Dear Editor,

We adjusted the Stage 1 manuscript to reflect the reviewers' comments. In this rebuttal letter, for each question/comment by the reviewers, we created a point 1), 2)... and responded immediately below. All modifications are highlighted in yellow in the updated manuscript.

As recommended, we:

A. Increased the control group size by 10%
B. Dropped one life stage (young professionals) and are now focusing on optimizing our intervention on high school and university students
C. Introduced positive emotions in our psychological questionnaire

We thank you and the reviewers for your time and help in significantly improving the manuscript.

Anna, Cameron, Gianluca, and Andrea

Reviewer 1: Jana Kesenheimer (signed)

Dear Anna and greetings to the entire research team,

I had the pleasure of reviewing your registered report in which you present a comprehensive long-term study involving approximately 50 participants undergoing an intervention, with effects compared to a control group of around 50 individuals. I believe the described approach is a valuable initiative to confront interested (yet not actively engaged) students and other young individuals with climate change, providing them with motivation for action. The theory-guided intervention based on the 12 factors is particularly promising. Additionally, the ideas for analyses considering interaction effects over time involving group, time, and person (as a random effect) levels are sensible.

While reading, some concerns and ideas arose, which I will describe below:

1) If I understand correctly, the control group is keeping diaries (EMS) for a total of 12 weeks (3 months). A 30% drop-out rate seems optimistic, especially given that there is a 3/100 chance (roughly) of winning €150 at the end. It might be worthwhile to research dropout rates from other environmentally focused diary methods, and adapt this rate, if necessary.

   Dear Jana, thanks for all your time and comments. Thanks also for suggesting to consider EMA in our dropout rate estimate. To clarify, the experimental group will do 12 weeks of EMA, vs. the control group only 6 weeks. Following your advice, we found a meta-analysis looking at dropout rates for longitudinal studies using EMA, ranging from low-effort sampling schedules to highly intensive sampling schedules (such as daily assessments carried over for 6 months; Wrzus et al, 2023). This meta-analysis suggests an average dropout rate of 10%. Additionally, we found two studies using experience sampling methods for environmental-related behavior, indicating 92% (183 participants out of 199; Richter et al., 2022), and 83% (Nielsen et al., 2021) compliance by the initial sample with daily diary keeping. This evidence supports that our 30% expected dropout rate for the experimental group may be appropriate, maybe even conservative, for this study. To reflect this information, we raised our control group sample size by 10%. We added this in the manuscript, on page 6, line 19:

> *"We expect a 30% drop-out for the experimental participants; this includes the dropout expected due to the intensive in-person intervention (30% as suggested by Castiglione et al., 2022) and to the extended behavioral reporting (10% as suggested by Wrzus et al, 2023; Richter et al., 2022; Nielsen et al., 2021). For the control participants, we expect a 10% dropout, due to the behavioral reporting alone. All considered, we raised our sample size to N=65 for the experimental group, and N=55 for the control group.''*

2) EMS has a significant drawback not yet mentioned: participants can mention anything, which, while acknowledged as an advantage, may lead to large subjective distortions when only frequency is queried. Personal biases (personality, attitudes, etc.) might not be reflected, potentially obscuring the results. For example, we conducted a study in which environmentally less conscious individuals reported partly mundane activities (e.g., using a lid while cooking to save energy), while others understood more elaborate behaviors (e.g., refraining from a flight and opting for a several-hour train journey): https://doi.org/10.3389/fpsyg.2022.883704. If both are equally weighted in terms of frequency, personal biases (personality, attitude, etc.) are not reflected, which could obscure the results. Honestly, I'm not sure how you could address this issue, but I wanted to inform you that such distortion could occur in EMS.

This touches on the important point of accounting for behavior *impact* in a frequency analysis. We address this issue by looking at individual versus collective behaviors (where individual actions tend to have a lower impact compared to collective actions). Further breaking down each behavior type (individual and collective) by impact in terms of "action elaboration" or "corresponding CO2 saved" (e.g. low, medium, high) would require a technical evaluation of each behavior reported by the participants that may go beyond our competences. However, we do want to avoid including behaviors that are so low-impact to be irrelevant, or even not directly related to emission reduction. In fact, it is often misunderstood which pro-environmental behaviors are specifically pro-climate behaviors (i.e. reducing emissions), which are the focus of our study series (e.g. Kause et al., 2019).

To address this issue, we compiled a *coding list* of the most relevant emission reduction behaviors, for each behavioral category considered (individual: transportation, energy consumption, diet, consumerism, and collective: education, civic engagement, advocacy, activism), using the following sources:

Individual behavior:

- Marchi et al., 2021
- Brandenstein et al., 2023
- Dietz et al., 2009
- Our World in data - Emissions by sector: https://ourworldindata.org/emissions-by-sector

Collective behavior:

- The Commons Social Change Library: https://commonslibrary.org/climate-activism-start-here/

Marchi, L., Vodola, V., Visconti, C., Gaspari, J., & Antonini, E. (2021). Contribution of individual behavioural change on household carbon footprint. E3S Web of Conferences, 263. https://doi.org/10.1051/e3sconf/202126305024

Brandenstein, N., Ackermann, K., Aeschbach, N., & Rummel, J. (2023). The key determinants of individual greenhouse gas emissions in Germany are mostly domain-specific. Communications Earth & Environment, 4(1), 422.

Dietz, T., Gardner, G. T., Gilligan, J., Stern, P. C., & Vandenbergh, M. P. (2009). Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions. Proceedings of the National Academy of Sciences of the United States of America, 106(44). https://doi.org/10.1073/pnas.0908738106

Specifically, we selected emission-reduction behaviors that are commensurate with a daily report such as EMA (e.g. decisions such as not having children may be too big and long-term to fit a daily behavioral report), and that take into account our student audience (e.g. students may not have decisional power to weatherizing their home as most rent out their homes). Any behavior reported by the participants fitting a description in this list will be retained. All the other behaviors will be excluded, unless they are clearly relevant for emission reductions, according to our sources and judgment; in this case, we will update the coding list by adding a description of the novel relevant behavior.

We added this explanation on page 18, line 4:

> *"....based on a coding list of relevant pro-climate behaviors, which we have compiled from multiple sources (see Appendix A)."*

And in Appendix A, on page 56, line 29:

> *"The behaviors listed by the participants will be screened using the following coding list of relevant emission reduction behaviors, for each behavioral category*

considered (individual: transportation, energy consumption, diet, consumerism, and collective: education, civic engagement, advocacy, activism). This list was compiled from multiple sources (see below). Specifically, we selected emission-reduction behaviors that are commensurate with a daily report such as EMA (e.g. decisions such as not having children may be too big and long-term to fit a daily behavioral report), and that take into account our student audience (e.g. students may not have decisional power to weatherizing their home, as most rent out their homes). Any behavior reported by the participants fitting a behavior description in the following list will be included. All other behaviors will be excluded, unless relevant for emission reduction, according to our sources and judgment; in this case, we will update the coding list by including a description of the novel relevant behavior."

Behavior type	***Behavioral category***	***Relevant behaviors***
Collective	*Civic Participation*	· Donate to a climate cause (e.g., a climate org, a climate campaign, a pro-climate political party) · Support a climate campaign (e.g., sign or circulate a petition) · Join educational events organized by a climate organization · Vote for a political candidate/party proposing emission reduction policy
	Advocacy	· Contact one's representatives to ask them to enact emission reduction policies · Collaborate with institutions and politicians to create emission reduction policies · Give professional advice to institutions and politicians on emission reduction policies · Join a climate organization that advocates for emission-reduction policies · Start a new project, campaign, program, or group/organization to communicate/collaborate with politicians to enact emission reduction policies
	Activism	· Organize events such as demonstrations, petitions, sit-ins, or campaigns to put pressure on politicians to enact emission-reduction policies · Attend events such as climate demonstrations, protests, sit-ins to put pressure on politicians to enact emission-reduction policies · Join a climate organization that puts pressure on politicians to enact emission-reduction policies, via demonstrations, protests, sit-ins · Start a new project, campaign, or group/organization to pressure politicians to enact emission reduction policies · Take legal action against big emitters (companies or institutions)
	Education	· Speak to others about the climate crisis and emission reduction behaviors · Circulate information about the climate crisis and emission reduction behavior on social media · Organize educational events about the climate crisis
Individual	*Food consumption*	· Reduce meat consumption · Reduce dairy consumption · Eat seasonal products · Eat local products · Reduce food waste
	Transportation	· Use means of transport that do not burn petrol (walking or cycling) · Use public transports · Substitute higher emission transports with lower emission ones (e.g. carpooling instead of driving alone, or carpooling instead of flying)

	Energy use	· Shift to renewable energy sources · Reduce electricity consumption (switching lights off, unplug appliances, reduce appliance use) · Reduce energy consumption (reduce home heating and cooling, reduce water use) · Reduce cooking gas use · Limit online data storage and transfer
	Finances	· Divest from fossil-fuel related services (e.g., move utilities provider from fossil-based energy providers to 100% certified renewable energy providers, move one's savings from a fossil-tied bank to a credit union) · Reduce the amount of unnecessary products bought · Buy fewer high-quality products instead of many low-quality ones · Buy products produced locally

Regarding personality type, we will explore the relation between the Big 5 and the behavior type participants engage in. However, given the low power, this is not part of the severe tests in Stage 1.

3) The effort required for participants seems substantial. Likely, only individuals already living environmentally friendly lives will persist, raising questions about the variance left for improving their behaviors – as you accordingly suggested in the report. Considering the anticipated ceiling effects and drop-out rates, I personally find the sampling size to be a bit small. Additionally, the power analysis mentions 14 replicates, while a minimum of 4 replicates is discussed on page 15. Should this number be reduced in the sampling accordingly?

Our sample may self-select based on a pre-existing concern for environmental issues and, accordingly, already lead a pro-environmental lifestyle. To address the possibility of a ceiling effect, it is important to consider the reviewer's point 2). The participants attracted to this study may already engage in low-effort pro-environmental behaviors, such as low water consumption, recycling, etc., but not necessarily in more sophisticated *pro-climate* behaviors, i.e. all those behaviors aimed at reducing greenhouse gas emissions. Therefore, we do not expect a high frequency of emission reduction behaviors, such as those in our coding list. Additionally, one of our prerequisites for recruitment is that people are not collectively active on the climate crisis; therefore, at least for the collective behavioral categories, there will be room to improve (note that 50% of the intervention content aims to increase collective engagement). For the above reasons, we do not expect a critical ceiling effect. We clarified this in the text on page 32, line 20, as follows:

> *"The highly motivated participants joining and completing the study may start off with high pro-environmental values and a pro-environmental lifestyle, posing the risk for a ceiling effect. However, even among those caring for the environment, it is often misunderstood what pro-environmental behaviors are specifically emission reduction behaviors, which are the focus of our study series (e.g. Kause et al., 2019). Most participants attracted to this study may already engage in low-effort pro-environmental behaviors (such as recycling, low water use,*

> *etcetera), but not necessarily in more sophisticated pro-climate behaviors (i.e. all those behaviors aimed at reducing greenhouse gas emissions). Therefore, we do not expect a high frequency of emission reduction behaviors such as those specified in our coding list, even among those caring about the environment. Additionally, one of our inclusion criteria is to not be civically engaged for the climate; therefore, at least for the collective behavioral categories, there is extensive room to improve (note that 50% of the intervention content aims to increase collective engagement). For the above reasons we do not expect a critical ceiling effect."*

Regarding the 4 replicates mentioned by the reviewer, these refer to the *maximum number of EMA surveys that can be missed, to avoid being excluded*; i.e. participants need to fill a minimum of 10 surveys out of 14 to avoid exclusion. However, in responding to point 12. of reviewer 3, we raised this number to 5 surveys that can be missed. We clarified this by modifying the text as follows, on page 17, line 3:

> *"A. they complete less than nine daily surveys out of 14 (i.e. they miss more than 5 surveys), in any of the baseline or final EMA behavioral tracking stages…"*

Given the incentives (point accumulation) to fill the EMA surveys, we expect most participants to fill around all 14 surveys, which is why we have accounted for 14 repetitions in our power analysis. But we ran our power analysis again on PANGEA, considering 10 repetitions and keeping the other parameters constant (*d* = 0.25, power = 0.85); this leads to a sample size of N=102, which is our final sample size already.

Therefore, considering both the dropout rate addressed in our answer to point 1), and the ceiling effect addressed here, we expect that N=65 experimental participants and N=55 control participants will grant enough power to detect the effects of interest.

4) It would be great to learn more about the sampling of the study you described on pages 2-3 (e.g., were the participants also young, and how many were there?).

We report here the sample characteristics from Castiglione et al., 2022:

> "*We recruited 170 UCSD students between 18 and 38 years of age through online department platforms and by flyers posted on the campus of UC San Diego. The inclusion criteria were that participants were enrolled at UCSD, that they believed in anthropogenic global heating and that they had no or low prior engagement in climate activism (see Screening Survey). The drop-out rate was higher than expected: 30% of the participants dropped out before the study began, and 13% dropped out during the three-month study period, so the final sample was N = 96 (22 males, 72 females and 2 unspecified). 20% of the participants were Caucasian, 51% Asian, 18% Hispanic/Latino and 11% had other ethnic/racial backgrounds. 75% of the participants identified as Democrats, 5% as Republicans and 20% as Other*."

5) Participants might be aware of what you are investigating, leading to biased responses in line with the expectation that the sessions should have an effect. How do you plan to account for this bias?

   This is an important point. We expect EMA to prevent this response bias, as it would be difficult for participants to keep track of their day-by-day responses from the baseline EMA period, and to calculate better responses in the final EMA period, which will be 6 weeks later. The daily report schedule forced by EMA makes it more difficult to lie to meet the researcher's expectations.

   The focus on negative affects regarding climate change (anxiety) is noted. It would be beneficial if you could elaborate on why positive emotions like hope are excluded. Another crucial negative emotion seems to be anger, which doesn't paralyze but activates (see: https://doi.org/10.1016/j.joclim.2021.100003).

   Thanks for this observation. Anger will be addressed by the intervention (it has a dedicated activity in Module 1), and is already included in the psychological questionnaire as part of the Affective Engagement factor. We added positive emotions to our psychological questionnaire (based on Landmann et al., 2023), as they are indeed relevant and are likely to be raised by multiple activities in our intervention. For example, empathy may increase in Module 1, where participants will research extreme events and their impacts on people living in their region. Hope may increase in the modules addressing system change and civic engagement. This change can be found in Appendix A, on page 48, line 34.

6) One last thought: considering your assumption that planning on one-day influences implementation on another, incorporating AR1 models might be beneficial.

   Thanks for this suggestion. We are interested in looking at Action Score and Plan Scores as separate aspects of behavior, with a dedicated linear mixed effect model accounting for the repeated measure of each score. However, we will still account for their dependence upon each other by analyzing the evolution of the percentage scores (i.e. Action Score/Plan Scores).

Best regards and good luck with your research! I am looking forward to follow this research progress.

Jana

Thank you again for the helpful suggestions!

=============

Reviewer 2: Helen Landmann (signed)

The registered report „A climate action intervention to boost individual and collective climate mitigation behaviors in young adults" proposes a set of five intervention studies with control groups and three measurement time points. I highly appreciate the attempt to investigate the causal effects of pro-environmental interventions in longitudinal studies. The design allows for pre-post comparisons, comparisons with a control group and the investigation of effects three months after the intervention. It is also fortunate that the studies will take place in different countries (Italy and the Netherlands). The study plan requires much effort and time. Such intervention studies are highly practically relevant but so far rare.

1) In the summary of the psychological obstacles of pro-environmental behavior I missed the point that most pro-environmental behaviors are embedded in a structure of a social dilemma (the behavior that provides the largest short term benefits for the individual is different from the behavior that benefits the collective in the long run; see Claessens et al., 2022; Steg et al., 2014).

   Thanks for this important point. We now added it to the text:

   > *"Some examples are A. the perception that global warming is far away in space and in time, and not directly salient to us (McAdam 2017; Weber 2006); B. the social dilemma inherent in the decision to act for the long-term collective versus immediate individual interest (Claessens et al., 2022; Steg et al., 2014); C. the paralyzing fear of a crumbling future (Clayton, 2020); D. the power structures supporting the current extractivist economy seem inaccessible to individual citizens (Schmitt et al., 2020); E. climate change is a widely-distributed problem (spanning from agriculture and biodiversity to socio-political stability) and the solution is global and complex, which reduces people's sense of personal efficacy (Castiglione et al., 2022)."*

2) As correlates of individual and collective pro-environmental behavior, you mention negative emotions, individual and collective efficacy, environmental identity, social norms, and cognitive alternatives. You may want to consider the role of positive emotions and activist identity in addition (see Landmann & Neumann, 2023).

   We agree. Positive emotions are an important addition. In this revision, we added hope, human connection and empathy to our psychological questionnaire (also see our reply to reviewer 1, point 6.). Regarding activist identity, our psychological questionnaire currently includes items measuring this construct, under the "Self-identity" factor, in Appendix A.

3) The distribution of studies in Italy and the Netherlands is a bit one-sided. Only one study is planned in the Netherlands, the other four in Italy. I could not find an argument for why this isn't more balanced.

Most authors in this series of studies are based in Trento, Italy, making this the primary site. Our main goal is to test the effectiveness of our intervention on young adult audiences at different life stages. To add a cross-cultural approach to our investigation, we included a sample from a different setting. However, we are unable to completely balance each study by life stage and cultural setting, due to our limited resources. We chose to replicate the intervention on university students in Amsterdam, given that one of the authors is based there. However, if we found strong differences in the way Italian and Dutch university students react to the intervention, this would motivate comparing the other life stages cross-culturally, in future studies. We added this to the text, on page 27, line 19:

> *"Different reactions may emerge across Italian and Dutch university students as well; this would highlight the importance of cross-cultural differences that could be later studied in high school students as well, and used to optimize the intervention further."*

4) I would prefer more information about the expected drop out rate - do you expect drop out only for the experimental group, not for the control group? I think you should calculate some drop outs in the control condition as well.

We increased the sample size of the control group to by 10% to account for the EMA effort and overall drop-out rate (also see reviewer 1, point 1.).

5) Please clarify whether participants are randomly assigned to the experimental or the control condition or whether they decide themselves if they want to participate in the intervention. You write that participants will receive a certificate for participating in the intervention. What about those in the control group? Do they have the same incentive for participation (including the certificate) or is the appeal for participation for those in the control group different?

Only the experimental group will receive the certificate, but both groups will be entered into a raffle where they may win money prizes if filling out the questionnaires according to our instructions (see Appendix B, incentive system).

The two groups will be recruited with two different flyers: the 65 experimental participants via a flyer advertising the study as a series of six 3-hour laboratories plus online questionnaires, and the 55 control participants via a flyer advertising the questionnaires only. This is because we do not expect 120 people to be willing to sign up for the meetings. We are aware that recruiting the experimental and control groups differently may cause different initial motivations: experimental participants are ready to engage in a series of in-person meetings, while control participants only fill out some questionnaires online. Lower initial motivation could cause the control group to drop out more easily; this risk may be mitigated by raising the control group size from 50 to 55 participants. Additionally, we will compare the mean frequency of actions planned and performed by the two groups in the first two weeks of EMA as a test of whether the initial motivation is similar between conditions.

We explain the above in the text:

> *"The experimental group will be recruited to participate in six 3-hour meetings on climate education and fill online questionnaires, while the control group will only complete the questionnaires due to the difficulty of recruiting 120 participants for the in-person meetings. However, we will account for potential differences in initial motivation by comparing the mean frequency of reported behaviors in the baseline EMA between the two groups."*

Beyond initial motivation, recruiting with different flyers should not hinder the group comparison, in the first study of each life stage. In these two studies, the goal of having a passive control group is to disentangle whether pro-climate behavior increases in the experimental group (from pre- to post-intervention) due to the intervention itself, or to some external event increasing behavior for everybody.

6) I disagree that the comparison with the control group completely cancels the possible effect of social desirability. The perceived expectation of reporting pro-environmental behavior might be higher in the experimental condition, in which participants are repeatedly confronted with environmental issues. However, assessing pro-environmental behavior by self-report is still suitable for the planned studies. If people are asked very specific questions about their behavior as in the EMA, it is difficult to lie. Thus, I think the self-report measures of pro-environmental behavior are fine, I just suggest to discuss its limitations differently.

Thank you for this observation. We have edited the text on page 13, line 14, as follows:

> *"One weakness of EMA is that it uses self reports that may be affected by a social desirability bias (Koller et al., 2023); however, in our design, the comparison with the control group should cancel this effect (especially for the studies employing an active control group, see Optimization contingencies and procedures section). Additionally, when people are asked to report their behaviors on a daily basis, it is more difficult to lie."*

7) The mixed effects linear model seems suitable to test H1. I wondered why you chose a different analyses for testing H2 (repeated measures ANOVA) – mixed effects would be also suitable here.

The DV in H1 is a repeated measure of behavioral scores, for which a linear mixed effect model is appropriate. On the other hand, the DV in H2 is a single score for each psychological factor (the mean of the questionnaire items for that factor), which requires an ANOVA test (e.g. the lmer4 function in R would automatically run an ANOVA test if used to analyze single-score data).

8) For testing whether the effects of the intervention differs between the studies (H4) you need to add condition as predictor as well as its interactions with study ID. This would make your planned regression analyses very complex. It may therefore be more

> convenient to test H4 with mixed effects models including condition and study ID as predictors.

Similarly to our answer to point 7), linear mixed effect models are not feasible with a mean (not repeated) action score. However, as pointed out by the reviewer, we added “Condition” and “Condition*Study ID” interaction as predictors in the regression of H4. We clarified this on page 21, line 8:

> *“We will run a regression model of the mean final Action Score for all the studies, with predictors: A. the mean baseline Action Score, B. Condition, C. Study ID, and D. the interaction Condition*Study ID. Study ID indicates the temporal order in which each study was conducted (1=first study, 2=second study…), within each life stage. We expect that the Condition*Study ID interaction will be significant, i.e. that the later the iteration, the greater the difference will be between the two conditions’ final Action Score (after accounting for the mean baseline Action Score of each).”*

I’m already looking forward to seeing the results!

> *Us too and thank you for this helpful review!*

Claessens, S., Kelly, D., Sibley, C. G., Chaudhuri, A., & Atkinson, Q. D. (2022). Cooperative phenotype predicts climate change belief and pro-environmental behaviour. Scientific Reports, 12(1), 12730. https://www.nature.com/articles/s41598-022-16937-2.pdf

Landmann, H., & Naumann, J. (2023). Being positively moved by climate protest predicts peaceful collective action. Global Environmental Psychology. https://doi.org/10.23668/psycharchives.13186

Steg, L., Bolderdijk, J. W., Keizer, K., & Perlaviciute, G. (2014). An integrated framework for encouraging pro-environmental behaviour: The role of values, situational factors and goals. Journal of Environmental Psychology, 38, 104-115. https://doi.org/10.1016/j.jenvp.2014.01.002

=============

Review by anonymous reviewer 1
Thank you for the opportunity to review this programmatic registered report. The goals of this work are timely and important.

Thank you for your time and these helpful suggestions.

Major feedback:
The overarching hypotheses are sensible but not specific. I understand that this may be challenging given the iterative nature of the design but given the larger number of outcomes and psychological factors, it would be helpful to know more about: 1) **how hypotheses will be refined over studies** (e.g., if some outcomes but not others are altered in study 1, will this prior information be incorporated in some way in study 2?) and 2) **how the false positive rate will be minimized across outcomes and moderators.**

1) This program of research has an implicit **causal model**—in which the intervention should increase the targeted psychological mechanisms, which in turn should change behavior—so it is unclear why methods for testing **causal pathways** are not actually proposed. In its present form, H1 maps onto the c path (X --> Y), H2 maps onto a path (X --> M), and H3 contains a combination of the b path (M --> Y) and the indirect path (X --> M --> Y). Rather than testing these hypotheses in separate models, it would be more parsimonious and informative (i.e., to actually test the implied indirect paths) to test these hypotheses in a single model path model. There are various methods to do this (see e.g. https://abcdworkshop.github.io/slides/ABCD*SEM*Theory_2021.pdf) and extensions that could be used for the EMA data through Bayesian multilevel modeling (in the brms package; see e.g. https://m-clark.github.io/models-by-example/bayesian-mixed-mediation.html). If a Bayesian framework were employed, this would also have the advantage of enabling the priors to be updated from one study to the next based on the results.

We thank the reviewer for the suggestion of running one comprehensive causal pathway model, such as SEM or regression-based mediation, which we considered while designing our analyses. We discussed this between the authors.

While not using a single model for all our analyses, in our current analytical pipeline we are still testing mediation. The mediation pathway we are testing is explained below.

1) X→Y (H1)
2) X→ M (H2)
3) X→M→Y (H3)

Where X=Condition (having or not participated in the intervention), M=Psychological factors, and Y=Behavior. In particular:

1) H1: here we test whether our intervention increased behavior for the experimental versus control group (via a linear mixed effects model)
2) H2: here we test whether our intervention increased the psychological factors for the experimental versus control group

3) H3: here we test whether the factors that were significantly raised by our intervention (we select only the factors significantly raised, as per H2), predict a change in behavior.

The main pitfall in our mediation test sequence was that the behavioral DV in the H1 analysis (repeated measure behavioral scores) is not the same as the behavioral DV in the H3 analysis (where we take the mean of the behavioral repeated measures). In fact, for testing mediation each variable should be kept in the same format across the three models (either mean or repeated measures). Therefore, we have now added an ANOVA model to our H1 testing, where the behavioral DV is the mean of the behavioral repeated measures (as in H3). We added this on page 19, line 7:

> *“**H1a**. For each behavior type (individual versus collective) we will compare the participants’ Action Scores tracked via EMA during the baseline, final and follow-up tracking period, between the experimental and control groups. First, to consider the breadth of the multi-day data (14 Action Scores per time point), we will run a linear mixed effect model, preserving the 14 repeated measures; this model will include as predictors Time (Baseline, Final), Condition (Experimental, Control), and the Time*Condition interaction, and Participant as random factor. To test the significance of the Time*Condition interaction, we will use sequential ANOVA decomposition of fixed effects, comparing the main model to a model including all the same predictors except for the Time*Condition interaction. Second, we will take the mean of the repeated measures of the Action Scores for each time point (baseline, final and follow-up), and run a repeated measure ANOVA, with factors Time* (within*: Baseline, Final, Follow-up) x Condition* (between: *Experimental, Control). ”*

We decided against fitting one SEM or regression-based mediation model for all hypotheses for the following reasons:

- We are testing a novel design (the only similar study to our knowledge is Castiglione et al., 2022, although the behavioral measure and intervention structure were different), with not much prior knowledge of which factors will be successfully raised by our intervention. Having control over each step of the mediation analysis gives us more flexibility and greater precision, i.e. to include in model 3 only the psychological factors that significantly increased due to our intervention (as per model 2).
- SEM is particularly helpful to test complex relationships among variables (IV, DV, moderators, mediators, covariates). At this very early stage of our investigation, we are more interested in the basic questions of: will this intervention increase any of the psychological factors and pro-climate behaviors? And if so, is there any relationship between the change in the factors and the behavior change? Our step-by-step analyses can answer these, and hopefully point towards more complicated relationships between these variables that could be tested with SEM in the future.
- SEM and regression-based mediation models are generally used in studies with large sample sizes; here, it might be very underpowered with only 120 participants and two conditions.

- None of the authors are experts in SEM, so this change would entail substantially more work than the current approach.

Lastly, Bayesian modeling requires enough prior knowledge to set up prior parameters, which we do not have in this novel design.

2. Regarding **sample size and power**, I am concerned that the estimated effect sizes for H2 and H3 are not realistic and will result in these tests being underpowered. Recent studies have shown that average published effects tend to be much higher than replication or preregistered effects (https://journals.sagepub.com/doi/full/10.1177/1745691612462588 ; https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00813/full). In the former study for example, preregistered between-person experiments had a median effect size of just r = .12 versus r = .34 for non-preregistered studies. If power is not increased, this could also negatively affect the iterative nature of the research program. That is, observing a false negative might prompt changes to the intervention that are unnecessary or detrimental, thus affecting subsequent studies. One approach for overcoming this issue would be to power study 1 to detect smaller effect sizes (similar to H1) to get a realistic sense of the effect size for this intervention, and then power subsequent studies accordingly.

We thank the reviewer for this suggestion. We agree that the effect size inflation is a potential concern. The main effect of interest in this study is behavioral change (H1), for which we have 85% power to detect an effect size of about d=0.25. This effect size is the average of many behavioral interventions analyzed by a meta-analysis (Nisa et al., 2019), and is likely to reflect a quite robust effect size compared to a single non-preregistered study. About H2 and H3, some of the studies we used to estimate our effect size were pre-registered, fully registered, or replications (e.g. for H2: Jachimowicz et al., 2018, Ramstetter et al., 2022, Castiglione et al., 2022, and for H3: Hamann et al., 2020, Jachimowicz et al., 2018, Wright et al., 2020); this reassures us on the robustness of this estimate.

Powering our study to detect a sample size as small as 0.25 for H2 and H3 yields very large sample sizes (e.g. 576 total participants for H2, considering f=0.12). Unfortunately, our resources do not allow for such sample sizes, given: A. the handling by our research team of six in-person 3-hour meetings with these many participants, and B. the high level of commitment required to participate, which makes recruitment particularly difficult. Additionally, such a small effect size for H2 and H3 may be out of scope, given the pre-registered literature mentioned above. So, while we agree that more power would be desirable, we suggest the studies are worth running regardless, if only to rule out the larger effects that this will be adequately powered for.

3. Regarding **attrition**, the authors state that they expect ~30% attrition, but as far as I could tell the target sample sizes do not include additional participants to make up for this. Given the very high level of engagement required for this study (and relatively low compensation), it seems likely that attrition will be high. It may also be wise to

preemptively plan for even greater attrition in study 1 and then adjust as needed in subsequent studies.

This makes sense. As specified on page 6, line 24, we increased our experimental group size from N=50 (as per power analysis) to N=65, to account for the 30% dropout rate due to the longitudinal design, and including the 10% dropout rate from EMA. Now, as suggested by the reviewers, we increased the control group size by 10% as well, to account for EMA dropout (see our reply to reviewer 1, point 1).

4. Another concern is the combination of iterative improvement and assessing generalizability different developmental and cultural samples. Each of these goals are useful, but when used in combination, they may make the result uninterpretable. That is, the authors expect that each iteration will improve the efficacy of the intervention (H4), but that may not be the case for more distal populations from the ones it was developed on. Thus, the final iteration of the intervention may not be the best because it simply works less well on high school students than it does for university students. Similarly, making changes in study 4 based on how well it performs in young professionals may result in a worse intervention for high school students who are in a different phase of life and have different affordances. An alternative approach would frame studies 3-5 as tests of generalizability and not use iterative improvement.

The reviewer raises a very important point. The main two goals of this programmatic research line are to: A. investigate the psychological dynamics underlying climate action across different young adulthood stages, and B. gradually update and optimize the intervention en-route, based on our sequential behavioral and psychological findings. But updating the contents and activities based on prior iterations makes the most sense if the intervention is tested repeatedly on the same audience (e.g. all university students, or all high schoolers or all young professionals).

To retain both these goals, we decided to limit our investigation to high school and university students, and run multiple studies on both. Because our current resources would not allow us to run two studies on all three life stages, we opted to forego young professionals, because:

A. Students have a more flexible lifestyle, with likely more room for individual and collective behavioral change (e.g. they have not yet picked a career, they likely have more time and opportunities to engage in collective action), making them a more promising population in which to look for behavior change.
B. Students have greater opportunities to bridge among different communities and generations (they may spend more time in recreational spaces or with their families relative to working people)
C. Students are more convenient to recruit, especially for the later iterations where we plan to employ an active control group; organizing a mock intervention will be more feasible in a school or university, e.g. taking advantage of already existing environment-related courses.

We added this to the text, on page 5, line 18:

> *“We chose a programmatic design to: A. investigate the psychological dynamics underlying climate action across different young adulthood stages (high school and university), in two cultures (Italy and the Netherlands), and B. gradually update and optimize the intervention within each life stage en-route, based on our sequential behavioral and psychological findings. The planned five studies are the following:*
> *University students:*
> - *University students in Italy (study 1)*
> - *University students in Italy (study 2)*
> - *University students in the Netherlands (study 3)*
>
> *High school students:*
> - *High school students in Italy (study 1)*
> - *High school students in Italy (study 2)*
>
> *Each study will yield one Stage 2 output, with results and discussion and a detailed description of the updates made to the intervention, and the motivations for those updates. In particular, this is an incremental Stage 1 submission; the current submission describes study 1 within each life stage, while the details of the following studies will be developed sequentially, based on the results of the previous Stage 2 output and the current state of the literature. The contingencies by which the intervention protocol will be updated along the way within each life stage are explained in the section “Optimization contingencies and procedure.” The updated Stage 1 protocol will be submitted again for re-evaluation, before studies 2-3 on university students, and before study 2 on high school students. All anonymized data, code, and materials will be shared openly through the Open Science Framework.”*

Following this change, we adjusted our H3 and the related analysis accordingly, see page 17, line 20:

> ***“H4: Intervention improvement over time.*** *The programmatic optimization process will make the intervention more effective at triggering individual and collective pro-climate behaviors within each life stage (high school and university students), over the iterations.”*

And page 21, line 6:

> ***“H4a.*** *For each life stage, and for each behavior type, we will assess the improvement in the behavioral engagement (both action and planning) induced by our intervention, over the iterations….”*

5. Relatedly, the current design does not include many individual difference measures or have any planned moderation analyses. Although the hypotheses are centered around average effects between groups, there will likely be a lot of heterogeneity in how effective the interventions are for individuals. Including a broader array of individual difference measures may help contextualize for whom the intervention is effective (and explain attrition) and generate moderation hypotheses that could be tested in subsequent studies.

The current design includes questionnaires on personality traits (see page 54, line 6). While we do not have enough power to justify any formal statistical analysis, we will visually explore possible trends of how different personality types respond behaviorally and psychologically to the intervention (i.e. more or less extroverted people experience a greater or smaller change in the psychological factors and in behavior). Because this will be purely exploratory, we do not include our analytical plan in the Stage 1 submission.

Other suggestions that may enhance this work:

6. Include measures of hope and resilience/well-being in addition to climate anxiety

Following the other reviewers' advice, we introduced positive emotions into our questionnaire, such as hope, empathy, and connection (see our reply to reviewer 1, point 6.). For what concerns resilience, we do not have a scale dedicated to this construct, but we do measure sub-factors of resilience. According to the resilience scale developed by Connor et al., (Connor & Davidson 2003) resilience is composed of sub-constructs such as personal competence (i.e. which we measure via our self-efficacy items), and control (which we tackle via our cognitive and behavioral control items). So, overall the coverage looks adequate.

Other questions that arose while reviewing:

7. The modules have solid coverage in topics but it is noteworthy that there is no focus on developing strategies for resilience and hope amidst the climate crisis. This seems to be an essential ingredient that buffers and enables people to continue to engage in climate action, particularly as they learn more and may feel anxiety, grief, hopelessness etc (see e.g. https://pubmed.ncbi.nlm.nih.gov/36502586/).

We agree that is important. Fortunately, the modules focused on civic engagement (modules 5 and 6) are rich with materials and activities based on effective collective action and teamwork. The contents and activities in these modules are designed to increase self- and collective efficacy, and cognitive and behavioral control, which, as mentioned above, are at the core of resilience (Connor & Davidson 2003).

8. How many people will be in each intervention module group? If there are multiple groups, how will intervention fidelity of the module be tested (e.g. between different leaders)?

Each module will be attended by all 65 experimental participants, with the same coordinator leading all intervention modules, in all the studies.

9. What happens if the intervention does not affect the psychological targets or change the behaviors, and it needs to be drastically overhauled? The current model suggests small incremental improvements but in the absence of pilot data testing the intervention modules (e.g. with manipulation checks to ensure they're targeting the processes intended), it seems possible that deeper changes may need to be made.

While the overall structure of the intervention will be kept, the current optimization protocol allows for deep changes. The changes currently allowed are A. the module contents and activities (expanded or removed), B. the targeted psychological factors and their questionnaires (added or substituted), C. the module themes, order and format (e.g. the ratio of presentation vs. activity time), D. the tutoring approach, and E. the reward system. Based on our results, we will be able to change the intervention deeply (according to the criteria above).

10. The current set of collective actions focuses primarily on traditional political targets, but what about collective actions in an individual's more local sphere (e.g., organizing to change the food served at the school cafeteria to reduce climate impact, working with neighbors to plant trees)? These may be more accessible and potentially increase self-efficacy since the effects of their actions are more proximal.

We agree that this would support action. There is a balance between intervention standardization and local context. While some of the theoretical content of the modules will address the history of social movements and the related traditional political targets, some will focus on local politics and institutions within reach of the participants. Additionally, in the group activities, the participants will research local institutions and their emissions, and practice developing campaigns to convince those institutions to cut them.

11. What happens if individuals do not plan any actions? Will their percentage scores be treat as NAs or 0s? And what happens if individuals engaged in unplanned actions? Will they be able to report those and will they be counted toward their action scores?

If participants open the daily survey but do not report any planned or performed actions, their Action and Plan scores will be 0. If they do not open the survey we will assign "NA" to both scores. We clarified this on page 18, line 14:

> *"If participants open the daily survey but do not report any planned or performed actions, their Action and Plan scores will be 0. If participants do not open the survey we will assign "NA" to both scores."*

As mentioned on page 13, line 12, participants will be able to report actions they did that day even if they had not planned them the day before. These actions will be included in the Action score, which we clarified on page 19, line 10:

> *"The Action Score will be the absolute number of relevant actions participants performed that day, across all eight categories (e.g. three actions performed). This will include actions that were not planned the day before."*

12. What is the justification for the exclusion thresholds listed? If possible, using methods that include all available data but adjust for differences between estimates of individuals with more versus less data (e.g. shrinkage in multilevel modeling) may reduce bias in the estimates.

We list three exclusion criteria: A. completing less than ten daily surveys out of 14, in any of the baseline or final EMA behavioral tracking stages, B. failing to fill any one of the baseline or final psychological questionnaires, C. failing to attend more than one out of six intervention meetings.

For criterion A., while we are aware that linear mixed effects models (such as the one used for our H1 test), can handle missing data e.g. by using maximum likelihood estimation, these estimations should also be taken cautiously (e.g. see this article: https://rpsychologist.com/lmm-slope-missingness, and the related PhD thesis: https://openarchive.ki.se/xmlui/handle/10616/46909). While we are not aware of a recommended threshold for which exclusion is recommended, missing more than ⅓ EMA measurements for any EMA phase (baseline, final or follow-up) seems a reasonable threshold to avoid misrepresenting our data. However, to reduce data loss, we raised this threshold from 4 to 5 missing EMA surveys, and specified this on page 17, line 3:

> *"A. they complete fewer than nine daily surveys out of 14 (i.e. they miss more than five surveys), in any of the baseline or final EMA behavioral tracking stages…"*

Regarding criterion B., failing to fill the psychological questionnaire will completely jeopardize the H2 and H3 tests, as we won't be able to measure the change in the psychological factors; therefore, it is inevitable to exclude the participants missing those questionnaires from our mediation analysis.

For criterion C., to decrease our conservativeness, we have increased this exclusion threshold to 2/6 meetings missed (on page 17, line 6). Missing more than two meetings out of six, i.e. more than 1/3 of the intervention content, may also jeopardize our hypotheses, which either directly or indirectly rely on intervention participation by the experimental group.

Thank you to the editor and all reviewers for these helpful comments.

The Authors