Manuscript number: https://doi.org/10.31234/osf.io/ycvxw

Article title: On the specifics of valuing effort: a developmental and a formalized perspective on preferences for mental and physical effort

Point-by-point reply

---------- comments Boris Cheval----------

I have now received two reviews of your manuscript *On the specifics of valuing effort: a developmental and a formalized perspective on preferences for mental and physical effort*, and I have read the paper carefully myself. As you can see from the notes below, the reviewers, who are highly qualified with respect to the topic of the paper, highlight particular strengths and note the importance of your work in helping us better understand some of the key questions revolving around whether effort valuations is domain-specific or domain-general. I fully agree with the reviewers that this is a very interesting question that has not been fully addressed in the current literature. However, they also raised some questions and concerns that need to be addressed before I can recommend your paper for PCI Health & Movement Science. I will not repeat all of the comments, but I agree with most of them. In particular, I concur with the following points that I believe need to be carefully addressed to improve the overall quality of the manuscript:

Dear Dr. Cheval, thank you very much for your feedback on our paper. Below, we detail how we have addressed the points that were raised by you and the reviewers. We are grateful to have received such constructive and helpful feedback and believe this has further strengthened the paper.

Add a definition of effort.

We fully agree that more information on how effort is conceptualized is important in order to avoid confusing matters here. We have thoroughly revised and extended this part, as we deem this a crucial aspect of effort research:

"Effort is instrumental for attaining goals [1–3], but the principle of least effort [1,4] posits that people try to avoid exerting effort if possible. This is because exerting effort feels aversive [5] and in turn carries an inherent cost [6,7], which is reflected in current definitions of effort (and the perception of it). Operationally, effort has been defined as a mediator between task characteristics and a person's task-specific capabilities on one side, and the achieved task performance on the other side [3]. To illustrate, say, a runner is asked to run five kilometers in 20 minutes (task characteristics) and has a personal record of 16 minutes over this distance (capabilities). Effort is what mediates between these characteristics and the achieved running time. Simply put, for a runner with a 16 minutes record, the task to run five kilometers in 20 minutes will probably require less effort than for a runner with a 19 minutes record. Focusing on effort's inherent costliness, some research emphasizes that effort relies on finite resources [8], whereas other research points towards functional processing constraints that make the exertion of effort costly [9]. Importantly, effort is not only conceptualized in terms of its objective properties but also with respect to how the exertion of effort feels [8,10]. Current definitions of this feeling also emphasize effort's costliness by referring to the perception of effort as the "instantaneous experience of investing resources"

([8], p7), and conceive perception of effort as a meter for the "momentary cost of effort investment" ([11], p3). This begs the question of how people choose to utilize this costly instrument. Contemporary theories of human motivation propose that humans weigh the costs of effort against the rewards associated with its exertion [12,13]. Simply put, if going for a run is instrumental for my goal to become fitter, the effort that is needed to actually go for the run is weighted against how rewarding it would be for me to reach my fitness goal. The same reasoning applies to effort that is exerted in the cognitive domain: If my goal is to learn Italian, then I weigh the efforts needed to master the language (i.e., memorizing new words and grammar rules) against how valuable it would be for me to actually speak Italian.**"**

(2) Justify some inferences (e.g., the relationship between engaging in more or less effortful activities and the cost of effort, the use of school grades and sports performance as proxies for actual behavior).

In the revised manuscript we have explicated and/or justified these inferences at the respective places. Please see our point-by-point responses below, where we respond to the corresponding comments by the reviewers.

(3) Explain why motivation to engage in the task and perceived task difficulty were not included, despite the current study's reliance on motivational intensity theory, in which these two variables are critical in explaining effort (dis)engagement. If relevant, discuss the potential implications for data interpretation of not including motivation and perceived difficulty.

In the revised manuscript, we now devote a paragraph to discuss the relationship between effort, task difficulty, and motivation. Here, we also discuss potential implications for this relationship if a person would truly value effort in its own right. We further acknowledge the limitation of having no measure of motivation included in this study. Importantly, as the paper is constructed through the lens of Learned Industriousness theory, our focus is on people's reporting of valuing effort per se and not geared towards testing Motivation Intensity Theory:

"If some people value effort regardless of its instrumentality as a means for goal attainment, this poses intriguing questions about the relationship between effort, motivation, and task difficulty. For example, Motivation Intensity Theory proposes that the amount of effort a person should mobilize towards a task is a function of task difficulty and potential motivation [2,14]. Accordingly, depending on how motivating a task is, people tend to mobilize more effort, while trying to conserve resources as to not overexert themselves. A large body of research is consistent with the propositions of Motivation Intensity Theory [14]. Intriguingly, if effort can be valuable in its own right, then this suggests that its mobilization is motivating too. In this scenario, not only the motivating downstream properties of what the task is allowing one to achieve (e.g., winning a race or bragging rights for training harder than everyone else) define how much effort one should mobilize, but the effort itself would contribute to the motivation as well. Consequently, task difficulty might not set the ultimate upper boundary for how much effort one should mobilize; rather, people might exert more effort than required. Evidence for such a decoupling of effort mobilization from outcome-specific motivation and task difficulty comes from research in the sports context showing that people overshoot required effort targets [2] or from athletes who overexert themselves in ways that are detrimental to their long-term performance (e.g., the phenomenon of overtraining). The interpretation that these are cases of people valuing effort – and in turn being motivated by its mobilization – is a highly speculative one. It is also plausible that other concurrent factors motivate, and thereby license, excessive effort mobilization. To better understand this, it is important to not only track preferences for

more or less effort, but to also assess the motivation to do so. Unfortunately, our study did not include a measure of motivation and further research is needed to address this limitation."

4) Review the data and the R script. This includes documentation of the data set (codebook).

We have thoroughly revised the R script, as well as the documentation of the analyses. Further, we have uploaded supplementary materials that contain further ancillary analyses to the OSF page.

- I wonder how the results of Study 1 and Study 2 can be integrated, since the samples are drastically different (e.g., differences in the mean scores on the scale and their reliabilities). Moreover, since the results of Study 2 are, in my opinion, rather inaccurate proxies for the valuation of effort (effort and performance can be drastically different, especially in math and sports), I wonder whether the second study adds more "value" than "harm" to the current article. But of course, this is a decision for the authors. I just raise this point from my outside perspective.

Thanks a lot for raising this important point. We feel that the consistency of the general results and interpretation across the two different studies gives more trust in the robustness of the general conclusions we are drawing from this paper. However, we agree that these methodological differences could also have the opposite effect. In the revised version, we are now more explicit about this conceptual consistency across these vastly different samples/measures, highlighting that very proximal (choices in the ring task) and also more inaccurate, watered down proxies (liking and performance of certain school topics) lead us to the same general conclusions. For example, in the discussion, we now write:

"From a methodological and a conceptual perspective, it is noteworthy how well aligned our different measures of the value of effort (questionnaires, decision tasks) and its behavioral consequences (boredom, grades) were. The choices in study 1 are a very proximal indicator of peoples' effort preferences, whereas topic-specific boredom and grades are much more distal proxies of one's effort preference. While the former might be considered as somewhat sterile and inconsequential (participants only indicated their preference for a hypothetical effort configuration), the latter might be perceived as rather inaccurate, since liking the effort to do something does not necessarily translate to performing it well. We believe that the consistency of the present results across the vastly different methodological approaches in study 1 and study 2 offers some interesting first implications for further directions on the generalization of effort across a wide range of effort valuation proxies. For example, preferences assessed with the VoPE scale and the NfC scale corresponded closely to the preferences for physical and mental effort determined in the Ring Task in Study 1. This provides a first insight into the validity of the Ring Task as a measure of the value of effort. Future research could therefore use the Ring Task to study preferences for effort using actual (rather than hypothetical) incentives by implementing participants' choices. For instance, participants could be told in advance that one of their decisions they make will be randomly selected and then implemented by giving them as task with the corresponding mental and physical demands. This would go beyond the focus on self-reported preferences in questionnaires. Second, preferences for cognitive versus physical effort were meaningfully associated with boredom and grades in Study 2. This indicates that effort preferences matter for decision-making and behavior in daily life already at a young age. From a more applied perspective, it could therefore be worthwhile to assess (and foster) young peoples' specific effort valuations. However, it is important to note that while the consistency in results across different methodological approaches offers some intriguing implications for our understanding of effort generalizations, further research needs to

test how effort valuations translate to different more or less direct proxies of effort preferences.″

- One last question/comment: Do you think that, all things being equal (i.e., the level of expected reward associated with the behavioral alternatives is identical or no reward is expected), people may choose to **repeatedly** engage in the more effortful behavioral option? I would appreciate the authors' thoughts on this comment.

This is a super interesting question. Based on LI theory, we would expect people to routinely prefer the effortful option if such effort preferences have been learned. However, as with any secondary reinforcer, the long-term persistence of this pattern is likely to depend on the (at least occasional) occurrence of rewards that are coupled with exertion of effort. However, if the effortful option continuously does not yield additional rewards, we would probably expect this secondary reinforcer to lose its power, no? So, it probably depends on the timescale of our analysis. This would be an interesting experiment to run. Especially with a focus on the gradients in effort preference building and then potentially disappearing over time, no?

---------- comments James Steele ----------

In general I found the manuscript to be very well written, succinct and clear. In some places though, **particularly in the methods**, I felt there could be **more detail provided** which could be **added to the supplementary materials** so as to keep the manuscript itself succinct. The title and abstract reflect the content of the article well. The introduction does a good job of presenting the theoretical rationale for the study and the research question is clear, though **you could more explicitly state you're a priori hypotheses,** though these are deducible to an extent from the introductory discussion of theory and prior empirical work. As noted, there are elements of the methods I think could be expanded upon for clarity and I note these more specifically below along with some questions from myself. The results are clearly reported and visually presented, though I will suggest some alternative approaches to presentation that might aid interpretation. The discussion is appropriate and clear, relating the findings back to prior hypotheses. The scope of the findings and limitations are also clearly discussed.

Thanks a lot for the positive assessment of our work and we greatly appreciate the suggestions for improving it. These were extremely helpful! Below, we outline how we addressed them.

Principle or law of least effort - In mentioning the principle/law of least effort I would try to be consistent in what word you use throughout. To be honest I would lean towards principle as the existence of something like the 'effort paradox' implies that simple effort minimisation is not law-like.

We agree that consistency in terminology is key, we follow your suggestion and refer to the principle of least effort throughout in the revised version of the manuscript.

Mental or cognitive - Minor point, and depending on how mental is defined might just be semantic, but 'cognitive' might be a better word... Bruya and Tang discuss this in their interpretive analysis of Kahneman's Attention and Effort - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6136270/ - In fact, cognitive might be better and align with your use of the Need for Cognition Scale.

This is a very good point, indeed. After looking more into the provided reference, we agree that cognitive might indeed be the better term here. We have updated this throughout the manuscript.

Effort definition – You don't define effort at any point during the manuscript. Given the variability of definitions across the literature it might be worth offering the definition you are employing in this work. Also, whether or not you are referring to effort or its phenomenology.

We fully agree that more information on how effort is conceptualized is important to avoid confusing matters here. We have thoroughly revised and extended this part, as we deem this a crucial aspect of effort research:

"Effort is instrumental for attaining goals [1–3], but the principle of least effort [1,4] posits that people try to avoid exerting effort if possible. This is because exerting effort feels aversive [5] and in turn carries an inherent cost [6,7], which is reflected in current definitions of effort (and the perception of it). Operationally, effort has been defined as a mediator between task characteristics and a person's task-specific capabilities on one side, and the achieved task performance on the other side [3]. To illustrate, say, a runner is asked to run five kilometers in 20 minutes (task characteristics) and has a personal record of 16 minutes over this distance (capabilities). Effort is what mediates between these characteristics and the achieved running time. Simply put, for a runner with a 16 minutes record, the task to run five kilometers in 20 minutes will probably require less effort than for a runner with a 19 minutes record. Focusing on effort's inherent costliness, some research emphasizes that effort relies on finite resources [8], whereas other research points towards functional processing constraints that make the exertion of effort costly [9]. Importantly, effort is not only conceptualized in terms of its objective properties but also with respect to how the exertion of effort feels [8,10]. Current definitions of this feeling also emphasize effort's costliness by referring to the perception of effort as the "instantaneous experience of investing resources" ([8], p7), and conceive perception of effort as a meter for the "momentary cost of effort investment" ([11], p3). This begs the question of how people choose to utilize this costly instrument. Contemporary theories of human motivation propose that humans weigh the costs of effort against the rewards associated with its exertion [12,13]. Simply put, if going for a run is instrumental for my goal to become fitter, the effort that is needed to actually go for the run is weighted against how rewarding it would be for me to reach my fitness goal. The same reasoning applies to effort that is exerted in the cognitive domain: If my goal is to learn Italian, then I weigh the efforts needed to master the language (i.e., memorizing new words and grammar rules) against how valuable it would be for me to actually speak Italian."

At the beginning of the 3rd paragraph of the introduction I would add "e.g.," to the examples in parentheses.

We have added this to the examples.

Sample size – As the study was not pre-registered for particular model specifications and to test certain hypotheses for particular parameters I would remove the following comment:

"The achieved sample sizes exceeds the required threshold for conducting the planned statistical procedures. For example, our sample is sufficiently large to provide high power to quantify significance of small effects in regression analyses."

I would instead be explicit in stating that the study was exploratory, though you may have had theoretically driven hypotheses to examine, and that sample size was necessarily resource constrained (either by time, funds, access to participants etc.)..

We agree that we could not specify a priori power that would be specific to test certain hypotheses with a particular model. We have followed your advice and adjusted this paragraph accordingly:

"With respect to sample size determination, we followed the recommendation by Funder and Ozer to aim for samples as large as project resources permit in order to obtain the most stable estimates [32]. This study was not pre-registered and therefore no power analysis for specific statistical models were performed a priori. However, as all analyses would be carried out within a regression analysis framework, we aimed for a sample size that is sufficiently large to provide high power to determine significance of small effects in regression analyses."

McDonald's omega – For those not familiar, I would specify that this is what is being reported when you refer to the different scales reliabilities. (Also it's just an observation, but I wonder why these scales haven't been explored in the context of Item Response Theory... for example, it would be interesting to explore the fit of the partial credit model to them)

We have replaced the Greek Omega with McDonalds Omega to avoid any confusion. With respect to IRT theory: Here, our focus was not on developing a new research instrument or diagnostic tool, for which an IRT analysis would indeed be a great approach. Rather, we focused on testing a set of conceptual predictions with our scales. We believe for this research question, an IRT analysis does not play a pivotal role. That being said, we agree that further research on the psychometric properties of the scales are worthwhile.

Study 1 multinomial model - I had a quick read of the papers cited as I am not wholly familiar with the ring task... It wasn't clear to me why participants should be classified and then this be the variable modelled - classification of a continuous variable inevitably loses information, some people might be just on the edge of a category whilst others are more clearly within it, so why not model the resultant angles directly? Or the coordinates for PE and ME? Some further justification of this choice might be useful (and perhaps exploring different model specifications and the sensitivity of your substantive conclusions to them as this is exploratory research). If you opt to stick with the multinomial model as your main one though I'd probably suggest reporting the results as probabilities for each category as opposed to odds ratios as everyone tends to find the latter less interpretable as far as I know. Or if you went for the continuous model of the angle you could report predicted angles and their CIs for varying levels of each predictor and then post modelling interpret what categories they tend to fall in.

This indeed raises an interesting point regarding the information that is contained in the angles. In the present study, we adhered to common standards of interpreting results from the ring task (https://doi.org/10.1177/1088868313501745), which uses the angle to classify participants into distinct categories. The reason is that the measure is multidimensional, such that lower or higher values of the angle have no unequivocal interpretation. To circumvent this problem in the domain of social preferences, decision tasks have been developed that provide unidimensional scale of SVO at the ratio level (https://psycnet.apa.org/doi/10.1017/S1930297500004204). This works

because people's social preferences can be efficiently assessed by only a small portion of the circle, on which lower and higher values of the angle directly translate into less versus more prosocial preferences. This is, however, not possible for our adaption because people preference configurations for mental versus physical effort are distributed across the entire circle.

With respect to probabilities vs. odds ratios: We have discussed this among ourselves and found the odds ratio (i.e., odds in relation to the chosen reference category) to be more informative in this context, as it expresses probably of category membership relative to the category that would reflect the principle of least effort. We have moved a sentence from a footnote to the main text now to make this more explicit in the Methods section. However, we are not married to this choice. So, if you and the editor both deem probabilities to be more apt in this context, then we are happy to adjust this in another revision.

"A multinomial regression predicts a nominal dependent variable with k levels through a series of k-1 dichotomous comparisons to a reference category. In our analyses, we chose "minimize PE + CE" as the reference category and compared it to the remaining seven configurations. Minimizing both types of effort was chosen as the reference category because it aligns with the presumed principle of least effort that is understood to mostly govern behavior[3]. The multinomial regression quantifies the likelihood of being categorized as not belonging to the reference category as a function of the predictor variables. For ease of interpretation, we use odds ratios to indicate how much the likelihood for not belonging to the reference category changed if VoPE and NfC values changed."

Dataset - Mind you, in thinking about the angle model I was trying to figure out how the pmeo variable in the dataset (which I assume is the angle from working through your analysis script) was calculated from the rest of the data but it appears this has been processed prior to uploading the data. I think the accompanying dataset may need a data dictionary to explain what each variable is to enhance reproducibility... also, for each of the presented 24 choices would be good to know so I would upload that to the supplementary material or add the PE and ME to each choice in the data. At the moment I assume it is just 1 or 2 depending on which option they chose.

We have added a codebook now to make it easier for readers to understand what each variable in the dataset means. In addition, we have uploaded a PDF to the OSF project where we have listed the 24 choices the participants faced.

Study 2 multivariate model - I got the impression that the grade outcomes, and the boring-ness responses, were separate substantive hypotheses. So, I don't think it is necessarily needed to model it as a multivariate outcome. It might be worth clarifying why this approach was taken by relating it back to your substantive hypotheses more clearly. Also, both outcomes seem to me to be ordinal in nature (I assume for the grade not really knowing whether any particular measurement models are employed for German school grades) and so an ordered logit or probit model might be more appropriate. You could again present predicted probabilities over the range of predictor values for each ordered category.

We have favored the multivariate route because it made the analysis a bit more parsimonious and we felt it would make it more justified to compare coefficients across dependent variables if one would like to. However, we agree that there is also a point in analyzing these research questions separately (and

frankly, the results and inferences are qualitatively the same). Regarding the ordinal nature of our outcome variables: This is a very good point and we agree that these variables might be also be adequately modelled as being ordinal in nature. We have performed additional ordered logit regressions, which are included in the updated R script on OSF. As they have not produced findings that are substantively different to the previous analyses, we refer to these analyses in a footnote of the paper (here, we mention that two coefficients became significant in the ordinal logit regression that were on the threshold before).

"To assess the link between VoPE and/or NfC with each of these variables, we conducted a multivariate linear regression analysis with VoPE and NfC as predictor variables. It can be argued that these dependent variables might be better modelled in an ordinal fashion. To account for this, we replicated the multivariate linear regression analyses with ordered logit regression analyses. As this did not meaningfully change results, we will not report these additional analyses in the results section but have included the respective R code along with the results into the uploaded data analysis script to the OSF page of this paper."

Descriptive statistics and t-test – I wonder about the appropriateness of the t-test between scales. It's not clear to me that these scales are necessarily comparable in anything other than an operational sense. Also, even if we grant that the operations are similarly capturing the value of effort in the same scale for each domain, the difference seems small even though significant. I am also not sure that you really need to report this given the correlations and network model.

We fully agree with this point and have removed the t-test. Indeed, it does not really add any meaningful insights in this context. We have however kept it in the R script and supplementary files for informational purposes. We note in the R script that this is not reported in the paper to not create confusion.

Limitations – To me it is a very important point when you note "… school grades are a proxy of actual behavior, as the effort invested into school tasks does not necessarily lead to better grade.". Some people may have low ability but still value the effort required for those tasks and so put in a lot yet still produce low grades. This links to some extent to the manner in which you conceptualise effort though, and perhaps a reason why offering your concept definition is valuable.

Yes, grades are a very distal proxy of the efforts students devote to school tasks. Along with providing a more thorough treatment of effort in the introduction, we now discuss in more detail that our studies capture the assumed domain-specificity of effort valuation in two ends of the proximal-distal continuum. With the choices in the Ring Task being a very proximal (albeit inconsequential) proxy and school grades/boredom a much more distal (and potentially more diluted) proxy. Through this lens, it appears particularly interesting how consistent our findings are across these two different samples. We discuss this – and the need to investigate the domain-specificity of effort valuations in further contexts – in more detail now:

"From a methodological and a conceptual perspective, it is noteworthy how well aligned our different measures of the value of effort (questionnaires, decision tasks) and its behavioral consequences (boredom, grades) were. The choices in study 1 are a very proximal indicator of peoples' effort preferences, whereas topic-specific boredom and grades are much more distal proxies of one's effort preference. While the former might be considered as somewhat sterile and inconsequential (participants only indicated their preference for a hypothetical effort configuration), the latter might be perceived as

rather inaccurate, since liking the effort to do something does not necessarily translate to performing it well. We believe that the consistency of the present results across the vastly different methodological approaches in study 1 and study 2 offers some interesting first implications for further directions on the generalization of effort across a wide range of effort valuation proxies. For example, preferences assessed with the VoPE scale and the NfC scale corresponded closely to the preferences for physical and mental effort determined in the Ring Task in Study 1. This provides a first insight into the validity of the Ring Task as a measure of the value of effort. Future research could therefore use the Ring Task to study preferences for effort using actual (rather than hypothetical) incentives by implementing participants' choices. For instance, participants could be told in advance that one of their decisions they make will be randomly selected and then implemented by giving them as task with the corresponding mental and physical demands. This would go beyond the focus on self-reported preferences in questionnaires. Second, preferences for cognitive versus physical effort were meaningfully associated with boredom and grades in Study 2. This indicates that effort preferences matter for decision-making and behavior in daily life already at a young age. From a more applied perspective, it could therefore be worthwhile to assess (and foster) young peoples' specific effort valuations. However, it is important to note that while the consistency in results across different methodological approaches offers some intriguing implications for our understanding of effort generalizations, further research needs to test how effort valuations translate to different more or less direct proxies of effort preferences."

Domain specificity of effort psychophysics? – A thought crossed my mind that I just thought I'd try to flesh out and share. Feel free to ignore or if you think relevant maybe discuss.

The interpretation of these results to some extent may differ depending on whether the psychophysics of effort is general or domain specific. If general (and assuming strong identity of phenomenology), then results such as yours suggest people may well value effort in different domains differently. But if the psychophysical relationship were domain specific also (differing in the strength of its identity perhaps across domains and at different levels of actual effort), then apparent differences in valuation might be confounded by this. For example, let's say that at an equivalent level of effort (I am assuming my own conceptualisation of actual effort here - https://psyarxiv.com/kbyhm) across both a physical and cognitive task is attempted but a person perceives it differently in either domain. Or that different tasks requiring different actual effort are perceived similarly effortful. It then becomes difficult to tease apart whether their valuation from a binary decision task is due to the differences in actual effort required, or what they perceive it to be. Of course, this may not be an issue for interpretation of your results specifically as you have used hypothetical tasks and we could perhaps assume that participants imagined tasks based on your instructions and thus remembered the associated phenomenology, or if they had not experienced such tasks they instead attempted to forecast what it might feel like. So, their valuation would be assumed to be based on how effortful they thought it might feel. This is where I like your suggestion of examining this with real tasks. In this it might be possible to try to present choices where the actual effort required for the tasks presented is known and on the same scales for physical and cognitive (though, whilst easy for physical it's much harder for cognitive as I have found e.g., https://psyarxiv.com/6pvht), and you could ask participants to both forecast how much effort they thought they would perceive the task would require and also ask them to report it when completing the chose task.

This is a very intriguing point (for which we have no satisfactory response at this point in time)! As you have emphasized, the real difficulty seems to lie in the challenge of objectifying the cognitive task demands in the actual effort they require. As you have mentioned, this is far from trivial. An additional point that we feel complicates these questions is that for many tasks we might

assume different fatigue and/or learning gradients across tasks, no? For example, during some cognitive effort tasks (e.g., a Stroop task) people might get progressively better as the task wears on due to learning and the actual effort required to do the task might go down. Complicating matters further, tasks that become easy (i.e., require less actual effort) might feel boring & then require additional effort to keep attention on track (for this argument see for example https://pubmed.ncbi.nlm.nih.gov/32697921/). Physical tasks that do require a constant external effort (let's say squeeze dynamometer at 100N) don't get easier over time, but rather feel the opposite. So, to make a long story short: we fully agree that there are some caveats when it comes to comparing effort costs and value across domains/tasks. We would probably avoid opening this topic in the current paper as we cannot really offer something to the discussion and our tasks were more about hypothetical efforts and less about actually completing the effortful tasks. However, going forward, these questions are certainly super interesting and we would be very interested to investigating them further (e.g., how to construct tasks of equivalent and predictable effort requirements? How to have tasks with similar effort-dependent temporal gradients? Etc.)

Code – I would double check your code script as I noticed at least one typo (line 101). I also can't seem to reproduce your network graphs (I seem to also get three communities for study 1). There are also some arguments that seem unnecessary in certain parts (e.g., in the raincloud plot scale_X_discrete colors).

We have checked and improved our code for typos. Thank you very much for pointing these out! Further, we added a codebook in OSF, where you can find each variable, its meaning, valid values, and the value labels. Regarding the network plot: yes, this produces three communities for both studies. Importantly, these two communities consist of items from the NfC scale (primarily reflecting positive/negative wording). We now make this clearer in the Descriptive Statistics section & the Figure caption (We have also adjusted the color in the Figure to make the different communities more visible):

"To further investigate if NfC and VoPE scales represent differentiable constructs, we conducted exploratory graph analyses. For both studies, the exploratory graph analysis for the VoPE and NfC items formed three clusters (Figure 2c and 2d): One VoPE community and two NfC communities. (With the exception of two items in study 2, community membership for the NfC items could be fully explained by the direction of item wording.) Bootstrapping indicated that the communities are stable (1000 iterations), as all items were assigned to the same communities in 100 % of the cases for study 1. For study 2, the VoPE items were 100 % stable, whereas the NfC items showed a less stable community assignment (range between 50 % to 85 %)."

"Figure 2. Panel A (study 1) and B (study 2) show scatterplots depicting the relationship between the VoPE and the NfC scores. Panel C (study 1) and D (study 2) depict the estimated networks for the VoPE and NfC items. For study 1 the answer scales ranged from 1-7, whereas for study 2, they ranged from 1-5. Complementing the raw data, the regression line along with the 95% confidence interval is included into Panel A and B. Blue community = Value of Physical Effort Scale, red and yellow communities = Need for Cognition Scale. The red and yellow communities primarily differ with respect to positive and negative item wording."

---------- comments Reviewer 2 ----------

In this article, the authors explored the preference individuals have for exerting mental and/or physical effort and the potential impact this can have on their performance (as reflected in academic results), and the boredom experienced in these two types of activities. To do this, they conducted two studies. The first was conducted online and aimed to demonstrate that there is a distinction in preference between mental and physical effort. This was achieved by using the ring task and two scales. Following this initial study, the authors applied their findings to a real-world context: school. Students filled out these scales, and correlations were drawn with their performance in mathematics and physical education (serving to establish a connection between mental and physical effort). They also reported their usual level of boredom in these two academic activities.
I truly enjoyed reading this article. It provides a missing piece of information in the literature, namely that the two types of efforts can be perceived and evaluated differently depending on the individual. The two studies presented in this article are an important first step toward delving deeper into the exploration of this dichotomy. I also greatly appreciated the outlined limitations of these two studies, showing that the authors do not make claims beyond what the data indicates. This is pleasing.
However, I have several comments to make regarding this article. You will find them below, with the first section containing general comments, the second section containing specific comments, and the final section including a few remarks concerning the data and the R script.

Reply: Thanks a lot for the very positive assessment of our work and the constructive comments for further improving it. This was very helpful in our revision. Below, we outline how we have addressed each point.

1) I am somewhat surprised not to see a measure of motivation, particularly in the second study. Indeed, a person is more inclined to invest effort when the motivation to perform the task is high. On page 15, the authors mention a grain that could be a bit finer to measure the value of effort in more specific domains. Does this not refer to the motivation to engage in this activity? Aren't preference for an activity and the motivation to undertake it two sides of the same coin?
I noticed that the authors used the Motivational Intensity Theory in their reasoning (references 1 and 9). The Motivational Intensity Theory (Richter et al., 2016) suggests that potential motivation is the limit of the effort one invests in a task. Thus, effort can only be invested when there is a motivation to exert it. In the same vein, the authors' first sentence in the abstract is that "Effort is instrumental for goal pursuit". This indicates that there is motivation involved in achieving a goal. I understand that these two studies were not designed to measure motivation, however, I believe that a paragraph in the discussion, in the implications section for instance, would provide significant additional insight. Another option could be to discuss this point in the limitations / perspectives section as it was not measured, but is highly relevant for future research.

We fully agree with this point! Unfortunately, no measure of motivation was included as the research was conducted with a rather exclusive focus on effort. However, we completely agree that effort and motivation are very closely related and understanding this relationship is a highly relevant and fascinating question. We have followed your advice and discuss this question, along with the limitation of not having a measure for motivation in the present study, in the Implications section:

"If some people value effort regardless of its instrumentality as a means for goal attainment, this poses intriguing questions about the relationship between effort,

motivation, and task difficulty. For example, Motivation Intensity Theory proposes that the amount of effort a person should mobilize towards a task is a function of task difficulty and potential motivation [2,14]. Accordingly, depending on how motivating a task is, people tend to mobilize more effort, while trying to conserve resources as to not overexert themselves. A large body of research is consistent with the propositions of Motivation Intensity Theory [14]. Intriguingly, if effort can be valuable in its own right, then this suggests that its mobilization is motivating too. In this scenario, not only the motivating downstream properties of what the task is allowing one to achieve (e.g., winning a race or bragging rights for training harder than everyone else) define how much effort one should mobilize, but the effort itself would contribute to the motivation as well. Consequently, task difficulty might not set the ultimate upper boundary for how much effort one should mobilize; rather, people might exert more effort than required. Evidence for such a decoupling of effort mobilization from outcome-specific motivation and task difficulty comes from research in the sports context showing that people overshoot required effort targets [2] or from athletes who overexert themselves in ways that are detrimental to their long-term performance (e.g., the phenomenon of overtraining). The interpretation that these are cases of people valuing effort – and in turn being motivated by its mobilization – is a highly speculative one. It is also plausible that other concurrent factors motivate, and thereby license, excessive effort mobilization. To better understand this, it is important to not only track preferences for more or less effort, but to also assess the motivation to do so. Unfortunately, our study did not include a measure of motivation and further research is needed to address this limitation."

2) Following on from the first comment, the Motivational Intensity Theory (Richter et al., 2016) predicts that the effort invested in a task is a function of the perceived difficulty of that task, with potential motivation being the limit (the maximum one invests). How can this statement be reconciled with the fact that here there is a preference for investing effort in one domain rather than another? Is the perceived difficulty different?

This a very interesting point. As we study effort through the lens of Learned Industriousness theory in this paper, we are not equipped to truly asses how the present findings relate to the propositions of MIT. In the discussion, we now devote one paragraph to touch on the interplay between effort, motivation & task difficulty and discuss how "truly" valuing effort might lead to some counter-intuitive outcomes (see also comment above). Through a Learned Industriousness lens, we would think that the perceived difficulty does not need to be different to prefer one task over another (in the sense that if people really learn to value effort exertion irrespective of outcome, then perceived difficulty should – within a reasonable band – not deter the person from it). However, this is a really interesting research question. If we would translate the ring task to the lab, we could ask people also for the perceived difficulty of their choices. This would be a super interesting line of future research. However: in lay reasoning of people and in the way the questions of the questionnaires are phrased, difficulty and effort are likely to be tightly coupled.

"If some people value effort regardless of its instrumentality as a means for goal attainment, this poses intriguing questions about the relationship between effort, motivation, and task difficulty. For example, Motivation Intensity Theory proposes that the amount of effort a person should mobilize towards a task is a function of task difficulty and potential motivation [2,14]. Accordingly, depending on how motivating a task is, people tend to mobilize more effort, while trying to conserve resources as to not overexert themselves. A large body of research is consistent with the propositions of Motivation Intensity Theory [14]. Intriguingly, if effort can be valuable in its own right, then this suggests that its mobilization is motivating too. In this scenario, not only the motivating downstream properties of what the task is allowing one to achieve (e.g.,

winning a race or bragging rights for training harder than everyone else) define how much effort one should mobilize, but the effort itself would contribute to the motivation as well. Consequently, task difficulty might not set the ultimate upper boundary for how much effort one should mobilize; rather, people might exert more effort than required. Evidence for such a decoupling of effort mobilization from outcome-specific motivation and task difficulty comes from research in the sports context showing that people overshoot required effort targets [2] or from athletes who overexert themselves in ways that are detrimental to their long-term performance (e.g., the phenomenon of overtraining). The interpretation that these are cases of people valuing effort – and in turn being motivated by its mobilization – is a highly speculative one. It is also plausible that other concurrent factors motivate, and thereby license, excessive effort mobilization. To better understand this, it is important to not only track preferences for more or less effort, but to also assess the motivation to do so. Unfortunately, our study did not include a measure of motivation and further research is needed to address this limitation."

1) Page 4: An inference is made between the preference for engaging in more or less physical/mental effort and the cost of that same effort ("a preference for less mental and/or physical effort will be a proxy that the corresponding effort is costly…"). This is a logical inference made here and at first glance seems valid. However, to the best of my knowledge, the link between the two has not been demonstrated, and no citation is given for this sentence. Therefore, this inference should be included in the limitations of the article since the connection between the two does not seem to be demonstrated at the current time, or a reference should be added to support this statement

This statement indicates that humans and no-human animals try to minimize efforts because it is costly and tend to pick – all things being equal – the tasks that minimize effort. Now, we agree that seeking out less mental/physical effort might not map one to one on its perceived costliness as other factors might be at play here too. We have now slightly rewritten this sentence, to a) provide a reference for the fundamental observation that an organisms' choices can often be explained in terms of cost minimization (we use Tsai's often overlooked work as a reference for this), and b) made more clear that this might not be a perfect proxy nevertheless:

"In a first study, we investigate people's general preferences for the allocation of physical and cognitive effort. In line with classic research on the principle of least effort[1], a preference for less cognitive and/or physical effort will be interpreted as a rough proxy that the corresponding effort is costly. A preference for more cognitive and/or physical effort will be interpreted as a rough proxy that the corresponding effort is valuable."

2) Page 5: The authors chose to conduct the first study online via MTurk. The justification for this choice is not provided in the body of the text. Is it to have more participants more quickly? There are also other platforms besides MTurk for conducting online studies (e.g., Prolific, CrowdFlower, etc.). Some of them seem to yield higher-quality data (Peer et al., 2017, 2021). Can the authors justify the choice of MTurk? The study here is relatively simple and therefore probably minimally impacted by the quality of the respondents, especially with the verification questions posed to participants.

We have now added a sentence that we used MTurk to gain quick access to a reasonably large US sample.

"We used MTurk, as it allowed us to optimize project resources and sample a large enough US sample withing a relatively short timeframe."

3) Page 5, it is indicated that there are 37% female in the first study, and 62% in the second. Please also indicate the number of individuals who chose not to respond, as well as the number of "other" gender responses (1 in the first study I believe, and several in the second). Also, indicate the rate of missing data. I also wanted to know if it was sex at birth or gender that was asked of the participant. I assume it is the gender with which the participants identify, given the dataset and the statement on pages 6 and 7 about additional measures. If this is the case, the word "female" (sex) should be replaced with "women" (gender).

We have added the requested additional information to the paper. Regrading the Sex/Gender question: We asked the students about their gender and they responded with male/female. In the online study, we asked for gender and response options were male/female. So, we were not consistent in our terminology here. However, participants did not report any issues with the reporting format. We would therefore keep this description as it is, in order to be fully transparent about what we asked and what answering options were provided. To acknowledge this inconsistency, we have added a footnote:

"Please note, that in the questionnaire we asked for gender but provided the answering options female/male/other/prefer not to say."

4) Page 5: it is mentioned that the VoPE scale contains 4 items. In the dataset available on OSF, I can see 10 columns named vope or voes. Is this the same scale? Why is there a discrepancy between the 4 mentioned in the text and the 10 present in the dataset? If it is indeed the same scale, why administer all 10 items of the scale instead of just the 4 of interest?

Well spotted. We initially had more items developed for this scale but the final version of the VoPE consisted of four items. As our study was setup when VoPE was still in development, we chose to include all items that were initially developed. We have added this information now to the R-script on the OSF.

5) Page 5, following up on the previous question: I am wondering about the difference in the number of items between the two scales, particularly when it comes to the clusters that can be found later in the article. For the NfC scale, the authors observe two clusters, but only one for the VoPE. Since the VoPE only has 4 items (used), it is more challenging to find multiple clusters. Isn't this a limitation to the cluster analysis? Without a counterargument, I believe that this is a limitation to be noticed.

Good question! No, the number of communities is not necessarily driven by the number of available items. The number of communities one finds will depend on the (dis)similarity of the items. So, we could in theory have four communities with four vastly different items and one community with 40 very similar items.

6) Page 6 and 7: The authors explain on page 6 that the middle of the scale is 50 (and therefore the center of the circle has coordinates (50,50)). I think it would be interesting for the reader to see the value of 50 appear on both the x-axis and y-axis in Figures 1B and 3.

The description of the coordinates only serves to explain how we computed the angles for categorizing participants. However, the histogram is independent of this (arbitrary) choice. We would therefore prefer to not add these coordinates as an axis label, as this might be misleading.

7) Page 7: Figure 1A is too small; once printed, it is no longer legible. Please enlarge it (perhaps by rotating it 90° and extending it across the width of the page).

Indeed! We have enhanced its quality and size as much as possible now. In addition, we made a note in the Figure caption that a full size Version of Figure 1A is on the OSF page.

"Figure 1. Panel A shows the instructions for the Ring Task with an example task (For a full-page view of Panel A, please go to https://osf.io/uz7gt. Panel B shows the eight idealized effort preferences in a coordinate system. Note. PE = physical effort, CE = cognitive effort."

8) Page 8: It is not indicated whether the authors checked for outliers. I assume there are none in this case. Can they confirm this?

We checked our data for outliers using the interquartile range criterion and found no outliers in the data of study1 for the VOPE and the NFC scale and the data of study2 for the VOPE scale. However, we found three outliers in study2 for the NFC scale. Redoing the analyses for study2 without these outliers did not change our main findings. It has to be noted that we would observe one additional significant effect when redoing the analysis without these three statistical outliers: here, boredom in sports is now predicted not only by VoPE but also by NfC. However, we would be cautious to interpret this additional effect, because it does not appear to be very robust if its appearance depends on these three cases. For full transparency, we have included these additional analyses as a supplementary information to the papers OSF page (SI 5).

9) Page 8: In the dataset for study 2, I noticed that some data were missing for certain participants. This is typically observed when working with children, not an issue for me. This information is indirectly reported on page 8 during the t-test between NfC and VoPE with n = 284. However, I believe this information should appear somewhere, possibly detailed in supplementary data.

We have added a detailed list of the variables in the supplementary materials. Here, we also detail how many participants responded to each variable. This allows for maximum transparency.

10) Page 8: Could you please add the effect sizes of the t-tests? This allows researchers who wish to conduct meta-analyses to do so more quickly, in addition to indicating the strength of the difference to the readers.

Yes, in the initial version of the manuscript the effect sizes were indeed missing. Please note that we have now moved this part to supplementary materials as we agree with R1 that the comparison of the scale values is not very informative.

11) Page 8: The authors state that the correlation between VoPE and NfC is weak and therefore physical effort is, on average, evaluated differently from mental effort. I understand what the authors are trying to convey. However, since the correlations are significant, even if it's weak, it means that a higher score on one scale is associated with a higher score on the other. This sentence needs to be revised. Moreover, this raises the question of the threshold from which the authors would indicate that the two scales vary simultaneously (e.g., r = .20 or r = .30). This value would inevitably be arbitrary.

Good point. We have now rephrased this a bit, focusing more on the amount of shared variance between the concepts:

"In both studies, the VoPE and NfC scale were weakly correlated, r = .18, p = .002 (study 1), r = .13, p = .030 (study 2). Thus, although higher VoPE scores were associated with somewhat higher NfC scores, the shared variance between both concepts is < 5% in both samples, indicating that physical effort is valued differently to cognitive effort (Figure 2a and 2b)."

12) Page 9: Is it possible to create clusters with more distinguishable colors? It is mentioned on page 8 that there are three clusters, two of which come from the NfC scale. However, the color does not allow for easy distinction between the two clusters, especially in the case of study 2. Finally, regarding the graph, the scale is referred to as NfC in the body of the text, but as ndfc in the figure. I suggest that the authors harmonize the two or explain the difference.

We have adjusted the Figure accordingly.

13) Page 10: Figure 3. This is just a suggestion. While reading (on paper), I wrote the percentages indicated in the text on the figure to better remember them. It might be interesting to include them directly for the reader, for instance in the blue part of the figure.

Great idea. We have added this to the Figure!

14) Page 11: The authors mention that they measured gender, but it is not indicated whether this had an impact on the results. Does gender have an effect on the correlations or on the perceived boredom? Moreover, the chosen school subjects (mathematics and sports) are seen as more masculine or at least preferred by men (Eccles et al., 1993). Could this have an impact?

We re-ran the analyses and checked the effect of gender on our results. We have added this to the supplementary information (SI 4). No effects of gender were found.

15) Page 13 at the top: The score on the VoPE also predicts boredom in mathematics, even though this is to a lesser extent. Given the huge expertise in boredom of some authors of this article, it might be interesting to mention and briefly discuss this. It would also be interesting to put these results in perspective with those of study 1, particularly with figure 3 where we can see that there are very few people who only maximize physical effort, but those who do maximize this effort also seem to maximize mental effort.

Thanks a lot for this comment! This indeed is an interesting observation! Based on our data, it is difficult to say whether this finding reflects a co-occurrence of the value of physical and cognitive effort (as you suggest) or maybe the presence of a domain-general trait boredom. Or something else, as both interpretations would suggest a small correlation between NfC and sports-related boredom as well. Given that we would have to speculate on quite shaky grounds, we would rather refrain from including these considerations in the present manuscript. It would be interesting to test these different interpretations against each other in future research though!

16) Page 13: "Accordingly, researchers should be specific about the kind of effort they are addressing in their research". I couldn't be more in agreement with this statement!! This sentence could maybe be in a better position in the conclusion as a final take-home message.

Good point! We have moved this sentence into the Conclusions section:

"Taken together, the present research highlights important differences in how people value cognitive and physical effort, that the value assigned to effort in a specific domain (i.e., cognitive or physical) maps directly onto choices, experiences, and outcomes in activities that require this type of effort. Accordingly, researchers should be specific about the kind of effort they are addressing in their research."

1) In the data from study 2, I noticed two errors that seem to be a typo. Could you please check the other data and redo the analyses correcting for these errors? I imagine that this should not fundamentally change the results since only two data points appear to be erroneous.

Thank you for noticing and mentioning. We looked up the values from the original paper-pencil version and changed accordingly. The corrected values did not cause any major changes in our results.

Participant 83 has a value of 6 for item 1 of the NfC scale, while the German version of this scale indicates on page 7 that the scale ranges from 1 to 5.
Participant 78 has a value of -5 for item 4 of the NfC scale.

Thank you for noticing and mentioning. We looked up the values from the original paper-pencil version and changed accordingly. The corrected values did not cause any major changes in our results.

2) A read-me file should be added with the data to explain the variable names. The data files are currently hard to understand, making it difficult to replicate the analyses. For instance, I do not understand what the columns pmeo, gspa, or gsma represent.

We have now added a codebook for all variables to the OSF folder.

3) Page 5: Regarding the age, it is correctly indicated in the article, but in the provided R script, it says "strange answers", and the results for nationality and level are not indicated either. This information is just for you, to harmonize.

We have updated this accordingly. Thank you!

4) Page 5: It is mentioned that the materials, including the questionnaires, are available on OSF, however, I could not find the file containing the questionnaires. Could you please add it, or remove the private option on OSF as the sentence indicates that they are available? If not, please adjust the statement.

We have now uploaded PDFs for the questionnaires from both studies.