprocity' highlighted at the end of the above quote, for example, as in these two quotes from Henrich, 2004, "*Unfortunately, the existing genetic evolutionary approaches can explain neither the degree [nor the pattern] of prosociality (altruism and altruistic punishment) observed in humans*.", and, "*Experimental findings from many small- and large-scale societies show that people will trust, cooperate and behave altruistically toward anonymous individuals in simple one-shot games*." (Henrich 2004).

Altruism is normally defined as having no material benefits nor genetic benefits except for relatives (but that explanation is ruled out in the quote's prejudice against existing genetic approaches). Strong reciprocity is defined as cooperating in situations that have no material or genetic benefits from cooperating (anonymous interactions with strangers). For example, Fehr & Henrich 2003 define it as, "*The essential feature of strong reciprocity is a willingness to sacrifice resources in both rewarding fair behavior and punishing unfair behavior, even if this is costly and provides neither present nor future*

economic rewards for the reciprocator" (Fehr and Henrich 2003). Laboratory experiments often show cooperation in situations that have no material or genetic benefits. It is the hallmark trait for those that argue human cooperation cannot be explained by standard genetical evolution. That is why we had said cultural explanations claim to explain human cooperation '*even when it is against material or genetic interest*'.

We have rewritten certain parts to avoid this 'misunderstanding'. Our abstract has simply cut the following, "~~natural selection should favour individuals whom prefer to discover and copy successful behaviours~~".

Our opening sentence is now (**new in bold**), "*Humans often appear to cooperate in ways that cannot **easily** be explained by genetical evolution **alone***"

Our second paragraph has changed the offending sentence to, "*A proposed explanation for such cooperation is that humans have also evolved culturally, through the behavioural, rather than genetical, copying of traits, to cooperate even ~~when it is against our material or genetic interest~~ **in anonymous one-shot encounters with strangers***".

We have also added the following quote to better align our writing with the relevant theory, "For example, it is argued that humans have evolved psychological preferences for "*avoiding behaviors that deviate from the common pattern*" and that such preferences are "*probably either the products of purely cultural evolution (driven by cultural group selection), or coevolved products of genes responding to the novel social environments created by cultural group selection*."(Henrich, 2004).""

We have also changed the following , to be simpler, more neutral and less controversial, *"However~~, cultural and genetical evolution can pull in opposing directions~~ while it appears obvious that individuals sometimes copy other individuals, especially in domains such as fashion and language, it is not so clear that individuals copy costly behaviours. Instead, ~~natural selection may have favoured~~ individuals **may use different cognition in cooperative contexts and prefer** to copy successful social behaviours."*

However, again, nothing in this critique can explain why our participants had a preference for learning from examples of success.

References

Fehr, E. and J. Henrich (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. Genetic and Cultural Evolution of Cooperation (Hammerstein, P., ed.)**:** 55-82, MIT Press.

Fehr, E. and I. Schurtenberger (2018). "Normative foundations of human cooperation." Nature Human Behaviour **2**: 458-468.

Henrich, J. (2004). "Cultural group selection, coevolutionary processes and large-scale cooperation." Journal of Economic Behavior & Organization **53**(1): 3-35.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton and D. Tracer (2005). ""Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies." Behavioral and Brain Sciences **28**(6): 795-815; discussion 815-755.

Henrich, J. and M. Muthukrishna (2021). "The Origins and Psychology of Human Cooperation." Annual Review of Psychology, Vol 72 **72**: 207-240.

Hertz, U. (2021). "Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms." Proceedings of the Royal Society B: Biological Sciences **288**(1952): 20210293.

Appendix B

RESPONSE TO REVIEWERS AFTER 2ND ROUND OF REVIEWS

Manuscript ID RSPB-2021-1590

07-Sep-2021

Dear Dr Burton-Chellew:

Thank you for resubmitting your revised manuscript to Proceedings B. I have now received two reviews as well as the comments of the Associate Editor. As you will see, the reviews are generally quite positive, and based on my own reading of your manuscript as well as the advice of the AE, I think that with another round of revision your manuscript will likely be acceptable for publication. However, you will see that despite their positive assessments, the reviewers have raised some concerns, and I invite you to address them. Both of the reviewers and the AE have written particularly clear and constructive comments, appended below, so I will spare you repeating them here.

We typically do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

Best wishes,

Dr Sarah Brosnan
Editor, Proceedings B

RESPONSE: thank you for the opportunity to resubmit our manuscript, the thoughtful comments from the AE and the reviewers have helped us improve the paper. We detail our changes below and include a copy of tracked changes at the end.

Associate Editor Board Member
Comments to Author:
This is a revision of a paper testing adults' decisions to follow a common strategy by copying group-level choices to cooperate, versus copying the most successful strategy. This is an important question and an interesting way to test ideas from gene-culture coevolution. The reviewers have provided thoughtful comments about the new version of the manuscript that should be addressed, and I provide some additional comments and highlight some specific aspects I view as particularly important.

Some of the issues raised by reviewers are partially addressed in the manuscript but should be better highlighted to make sure these points are clear to readers. For example, my understanding is that the use of a sample where punishment was possible is used as a model here because this is a well-known way to stabilize cooperation in order to even show examples where other people

cooperated, but it would be important to take care that the reasoning for doing so is clear in the paper.

RESPONSE: Your understanding is correct. We have written more in the methods to make our reasoning clearer (**new in bold**) "**We wanted multiple rounds of data so we could observe more learning and short-term cultural evolution. Therefore, we needed multiple rounds of 'model data'. However contributions normally decline in public goods games.** If individuals observe declining levels of cooperation, it is hard to test if individuals are learning from success or trying to match a declining group average. Our solution was to use **one of our own** previous experiments that used peer punishment to stabilize mean contributions (Burton-Chellew and Guérin 2021). **This way we could show this study's participants a relatively stable level of contributions.** This meant that if contributions declined in **this current** experiment, this could not be attributed to individuals attempting to match, even imperfectly, the common pattern. In contrast, declining contributions would be expected if individuals learned from either the Success pattern or the Free-Rider pattern.

Our participants in this study were not aware that punishment was a feature of the previous experiment. **They had no need of this information as it was not relevant to the decision they faced.** Examples of highest earners were taken from **each round of** the public-good decision stage of the experiment, before any potential punishment in the subsequent punishment stage.

It would also be important to clarify use of terminology like 'copying', 'social learning', or 'following others' as noted by R1. This seems especially important in that a lot of the cultural evolutionary work on social learning is focused on imitation actions (often of a more complex behavioral nature than responses in this kind of economic game), so being clear about what the study is testing and its implication is important. The revisions should also take care to accurately characterize claims about (imperfect) conditional cooperation made by other teams.

RESPONSE: We have better characterized claims about (imperfect) conditional cooperation. As about 50% or more of conditional cooperators are 'perfect' we disagree that our characterization was incorrect. However, as detailed below, we have changed the text accordingly from 'match' to 'condition upon' local levels, to be clearer.

We have removed the vague term 'follow' from the manuscript, e.g. we changed the title from 'a preference to follow' to, 'a preference to learn from..' and removed unclear uses of 'follow'. Instead, we changed the headings of the results sub-sections, e.g. from 'Following success…' to "Deviating from the common pattern."

We defined 'social learning' in our introduction, "humans learn by observing others (social learning)".

We defined our use of 'copying'. "*The cultural evolution of behaviour requires that individuals learn from or copy one another (social transmission).. Copying can be defined in various ways. To provide a strict and objective test to facilitate treatment comparisons we recorded how often participants' contributions exactly matched the previous social information we had shown them. Of course, social learning and cultural transmission can also occur in less exact ways, especially in a repeated scenario with non-constant examples*."
Copying is also used frequently in the literature on cultural evolution, without being defined precisely e.g. (Henrich and Muthukrishna 2021).

R3 raises several questions about the potential ambiguity of the information provided by the "common pattern". I think that if participants are told that a particular response is the most common response (rather than given direct access to the full distribution of participant responses) then some of the concerns raised by R3 may not necessarily apply as such. Nonetheless, clarification about what the common pattern is, and how this information was presented to participants, would be important, as well as some discussion of potential limitations of the current design and next steps that could build on this.

RESPONSE: Participants were not shown the distribution. They were told what the information was factually, i.e. the overall average of 20 participants. Showing them the average contribution is easily understood and consistent with the methods of studies of conditional cooperation which suggest individuals condition their cooperation on local levels of cooperation. If individuals are not interested in the average then this contradicts previous literature and adds to our understanding of what motivates human cooperation. We have now added this to the main text in the methods when explaining the treatment. Showing the average over time provides a common pattern over time, but it is true we do not show the distribution (pattern within a round) or median/mode. We have discussed this to our Discussion. "*Alternatively, our participants may have not been interested in the overall average contribution of 20 participants, and instead have preferred to see the median or mode, however this would contradict the results of previous studies which used the local mean.*"

The authors should also clarify the question about potential use of deception in the task; while deception can be important in psychology studies to address various questions, it is nonetheless something to clarify about the procedure.

RESPONSE: The rules against deception are primarily to prevent researchers using virtual/computerized players programmed to play in certain ways in place of real participants. Not telling participants the whole truth is not the same as deception, which deliberately misleads participants. We told our participants the previous experiment contained the same decision task, which it did, but we omitted the fact that there was a second stage of punishment. In hindsight we could have said 'similar decision' instead of 'same decision', or could have simply used the no-information treatment we planned to use, which was not quite as stable as the punishment treatment. However, we do not think this would have changed our results. Furthermore, our arguable use of 'deception' here will not pollute the participant pool, which is the primary concern of forbidden deception in economic experiments, and did not harm our participants welfare in anyway (if it had been the other way round, and we told them how to be successful in a game with punishment but using data from a game without punishment, that would have been economically harmful to them).

We now write, "*Although this omission of details about the punishment stage in previous experiment could be seen technically as deception, it is true that our participants and the participants in the previous experiment had to make the same basic decision, of how much to contribute to a public good... We could have sacrificed some groups in this experiment to act as model groups, perhaps playing with no information to generate almost stable contributions, but this would have been complicated, uncertain, and severely reduced our sample size. Telling our particiapnts about the punishment stage (which is complicated to explain) would have needlessly complicated this study for little or no scientific or ethical benifit. With hindsight it may have been better to write 'similar' and not 'same' decision. However we doubt this would have changed*

our results."

Finally, the revision has adjusted the use of social norm through the paper to the "common" pattern, which makes sense to me. But it still describes this common pattern as representing a social norm throughout, thus undermining the responsiveness of this change. For example, the introduction states that following common patterns means confirming to norms ("desire to follow common social behaviors (conform with social norms)"), later the common pattern is described as a "descriptive social norm" (a term which is not defined in the paper), and the methods and results use the terms social norm interchangeably with common pattern (e.g., "35% of individuals matched the social norm at least once"). Following the comments in the prior version of the manuscript, this work does not demonstrate that the common strategy has the force of a norm the way this is typically used by people working in gene-culture coevolution, and while this is not necessary for the paper to be an interesting contribution it is important to make this clear. A discussion of this issue, including a more nuanced discussion of how this is relevant for theories about gene-culture coevolution (which are actually focused on norms as such) would be relevant.

RESPONSE: we apologise for our failure to sufficiently update the manuscript's new terminology. We have removed the above example from the introduction and any mentions of 'social norm' from the results (that was an oversight, sorry). We now use our discussion to discuss the relevance of our approach "*…we would argue that the overall average of 20 participants from the same participant pool would often be seen as indicative of a local social norm, i.e. it would describe normal behaviour in such a situation (ergo the descriptive social norm (Hertz 2021)), however we did not directly ask our participants how they interpreted the social information. It is possible that if a normative standard of what one ought to do had been applied to our social information then our participants would have been more keen to not deviate from the common pattern."*

Reviewer(s)' Comments to Author:

Referee: 2

Comments to the Author(s).

The authors have taken serious effort in revising this manuscript. The presentation of the research has much improved. Extending the Methods section in the main text has enhanced the paper by motivating the cumulative design and the forced choice tests. I have a few (rather minor) lingering issues I would like to see addressed before I would recommend publication.

RESPONSE: thank you for your valuable time re-reviewing our manuscript and your kind words.

The paper seems to use 'copying', 'social learning', 'following others', and 'matching' interchangeably. In the literature on conditional cooperation, 'copying' (defined as exactly matching behaviour of others) is not as common as the authors make it sound. In particular, the statement that 'most [people] will match their cooperation to local levels' (l. 57-58) and that individuals 'prefer to match the common patterns' (l. 82, 83) is not supported by the cited papers. If anything, the cited papers (most notably Fischbacher et al 2001 but also Fischbacher and Gaechter 2010) show 'imperfect conditional cooperation', in which people undercut rather than

match average contributions. Later on, the authors state that 'matching' might be imperfect (l. 135) but it would be helpful to make the writing precise when referring to these concepts from the outset.

RESPONSE: As we mention above, we have made several changes on this request. We have also changed the wording on lines 57-58 from 'match' to "***condition*** *their cooperation to local levels (either perfectly or imperfectly)*", which is in agreement with the findings from those studies and avoids committing to a distinction between perfect and imperfect matching. While the distinction between perfet and imperfect matching is interesting, we disagree with the argument that the papers have not been interpreted to suggest support for the idea that most people are either free riders or those that will match/conform to local levels. For example, in Fischbacher & Gachter 2010 they write, "Free riders (located at 0–0) and **perfect conditional cooperators** (at 1–10) are relatively the largest group of subjects."[emphasis added]. Imperfect conditional cooperators are still paying a price, by not contributing 0, to follow in some manner, the local level of cooperation.

Fehr & Gaechter (2002) is about punishment and not about conditional cooperation as studied in the current paper, so this reference should be removed from the citations in l. 83.

RESPONSE: We have removed the citation here.

I appreciated that the revision discusses the differences between the previous and the main experiment in terms of punishment. However, I did not understand the argument around this on lines 143-148. In particular, I am not sure what you mean with a "correct" definition of success here. This passage also has a few typos.

RESPONSE: we mean the definition was consistent with the situation are current participants faced, of deciding how much to contribute to a public good with no punishment. We have made the text clearer and fixed the typos.

From reading the experimental instructions cited in the main text I worried that participants might have interpreted success information as stemming from the person who made most money *overall* rather than in that particular round (which would make the success information more about 'prestige'). From the SI I see that in a preceding paragraph, the instructions did make it clear that the success information stemmed from the corresponding round. It would be useful to briefly point out this fact in the main text.

RESPONSE: Apologies for not making this clear, we have now mentioned that it is information from each particular round at several places in the text. The quoted instructions in our main methods also already mention that it is per round, e.g. "*2) The average decision, **per round**, of the individuals with the highest earnings"*

Descriptions of the GLMM analyses have improved and presenting the model results in the SI is very helpful. One minor point: it might be better to avoid the term 'slope' to refer to 'period in the repeated game' because 'slope' is often used to refer to a coefficient.

RESPONSE: We have changed slope to period to remove this confusion.

It would be useful if the authors proof-read the paper once more. The revision has a number of typos. The ones I spotted are listed below:
- l 75: 'they' – we can't find this mistake?

- l 186: ‘individuals’ – fixed thank you
- l. 189: ‘pattern’ – fixed thank you
- l. 219: ‘group’ – fixed thank you
- l. 220: ‘or an’ – fixed thank you

RESPONSE: we have fixed 4 of the 5 mistakes above, one we cannot find. We have hopefully fixed all other typographical errors, sorry for these mistakes.

Referee: 4

Comments to the Author(s).

This is an excellent paper, touching on some important issues at the interface of evolution and human social behavior. Because the paper runs contrary to some popular views, it should be taken all the more seriously as a potentially important contribution. (It is a sad but seemingly inevitable fact of the practice of science--given the psychological and sociological contexts in which science is done--that dissent is pushed to the margins.).

The experiment was well-framed in the introducton, with a solid experimental design and appropriate analysis of data. The conclusions are accurate and not overblown. The paper is also exceptionally well written.

RESPONSE: thank you for your valuable time as a reviewer and for your kind words!

My comments are rather few. All of the comments can be addressed quickly in a very light revision.

1. Lines 131-140: I accept the fact that the data they used to show subjects how previous subjects had contributed in previous games were obtained from games in which high levels of cooperation were sustained by peer punishment. I understand the proximate rationale for doing so (they wanted to avoid giving subjects the expectation that declines in cooperation were in some way normative), but their ultimate rationale was a bit unclear. Did they go to all of this trouble in the first place? Couldn't they have devised other ways to show subjects' previous subjects' choices?

RESPONSE: We originally conceived to show our participants behaviour from groups that were playing with no feedback between rounds, as this can lead to quite stable contributions. We decided to use our own previous data from previous as yet unpublished studies in order to save resources in the current study (if we had to create reference groups just for this study we would have greatly reduced our sample sizes). However, upon inspection we realized that behaviour from a recent punishment experiment of ours was more stable than the no-information experiment so we used that. These ‘reference’ data were not created with the purpose of aiding this study, they were created in the baseline treatments of other experiments which we hope to publish soon. We have now made it clear that it was one of our own previous studies.

Line 285. The passage beginning "Meaning that the..." is a sentence fragment.

RESPONSE: We have edited the paragraph to make full sentences, thank you.

Lines 379-395. In the Forced-Choice test, approximately 38% of people chose to copy the common pattern rather than the successful pattern. In the Costly-Choice test, exactly 35% of the

subjects who paid for information did so in order to see the common pattern. These percentages, obviously are identical, suggesting that the free-choice vs. costly-choice manipulaton makes no difference in assessing people's informational preferences. For this reason, I would like to see the results of an binomial sign test in which the data from the two data sets were combined, giving a total of 39 subjects who wanted to see the common pattern and 64 who wanted to see the successful pattern.

RESPONSE: We understand your reasoning and in fact we had already reported this, but it was not very clear, sorry for not making this clear. In fact, 37 subjects (23+14), not 39, chose the common pattern, and 64 chose to success (38+26). So if, as you suggest, we exclude the 20 who chose to not pay for either, then the binomial sign test is 64 successes out of 101 (64+37) trials. The two-tailed significance value is 0.0093 (i.e. just under 1/100). Upon inspection we realize we had accidentally reported the 1-tailed p-value (i.e. 1/200 instead of 1/100, the software reports both, the one tailed first which is probably why we accidentally reported it), we have now fixed this mistake and made the logic of the test rational clearer, "*Combining both choice tests together, we find a significant preference to be shown the Success pattern (Binomial Sign Test excluding those 20 individuals that chose not to pay for any information, two-tailed P-value on 64 choices for Success pattern in 101 trials = 0.0093, Figure 5).*"

Referee: 5
Comments to the Author(s).

Review

The paper entitled "A preference to follow successful rather than common behaviors in human social dilemmas" tries to advance our knowledge about the evolution of cooperation. The authors use an experimental decision-making task to examine the use of different social learning techniques in the public goods game, and by extension for cooperative actions per se. Based on the experimental results, they conclude that success-based information is preferredly copied compared to a "common- pattern" behavior and that success-biased learning can significantly reduce cooperative action. The authors conclude that the previously stated hypothesis, that a psychological tendency to not deviate from a norm (operationalized in the "common pattern treatment"), can not be the explanation for the evolution of human cooperation.

I welcome the author's aim to question existing reasonings for the evolution of cooperation. I do think that theoretical and empirical tests of hypotheses associated with the evolution of cooperation are important contributions to move the field forward. I do, however, have several crucial reservations, both regarding the design, and the conclusion drawn from the experiment. My review will focus on the experimental design, as previous reviews have already discussed several theoretical implications. Below, you can find a list of comments to the authors. I start with the main comments in no particular order and follow up with some minor comments.

RESPONSE: thank you for your valuable time as a reviewer and thoughtful comments.

1. Control/baseline conditions not clearly defined

To study the effects in the "common pattern" condition, the authors compare the decision in the treatment condition to what would have happened if participants have chosen by chance among the 21 options for contributing (e.g. ll 318 & ll. 330). This does not appear to be a very good baseline condition.

Simply put, the observed behavior could vary because of the treatment, but also because of other peculiarities of the decision-making task. Most participants do not make random guesses in a decision-making task. For example, a number in the middle, or at the extremes, may be more commonly chosen. This skew would also accumulate in the averaged choices of the "common pattern". Simply put, the estimated treatment effects are possibly biased.

ll.318 ff

When we showed individuals just the Common pattern by itself, they matched it 9% of the time, significantly more than expected by chance, and thus consistent with a desire to not deviate from the Common pattern (Binomial sign tests, random probability of matching set to 1 in 21 options = 0.048: frequency of matching Common pattern was 50 in 560 trials, P < 0.001; [...]"

A control condition with simple public good game (without having access to social information) may be a more suitable baseline. Another solution could be to compare the behavior in the treatment condition to results from baseline conditions in previous public good experiments.

RESPONSE: We understand your reasoning and agree that the null hypothesis of 1/21 options for comparing frequency of matching the common pattern may be biased upwards (so although 9% matching was significant compared to 1/21, it may not be very strong evidence for matching the common pattern with a more appropriate baseline). However, as this bias goes against out hypothesis we originally used it to be conservative, but you are right, ideally a treatment where individuals played with no information could be used to compare frequency of matching the common pattern to 'chance'. However, our aim was not to estimate accurate rates of copying the common pattern in its own right, but to examine how rates shift when information on success is also available. When comparing matching to the success pattern, our cumulative design means we also have the common pattern, so here the comparison to just the common pattern is an appropriate baseline to detect shifts in behaviour between treatments. We include histograms of all decisions by round for each treatment in the supplementary materials so readers can take a closer look.

To keep things simple, shorter and focused on what matters, we have now removed the statistical results comparing to this null hypothesis, and maintained the between treatment comparisons. We have modified the text to explain that some matching is expected by chance and this may bias estimates upwards,

*"Copying can be defined in various ways. To keep things simple, we first recorded how often participants' contributions exactly matched the previous social information we had shown them (Figure 4**). Some matching is expected by chance, which will inflate estimates of copying, but we can still use this strict and objective test to compare relative rates of matching across treatments."***

We have also restricted the results to a more logical order, with the exact matching rates coming before the more general learning shown by aggregating each individual's behaviour across the five rounds.

1.1 Treatment condition "The most common pattern" is ambiguous

I am aware of the previous discussion in the first round of revision. As a response, the authors have changed the framing from "norm psychology" to "common pattern". I am afraid, however, that "common pattern" is not a good wording for the condition. It suggests that a common choice (maybe even the most commons choice) is observed. This is not necessarily the case. In the treatment condition "common pattern", the participants are presented with the global average of all choices. The average can, but may not align with what most participants have chosen (e.g the mode). This problem would warrant a reframing of the paper, as there is nothing "common" about a global average.

RESPONSE: the previous literature on conditional cooperation has typically used the group average and not the mode. Choosing the mode of 20 participants facing 21 options would not be very reliable or consistent round to round. The phrase 'common pattern' exists in the theoretical literature on cultural evolution (and is not well defined there either), we use that term in the paper to relate to those claims. To say 'there is nothing common about a global average' seems an extreme point of view. Future research can examine if our results hold when using the mode or median instead of the mean.

The treatment may also lead to a heterogeneous effect on the participants' strategies. Some participants may rightly infer that the average contribution is possibly useless information, even with regards to detecting a descriptive social norm (e.g. in a case of a binomial distribution with many decisions towards the two extremes). Instead, the treatment may be more correctly labeled as a "reference point" (Abeler et al., 2011; Fehr et al., 2011).

RESPONSE: Our free rider treatment controlled for providing a reference point. Showing participants the behaviour of the highest earners within each group (which is always the lowest contributors) should be useless information if individuals confidently understand the game or are motivated to match the common pattern.

One possible solution may be to present the most common choice in the most successful group. This appears to be a better comparison to the decisions of the most successful participants. It also solves the issue that the authors are comparing group-level averages (success condition) to global averages (common condition).

RESPONSE: we do not understand this reasoning, why restrict the sample from 20 participants to just 4 when trying to inform individuals about typical behaviour?

Overall, these reasons make it very difficult for me to understand the use of this treatment in this experiment, especially in comparison to the success-based information which holds information about the earnings of the models.

RESPONSE: it would be a confound if more than one treatment contained clues on how to make more money in the public good game.

1.2 Variation between the treatment conditions

The variation of information in the treatment dimensions is ambiguous. The study varies between information about:

1. Average **contribution** overall (common pattern)
2. Average of lowest **contribution** (free rider)
3. Average **contribution** of those with **highest earning** (success)

ll. 207-213

While Treatment 1 and 2 select the information based on the distribution of contributions, Treatment 3 detects the information based on the distribution of earnings. In this setting, we do not know what drives the treatment effects. We may as well title the paper "A preference to follow *information about earnings* rather than *information about contributions* in human social dilemmas"

For a cleaner comparison, the authors may have chosen another treatment for the common pattern. For example, as written above, the most common choice in the most successful group, which also includes information on contribution and on the earnings.

RESPONSE: we do not understand this objection. The key point is that we are testing if participants, that have had the usual instructions and have been assumed to understand the game, still prefer to learn about individual success.

1.3 Punishment in the model group

In the group that serves as a learning sample, peer punishment was possible. It is not clear, however, why it was allowed, or, by extension, why it wasn't allowed in the main experiment. The effects of punishment were not included in the selection of successful participants, and the participants from the main study are not informed about the opportunity for punishment in that round:

ll. *762-764*
*Instead, you and all the members of your group will receive some information on the decisions taken by other players in another experiment, who **faced the same decision** as you face today.*

There are issues with this design. The authors responded to the previous review arguing that punishment is not their main interest:

ll.142-143

"Although punishment is often part of gene-culture coevolutionary theories, it is important to realize that our use of punishment is not a key feature of this experiment."

And that it was used to stabilize the contributions:

ll. 140-131

"Our solution was to use a previous experiment that used peer punishment to stabilize mean contributions [64].."

I do understand this point, but it raises the question of why the authors decided to not include punishment as a feature in the main task. In the current form, especially without estimation of possible endogenous effects of the punishment levels on the results, the punishment leaves some form of noise to the social information, without serving a clear cause.

A previous reviewer had also raised the point of possible deception, to which the authors responded:

"We disagree, the information we gave participants was relevant to how to be successful in the decision phase of a public good game, which was the set-up they faced. Our information could not harm them. It would have been a problem if the current participants faced punishment and we told them how to be successful in an experiment with punishment but without including the consequences of punishment."

It is not clear to me, how telling participants that the previous group *"[...] faced the same decision as you face today",* when this is no true, is not deceptive at all. Positions on what amounts to deception may vary across different academic fields. But it is also true that many behavioral labs or journals will flag this as deceptive. At the least, the authors should have flagged this in their Ethics Review.

RESPONSE: the decision to contribute from 0-20 MU to the public good was the same. Games with punishment have two stages, but stage one was the same in both experiments.

1.4 Anticipating what other participants do

I have already presented a few arguments why measurement error may be present and why this is a fundamental problem for making causal claims. This may turn out to not be a problem, but the authors need to discuss possible heterogeneous causes for the effects they estimate.

To illustrate what I mean, we can ask: "Why do participants choose to copy the successful?" The answer may be that they want to increase their share of the earning? This is plausible and consistent with most arguments associated with success-biased learning. In the experimental setting of the paper, there may be at least one other explanation. When participants know that other participants receive the same information, they have good reasons to believe that co-

players could follow the success-based information and be less cooperative in the next round. In this situation, a player may not be copying successful information, but anticipate lower levels of cooperation. The authors have tried to control for this by adding the "free-rider" condition. Another possible control would have been allowing one group member to receive information.

RESPONSE: yes we controlled for this with our free rider condition. Your idea is interesting, although would introduce other difficulties and would have drastically increased the costs of the experiment. In our design the information was common knowledge, and our free rider treatment controlled for these other affects you mention.

Other Comments

1.5 Median vs mean

For some analyses, the authors chose to use the *Mean*, in others, they chose to use the M*edian*. A short discussion of why either was chosen would strengthen the paper. Perhaps, such a discussion may lead to the conclusion that the *Mode* would have also been a good choice to illustrate a "common pattern". However, this is a minor comment, as the authors may have chosen to keep the design constant from previous works.

RESPONSE: Yes, using the mode is problematic in such a large strategy space (21 options). We used medians in lines ~280-90 (section Learning from others' success.) because the data here were not normally distributed. We have made this clear in the text. It is customary in public good games papers to show the mean contribution per round, even though the median may be more justified. In our figures we provide both the mean and the median (either in main or supplementary figures) to give our readers more information.

1.6 Caption Figure 3

I can not find the information about the median in the Figure, as stated in the caption.

ll. 302-304

"Shown are the distributions, mean (large semi-transparent point) and median (large solid, black rimmed, point), of these individual mean contributions, depending on treatment."

RESPONSE: perhaps the figure did not reproduce correctly in the submission process (pdf conversion problem). They are both there.

1.7 Typos

ll 145 "siituation"
ll 146 "defition"
ll. 148 "invesitgating" ll 189 pattenr

RESPONSE: thank you and our apologies for these mistakes which we have now corrected.

1.8 Ambiguous wording

ll. 407

"Our results suggest that those previous studies [...]"

Provide a reference for "previous studies". Without knowing to which studies the authors are referring, it is difficult to evaluate this claim.

RESPONSE: our apologies, we have added 4 citations here (Fischbacher, Gachter et al. 2001, Gachter and Thoni 2005, Gunnthorsdottir, Houser et al. 2007, Fischbacher and Gachter 2010).

And:

ll. 41-43

"The theoretical mechanisms for such cultural evolution often rely on the hypothesis that humans learn by observing others (social 43 learning) and desire to follow common social behaviours (conform with social norms) [18]."

In the case of this formulation (i.e "often"), more references would be apt.

RESPONSE: we have added 5 more citations here. (Henrich 2004, Chudek and Henrich 2011, Richerson, Baldini et al. 2016, Apicella and Silk 2019, Smith 2020, Henrich and Muthukrishna 2021)

References

Apicella, C. L. and J. B. Silk (2019). "The evolution of human cooperation." Current Biology **29**(11): R447-R450.

Burton-Chellew, M. N. and C. Guérin (2021). "Supplementary Materials for Burton-Chellew & Guerin 2021: Decoupling Altruistic Punishment." OSF.

Chudek, M. and J. Henrich (2011). "Culture-gene coevolution, norm-psychology and the emergence of human prosociality." Trends in Cognitive Sciences **15**(5): 218-226.

Fischbacher, U. and S. Gachter (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." American Economic Review **100**(1): 541-556.

Fischbacher, U., S. Gachter and E. Fehr (2001). "Are people conditionally cooperative? Evidence from a public goods experiment." Economics Letters **71**(3): 397-404.

Gachter, S. and C. Thoni (2005). "Social learning and voluntary cooperation among like-minded people." Journal of the European Economic Association **3**(2-3): 303-314.

Gunnthorsdottir, A., D. Houser and K. McCabe (2007). "Disposition, history and contributions in public goods experiments." Journal of Economic Behavior & Organization **62**(2): 304-315.
Henrich, J. (2004). "Cultural group selection, coevolutionary processes and large-scale cooperation." Journal of Economic Behavior & Organization **53**(1): 3-35.
Henrich, J. and M. Muthukrishna (2021). "The Origins and Psychology of Human Cooperation." Annual Review of Psychology, Vol 72 **72**: 207-240.
Hertz, U. (2021). "Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms." Proceedings of the Royal Society B: Biological Sciences **288**(1952): 20210293.
Richerson, P., R. Baldini, A. V. Bell, K. Demps, K. Frost, V. Hillis, S. Mathew, E. K. Newton, N. Naar, L. Newson, C. Ross, P. E. Smaldino, T. M. Waring and M. Zefferman (2016). "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence." Behavioral and Brain Sciences **39**.
Smith, D. (2020). "Cultural group selection and human cooperation: a conceptual and empirical review." Evolutionary Human Sciences **2**: e2.