Dear Chris,

Thank you very much for your swift action on our submission and for inviting us to respond to the reviewer's comments. We agree with R1 that more must be said about statistical power consideration and sample size planning, although this reviewer seems to have confused our power calculations for the pilot study with that for the proposed fMRI study. We disagree with R2 that our research design is inconclusive with respect to our research question, and we clarified the implementation of experimental control conditions. On the following pages you can find our replies to the pasted reviewer's comments in the order in which they appeared in each review. Changes made to the manuscript are highlighted using the MS track changes function.

Reviewer 1

The study addresses a valid scientific question, and the utilization of the RR publication route, especially for an fMRI study, should be commended. I would also like to emphasize that the authors invested important resources to run a pilot study. The logic behind the main functional hypothesis is strong and the methods appropriate to address it. Yet, my opinion is that the study does not fulfill the minimal standards for registered report regarding outcome neutral conditions, degree of details in data (pre)processing, the power analyses and the link between the hypothesis and the statistical contrasts used to test them. I urge the authors to check previously published (fMRI or behavioral) RR to better figure out the level of details required by this publication format, and the RR guidelines of some journals regarding the need for positive controls and power levels.

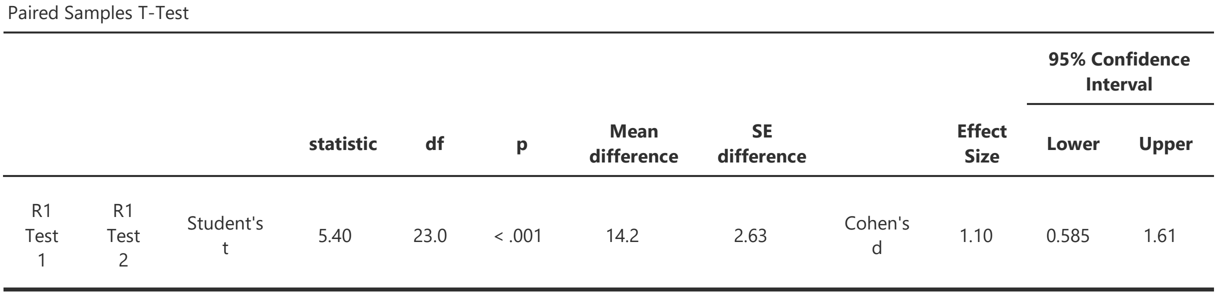
We thank the reviewer for the appreciation of the efforts that we made with preparation of this preregistration. This was our first submission of a peer-reviewed preregistration and we prepared it according to the guidelines of PCI:RR. We did not read the author guidelines of each associated journal that are often inconsistent, for example, with respect to target levels of statistical power, importance of pilot testing, and templates for documentation. We are also not aware of a preregistered fMRI study in our field of research that we could use as a template. In fact, preregistration of fMRI research is still very rare; consensual preregistration guidelines for MRI are not available; and established procedures for a preregistration of behavioural studies cannot be simply transferred to that for fMRI studies for a number of reasons (for a discussion see Flannery, 2018). We trust that our research proposal could still be recommended at PCI:RR, even if it does not meet the standards of each associated journal. We are however aware that failing journal-specific standards forfeits guaranteed publication in specific PCI-friendly journals.

- It is not perfectly clear which statistical designs were used for the power calculation (and throughout the study). Please include an exhaustive table listing all statistical designs and terms of interest and the related power analysis. This table should also allow to verify that the statistical design from the source study (e.g. Eder & Dignath (2016)) is exactly the same as in the present study. The parameter of the power analysis should be reported in full details to ensure its reproducibility (G*Power parameters). A power calculation should be conducted for each terms of interest, no statistical test should be conducted if not previously listed in the power analysis (and thus adequately powered).

The proposed fMRI study (and the pilot study) has a 2 (*Transfer Test*: before devaluation vs after devaluation) x 4 (*Pavlovian Relation*: CS1/Currency 1 *vs* CS2/Currency 2 *vs* CS3/Currency 3 *vs* CS-/no currency) x 3 (*Instrumental Relation*: R1/Currency 1 vs R2/Currency 2 vs R3/Currency 3) repeated-measures design. Due to the complexity of our research hypotheses (2-way and 3-way interaction effects in the omnibus analysis), we planned follow-up comparisons of statistically significant effects with separate ANOVAs and t-tests.

There are four statistical effects that are of particular theoretical relevance for the present research question. In the following, we will describe each planned statistical test and the corresponding statistical result obtained from our pilot study. The statistical effect of the pilot study was used as effect size estimate in a-priori power analyses for the fMRI study. For full transparency, we inserted snapshots of our GPower3.1.9.7 power calculations below.

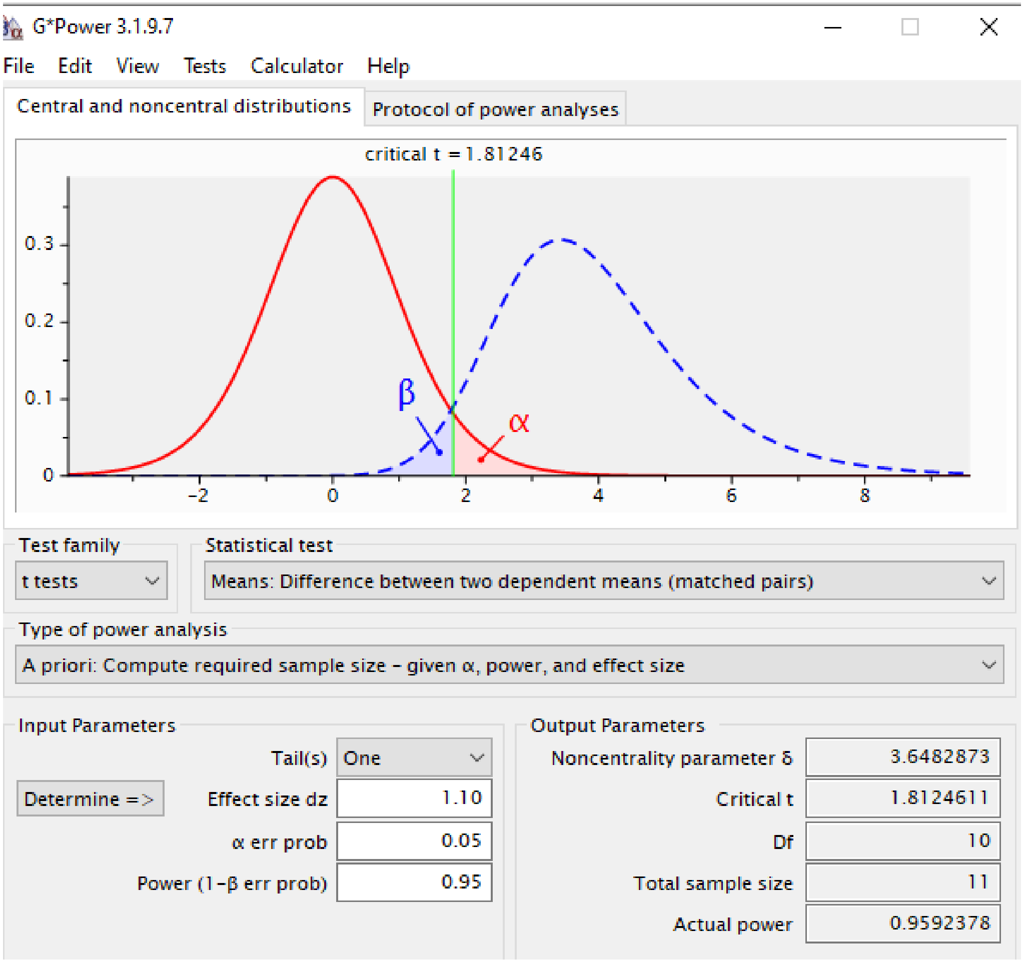
(1) **Effect of devaluation treatment**: If the devaluation treatment was effective, rate of working for the now-devalued outcome (R1) should be significantly lower in Transfer Test 2 (after devaluation) compared to Transfer Test 1 (before devaluation). In statistical terms, this means that numbers of keypresses (R1) during presentations of the neutral cue (CS-) is lower in Transfer Test 2 than in Transfer Test 1. This is indicated by a significant effect of CS- on R1 responding in a univariate ANOVA, which is identical with a paired t-test for R1 responding in Transfer Test 1 and 2. This is the result of the comparison in the pilot study:



Paired Samples T-Test

										95% Confidence Interval	
			statistic	df	p	Mean difference	SE difference		Effect Size	Lower	Upper
R1 Test 1	R1 Test 2	Student's t	5.40	23.0	< .001	14.2	2.63	Cohen's d	1.10	0.585	1.61

With this effect size estimate, G*Power gives following output:



Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size dz: 1.10

α err prob: 0.05

Power (1−β err prob): 0.95

Output Parameters

Noncentrality parameter δ: 3.6482873

Critical t: 1.8124611

Df: 10

Total sample size: 11

Actual power: 0.9592378

Thus, a minimum sample size of n = 11 would be needed to detect this effect.

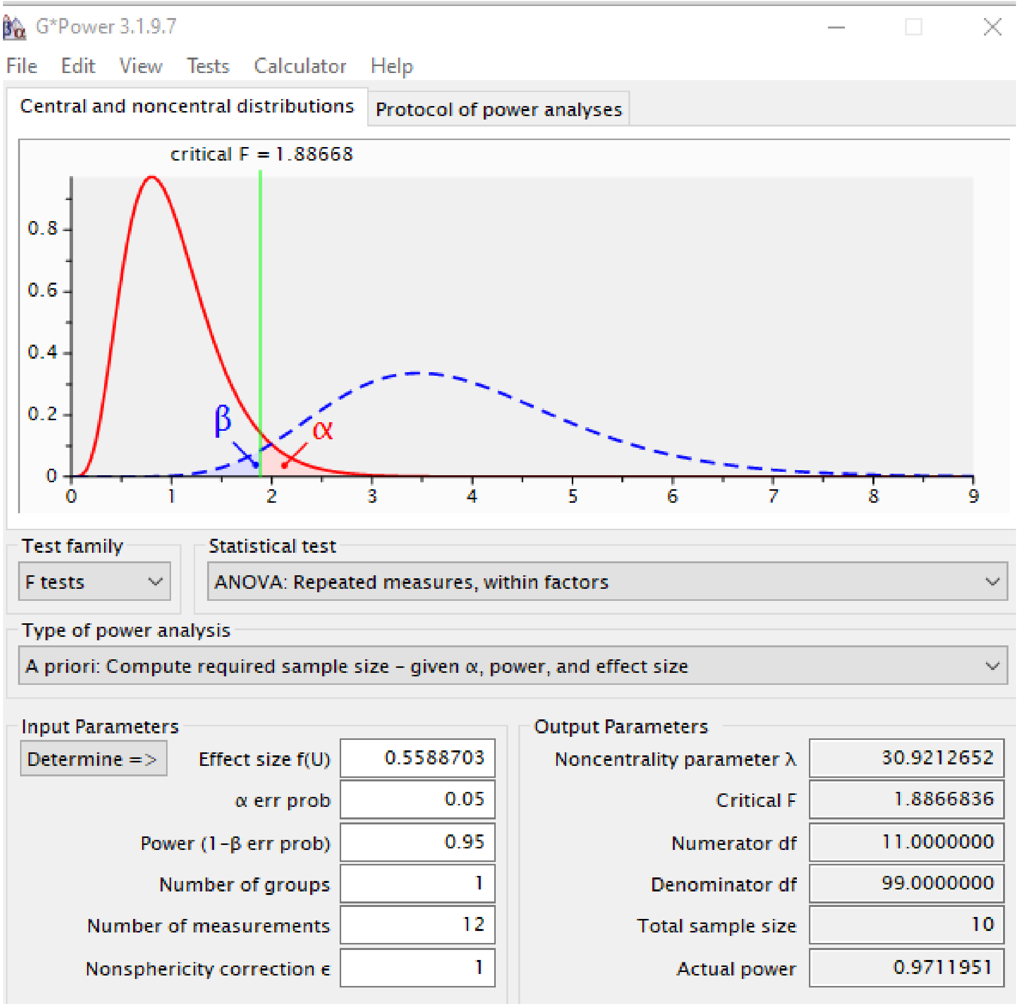
(2) **Outcome-specific PIT effect in Transfer Test 1 (before devaluation)**: Pavlovian cues (CS1, CS2, CS3) should specifically increase numbers of keypresses that were associated with the same outcome (R1, R2, R3) relative to the baseline condition (with CS- presentations). Statistically, this is expressed in a significant 2-way interaction effect between Pavlovian Cue (CS1, CS2, CS3, CS-) and Instrumental Relation (R1, R2, R3) (highlighted in red color in the table below).

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Cue	759	3	252.9	4.82	0.004	0.173
Residual	3623	69	52.5			
Response	1208	2	604.2	12.07	< .001	0.344
Residual	2302	46	50.1			
Cue ✻ Response	9753	6	1625.5	7.18	< .001	0.238
Residual	31236	138	226.3			

Note. Type 3 Sums of Squares

With this effect size estimate, G*Power creates the following output:



Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f(U): 0.5588703

α err prob: 0.05

Power (1–β err prob): 0.95

Number of groups: 1

Number of measurements: 12

Nonsphericity correction ϵ: 1

Output Parameters

Noncentrality parameter λ: 30.9212652

Critical F: 1.8866836

Numerator df: 11.0000000

Denominator df: 99.0000000

Total sample size: 10

Actual power: 0.9711951

Thus, a minimum sample size of n = 10 would be needed to detect this effect.

(3) **Outcome-specific PIT effect in Transfer Test 2 (after devaluation)**: Non-devalued Pavlovian cues (CS2, CS3) should still increase keypresses that were associated with the same outcome (R2, R3) relative to the baseline condition (CS-). Thus, the same ANOVA as in (2) is performed but this time without inclusion of the now-devalued CS1 and R1 (highlighted in red color).

Within Subjects Effects

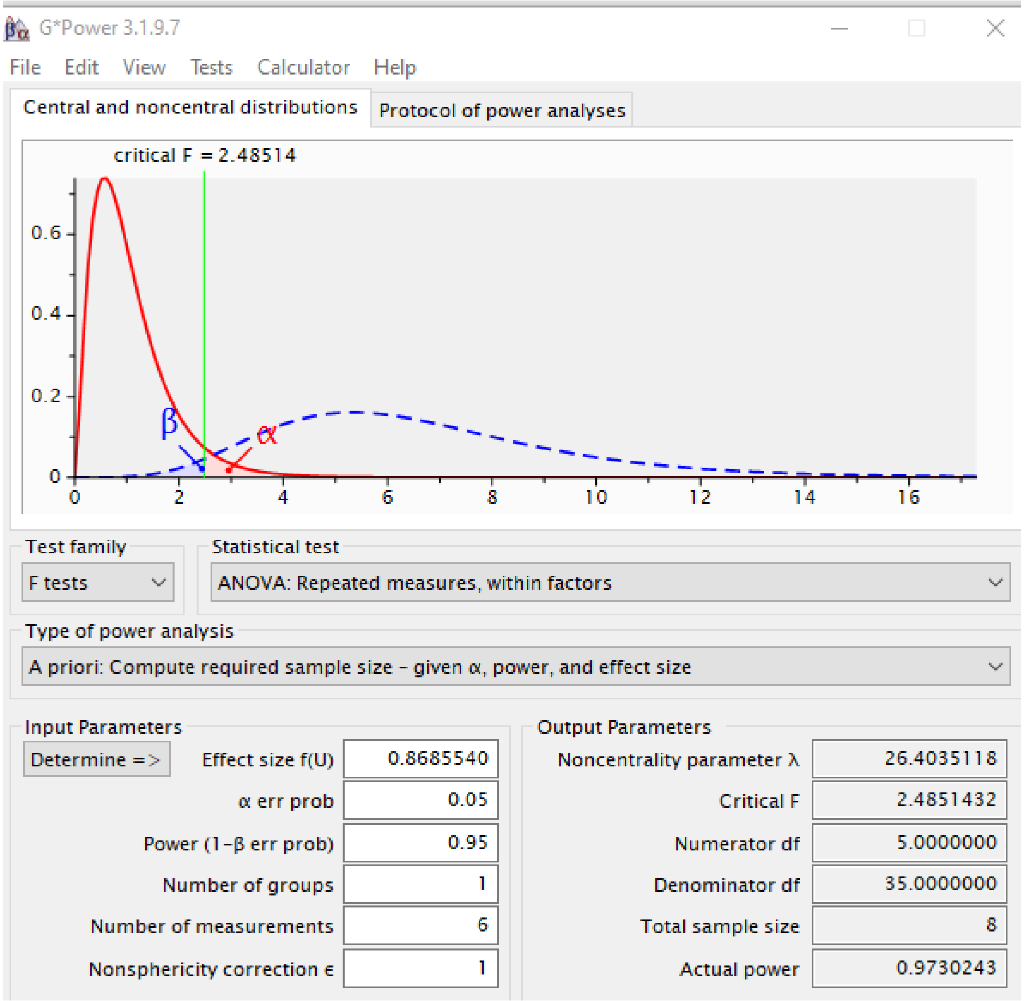
	Sum of Squares	df	Mean Square	F	p	η^2_p
Cue	1160	2	580.2	5.58	0.007	0.195
Residual	4783	46	104.0			
Response	685	1	684.7	7.81	0.010	0.253
Residual	2018	23	87.7			
Cue ✻ Response	8960	2	4479.8	17.35	< .001	0.430
Residual	11875	46	258.1			

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p

Note. Type 3 Sums of Squares

With this effect size estimate, G*Power creates the following output:


G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 2.48514
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
β
α
Determine =>

Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f(U)	0.8685540
α err prob	0.05
Power (1-β err prob)	0.95
Number of groups	1
Number of measurements	6
Nonsphericity correction ε	1

Output Parameters

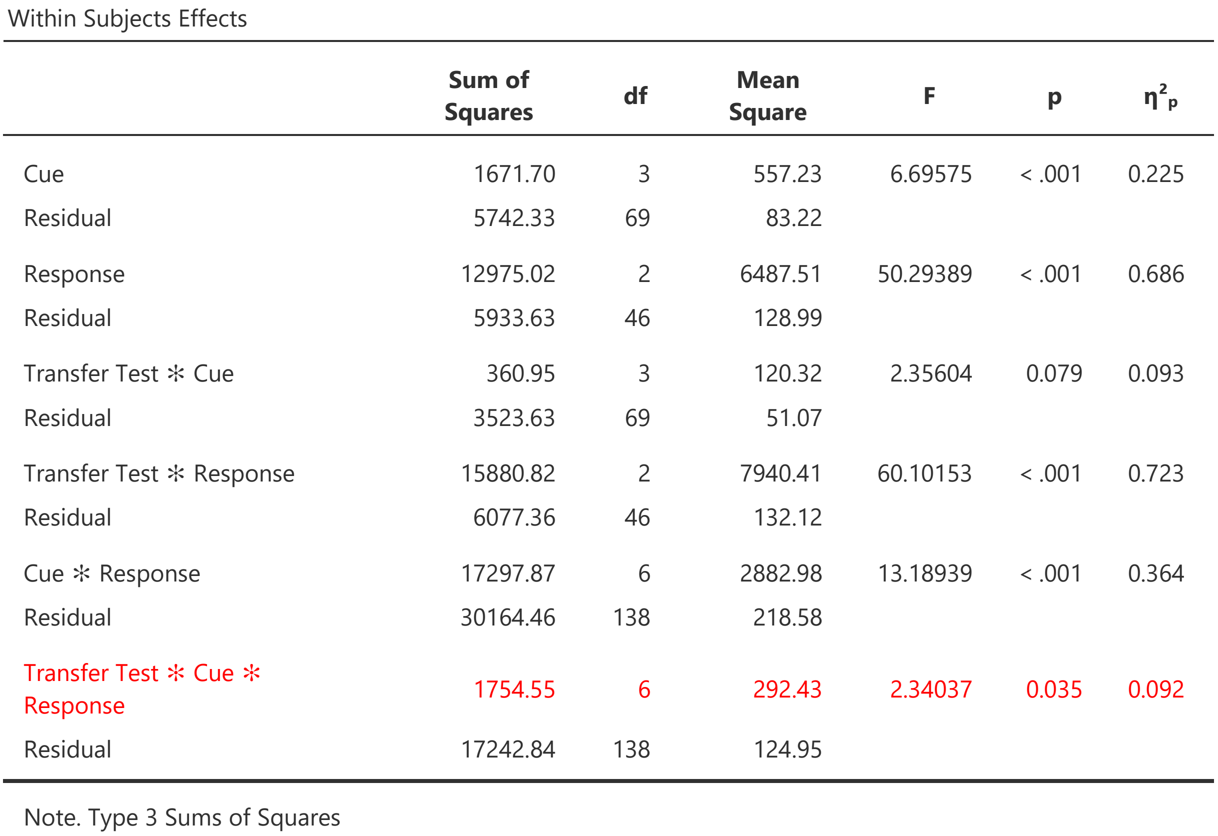
Noncentrality parameter λ	26.4035118
Critical F	2.4851432
Numerator df	5.0000000
Denominator df	35.0000000
Total sample size	8
Actual power	0.9730243

A minimum sample size of n = 8 would be needed to detect this effect.

(4) **Reduced PIT effect after devaluation of the associated outcome:** Statistically, this corresponds with a 3-way interaction effect between Pavlovian Cue, Instrumental Relation, and Transfer Test in the omnibus ANOVA. In the pilot study, this 3-way interaction effect was significant.

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Transfer Test	2.26	1	2.26	0.00795	0.930	0.000
Residual	6538.95	23	284.30			

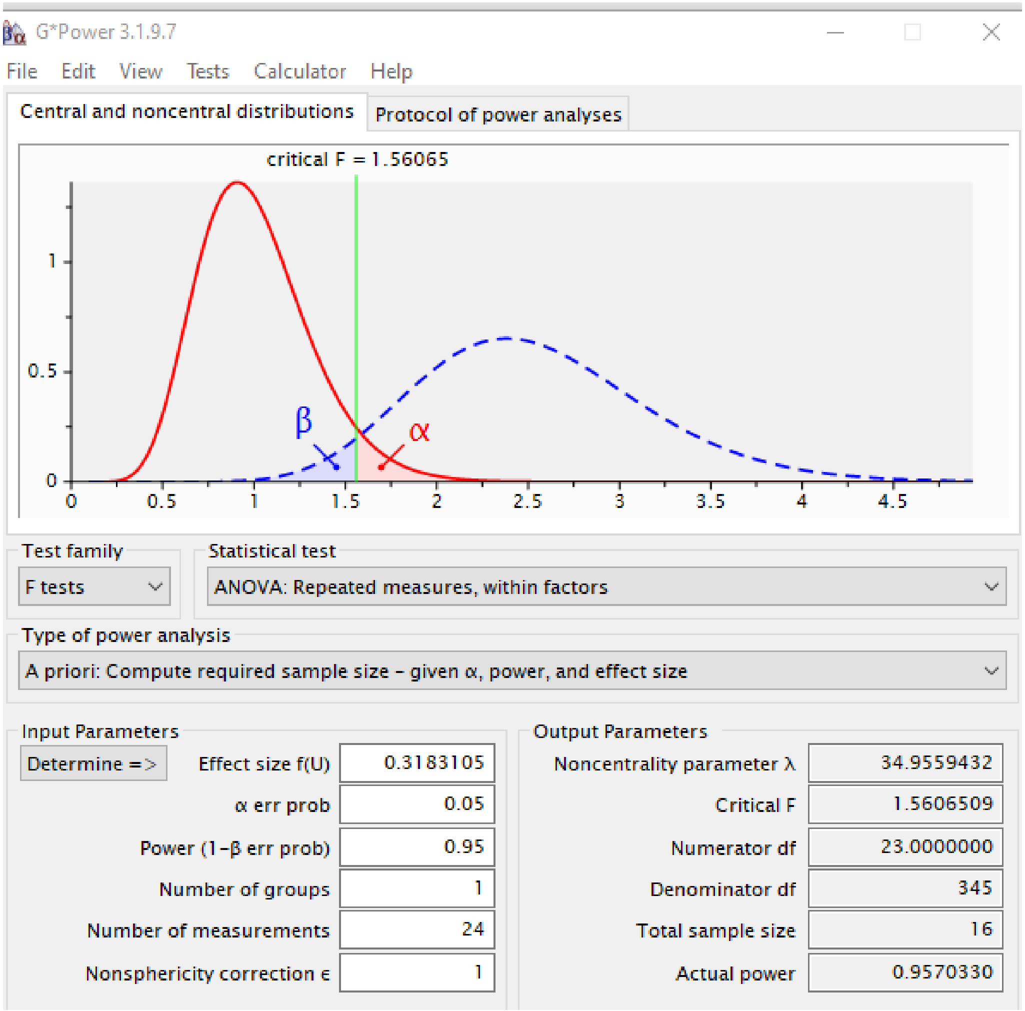


Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Cue	1671.70	3	557.23	6.69575	< .001	0.225
Residual	5742.33	69	83.22			
Response	12975.02	2	6487.51	50.29389	< .001	0.686
Residual	5933.63	46	128.99			
Transfer Test ✻ Cue	360.95	3	120.32	2.35604	0.079	0.093
Residual	3523.63	69	51.07			
Transfer Test ✻ Response	15880.82	2	7940.41	60.10153	< .001	0.723
Residual	6077.36	46	132.12			
Cue ✻ Response	17297.87	6	2882.98	13.18939	< .001	0.364
Residual	30164.46	138	218.58			
Transfer Test ✻ Cue ✻ Response	1754.55	6	292.43	2.34037	0.035	0.092
Residual	17242.84	138	124.95			

Note. Type 3 Sums of Squares

With this effect size calculation, G*Power indicates that a minimum sample size of n = 16 would be needed:


G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 1.56065
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
β
α
Determine =>

Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f(U): 0.3183105

α err prob: 0.05

Power (1-β err prob): 0.95

Number of groups: 1

Number of measurements: 24

Nonsphericity correction ε: 1

Output Parameters

Noncentrality parameter λ: 34.9559432

Critical F: 1.5606509

Numerator df: 23.0000000

Denominator df: 345

Total sample size: 16

Actual power: 0.9570330

This test is however not diagnostic regarding our main theoretical hypothesis (reduced PIT effect after outcome devaluation in Test 2), because the ANOVA would flag *any* significant difference between conditions as statistically significant.

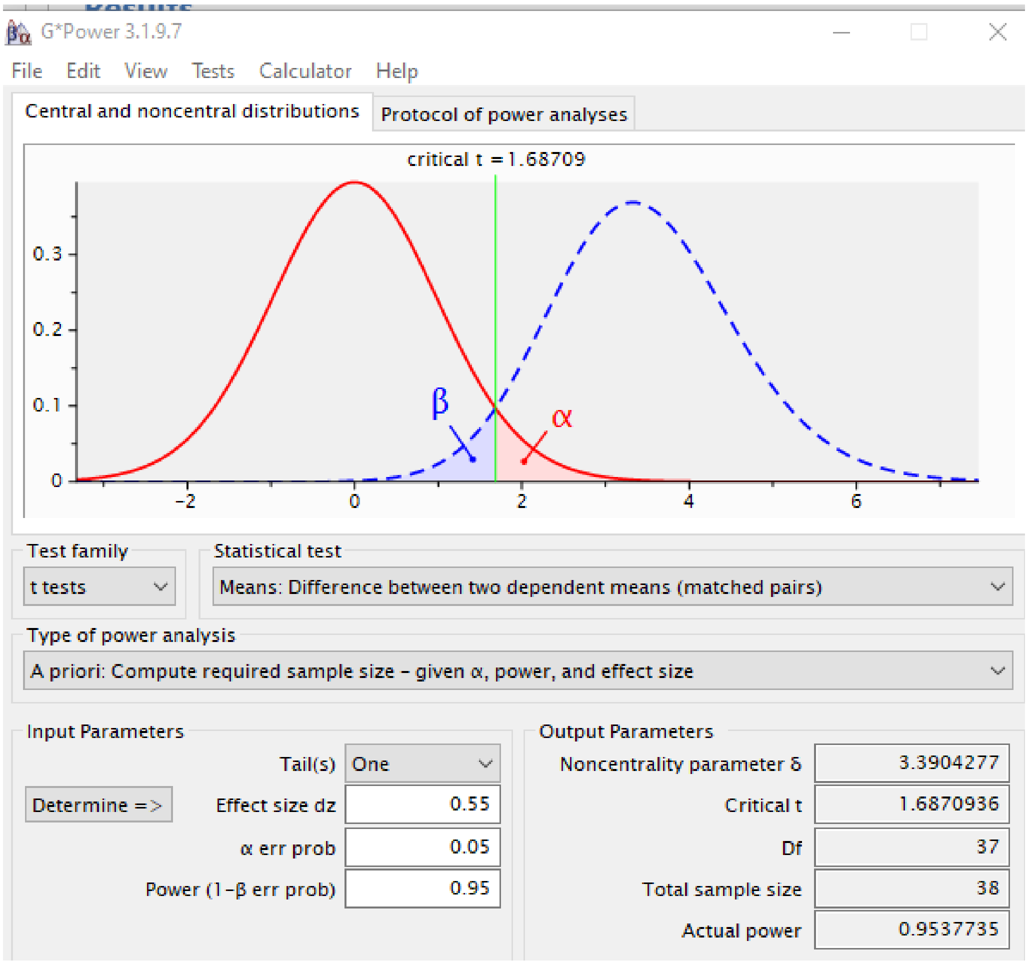
For a meaningful statistical test of the empirical hypothesis, a follow-up analysis is needed that directly compares responding for the outcome before and after devaluation (= R1 keypresses in Test 1 and Test 2). This is a 2 (Cue: CS1 vs CS-) x 2 (Transfer Test: before vs after devaluation) ANOVA of the R1 keypresses. If R1 responding is indeed reduced after devaluation of the associated outcome, as hypothesized, then the Cue x Transfer interaction effect should become significant in this ANOVA. For a fair test of this hypothesis, raw frequencies of keypresses must first be z-transformed to adjust for the overall difference in the base rates of key presses before and after the devaluation treatment (see Statistical Hypothesis 1 above).

Another statistical way to test the same two-way interaction effect is to compute difference scores that index the elevation of (z-transformed) R1 responding relative to baseline in each transfer test (i.e., PIT effect = number of R1|C1 keypresses minus the number of R1|C-keypresses) and to compare both difference scores in in a paired t-test. The p-value generated by this t-test is of course identical with the p-value generated by the 2x2 analysis. In our pilot study, this effect was significant in the hypothesized direction:

Paired Samples T-Test

										95% Confidence Interval	
			statistic	df	p	Mean difference	SE difference		Effect Size	Lower	Upper
PIT before	PIT after	Student's t	2.70	23.0	0.013	0.592	0.220	Cohen's d	0.55	0.115	0.975

Note that this was the statistical effect that we used for the a-priori power analysis presented in the submitted Stage 1 proposal:



Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	One
Effect size dz	0.55
α err prob	0.05
Power (1−β err prob)	0.95

Output Parameters

Noncentrality parameter δ	3.3904277
Critical t	1.6870936
Df	37
Total sample size	38
Actual power	0.9537735

This sample size planning is more conservative than the one provided by the ANOVA above (n = 16). It is also the appropriate statistical test because we linked this specific behavioral effect directly to the neuropsychological hypothesis of increased dACC activation.

To sum up, several statistical tests are planned that test for the effectiveness of the devaluation treatment (1); the effectiveness of our task procedures to implement “habitual” cue-instigated response tendencies in Test 1 (2) and Test 2 (3); and the main theoretical hypothesis that cue-instigated response tendency is reduced after devaluation of the associated outcome (4). A-priori power analyses demonstrate that the minimum sample size is n = 38 to detect the smallest effect of interest (dz = 0.55); this means, with n = 38 statistical power will be sufficient for ALL planned tests (1-4). This was also the reason why we only preregistered a power analysis for (4). However, we see now that this presentation was too condensed. Therefore, we expanded the section on sample size planning in the research proposal text with the information provided above. We also added hypotheses concerning manipulation checks to our summary of the research plan (Table 2 in the revised version).

- No power analysis is conducted for the key FMRI part of the study. Since the functional investigation is the main (new) outcome of the study, this aspect should be at least discussed (cf. e.g. link). An actual power analysis should actually be possible to conduct given the mostly ROI-based approach of the authors.

First, we fully agree with the blogger's statement that "preregistration in neuroimaging is a high stakes intervention" and that "one of the greatest challenges for fMRI research is power analysis". We assume that the reviewer's proposal of an "actual power analysis" refers to a power analysis based on extracted fMRI ROI data. We do not have these ROI data and there exists no comparable neuroimaging study that could be used to this end (i.e., our planned study is novel and original in this respect and does not replicate a previous neuroimaging study). In a guideline, Mumford (2012) recommends an fMRI pilot study with "somewhere between 6 and 10 subjects for a one-sample *t*-test" and to include the pilot data in the analyses of the final study at the cost of a slightly inflated Type 1-error. We could use this sequential approach; however, this would not compensate for the general problem of a poor and highly variable effect size estimate based on a small sample size. If we would run a large fMRI pilot study (*n* = 38 as suggested by our power analysis based on behavioral data), then we would run the proposed study twice, which is not an option due to the high financial costs of MRI research.

To give a concrete number: one testing hour at our fMRI facility costs 300 EUR (340 USD). With n = 38, a total cost of 11.400 EUR (13.000 USD) incurs for scanning alone, with additional costs for personnel, replacement of dropouts, and monetary compensation of participants. This is a huge budget associated with great professional risks (we must pay the bill from our own purse). We hope that the reviewer can understand that these, admittedly mundane, financial reasons prevent us from running a large fMRI pilot study, even though it means that we cannot implement the best-possible approach for a power analysis.

Is our power calculation based on behavior data reasonable? Yes, we believe it is because we hypothesized a logical relation between the behavioral effect and the neural systems subserving these behaviors (for a detailed theoretical justification of this link see Eder & Dignath, 2019). We do not know the strength of the brain-behavior association, but this an uncertainty that we (and many fMRI researchers before us) are willing to risk.

- A power pf 0.8 with alpha 0.05 is targeted, which appears to be below the usual standard for RR (0.9 or 0.95 power ideally with alpha of 0.02), not sure what the PCI RR guidelines are at this level, but likely more stringent than 0.8/0.05 given that an IPA for several journals with more conservative thresholds is granted at the end of the stage 1 review process.

It seems that the reviewer has misread the target power level for the conducted pilot study (P = .80) for that of the proposed fMRI study (P = 0.95) (emphasis in red color):

“The a-priori power analysis showed that *N* = 38 participants will be needed to detect an effect of this magnitude and larger with sufficient statistical power (1-beta = 0.95) in a one-tailed matched paired t-test with α = .05.” (Sample Size Calculation, p. 21).

“N = 38. A-priori power analysis for the detection of an increased dACC activation after relative to before the outcome devaluation in a one-tailed matched paired t-test with 1-beta = 0.95 and alpha = .05.” (Sampling plan, Table 3)

- Relying on small-n pilot studies for power calculation might be problematic, e.g. https://www.sciencedirect.com/science/article/pii/S002210311630230X), this aspect should at least be discussed. Likewise, given the typically over-inflated effect sizes in previous literature, it is generally recommended for RR to conduct power analyses based on principled grounds, e.g. determining a smallest effect size of interest and then how this SESOI can be detected given the specific task/population intra- and inter-subject variance.

While we agree with the reviewer that a large sample size provides better precision for effect size estimation, it must be also noted that we based sample size planning for our pilot study on statistical power considerations. Specifically, the pilot study was reasonably powered (1-beta = 0.80) to detect an effect obtained in a source study that used comparable study procedures (Eder & Dignath, 2016). On the basis of this explicit power calculation, we believe that the sample size of our pilot study was adequate (i.e., sufficiently large) for our research purpose (which were: detection of an effect that exists with reasonable power 1-beta = 0.80 and proof of concept).

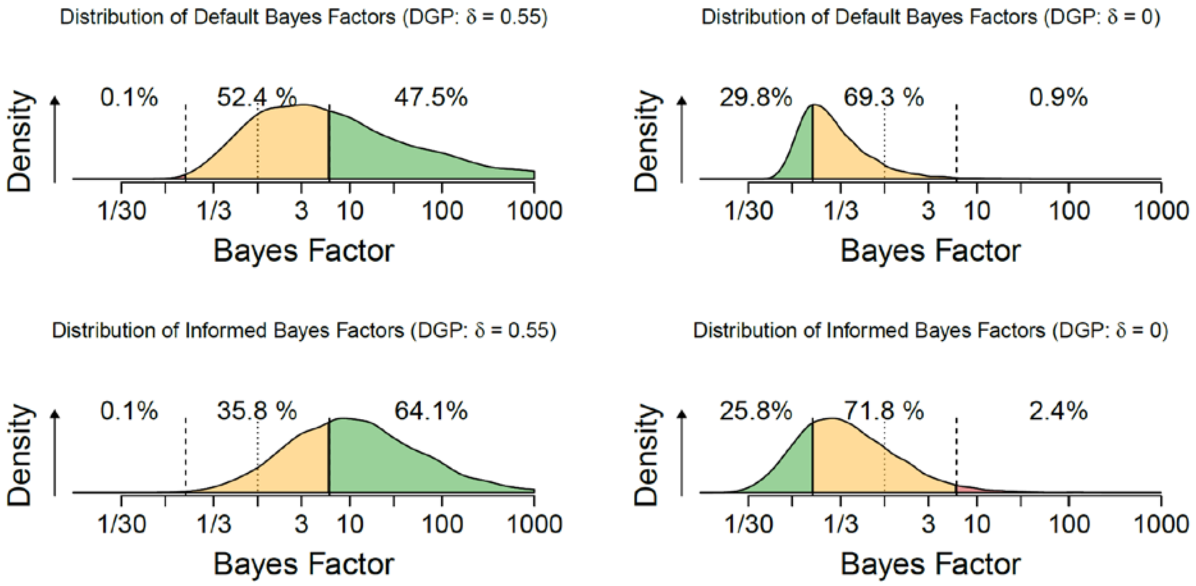
Effects sizes for the main test (see Statistical Hypothesis 4 above) were very consistent across the source studies (pilot study: d_z = 0.55; Eder & Dignath , 2016: d_z = 0.53). Hence, there is no indication of inflated effect size across studies with the proposed research design (we do not have a file-drawer with unpublished studies).

- Exploratory analyses are planned (e.g. p23). Such analyses must usually not be included in stage 1 RR to avoid blurring the limits between planned/confirmatory vs unplanned contrasts. The authors should check the specific PCI RR guidelines or contact the editor on this regard.

To the best of our knowledge, PCI:RR guidelines do not prohibit statements about exploratory analyses as long as they are clearly identified as such. In fact, we believe that their explicit description could be meaningful with respect to the preparation & comprehension of the researchers’ data-analytic approach. However, if the editor (PCI recommender) believes that this statement is grossly misleading for a Stage 1 research proposal, then we would be willing to delete this [last] paragraph from the text.

- Equivalence tests (link) or Bayes factors analyses should be planed, subjected to a power analysis, and conducted for hypothesis on a lack of difference (e.g. control for the expected absence of difference of ratings before devaluation, cf. also the point on outcome neutral conditions for the decisions in case the ratings actually differ). See p.3. of the PCI RR guidelines (link)

We generally agree with the reviewer that complementary Bayes Factors analyses could be useful for the statistical evaluation of null effects, especially in respect to our main research hypothesis of increased dACC activation after outcome devaluation. However, after an extensive discussion we reached the conclusion that our research plan would not benefit from this test for two reasons: Firstly, a Bayesian Factor Analysis (BFDA performed with http://shinyapps.org/apps/BFDA/; for a documentation see Stefan et al., 2019) using a fixed-N design with n = 38 (see sample size planning above), a population effect size estimate of ES = 0.55 (based on our pilot research), and decision boundaries = 6 (for BF10 and 1/6 for BF01, respectively) revealed that the probability for a decision favouring a true null effect (d = 0) would be only around 30% with a default prior and 25% with an informed prior, respectively (see the figure below). Thus, while our planned sample size (n = 38) is sufficient for NHST, it clearly is insufficient for the detection of a true null effect in a Bayesian test.



Distribution of Default Bayes Factors (DGP: δ = 0.55)

Distribution of Default Bayes Factors (DGP: δ = 0)

Distribution of Informed Bayes Factors (DGP: δ = 0.55)

Distribution of Informed Bayes Factors (DGP: δ = 0)

Secondly, the BFDA also showed that, if the effect size is 0.55 and the default prior on effect size is used for analyses, we would need at least 115 observations to obtain a Bayes factor larger than 6 with a probability of p = 0.95. If H0 is true and the default prior is used for analyses, we would need **more than 500 observations** to obtain a Bayes factor smaller than 1/6 with a probability of p = 0.95. These high Ns are unreachable for fMRI research, which is why we decided against the inclusion of Bayesian analyses in our research plan.

- Please specify whether/how excluded participants will be replaced to maintained the minimally required sample size. For instance, if one of the participants must be excluded from one stage of the procedure (should it be only for a technical reason), will it be excluded for the whole study?). Likewise, details are missing on the criteria for to exclude data at the group and individual level (minimal response rate, range of interpretable mean frequency of key press during the CS, etc.).

In the revised manuscript, we now clarify that participant dropouts will be replaced until reaching a final sample size of n = 38. Reasons for exclusions are technical failure, large head movements during MRI measurement (> 2mm translation or >2° rotation within one of the Transfer Test phases) and prolonged difficulties in learning the correct Pavlovian or instrumental relations. In the revised version, these criteria are now summarized in a separate section (*Data Exclusion on the Participant Level*).

- There is a crucial need to include outcome neutral conditions, especially given the 8 stages procedure used in the study. The authors should include sanity checks to ensure that each stage produced the expected outcome, which in turn allow the outcome of the next stage to be interpretable. For example, what would the authors do if the devaluation procedure on explicit ratings of monetary outcomes is not effective? For stage 3, what would be decided if the assignment is still incorrect after several instrumental training?, etc.

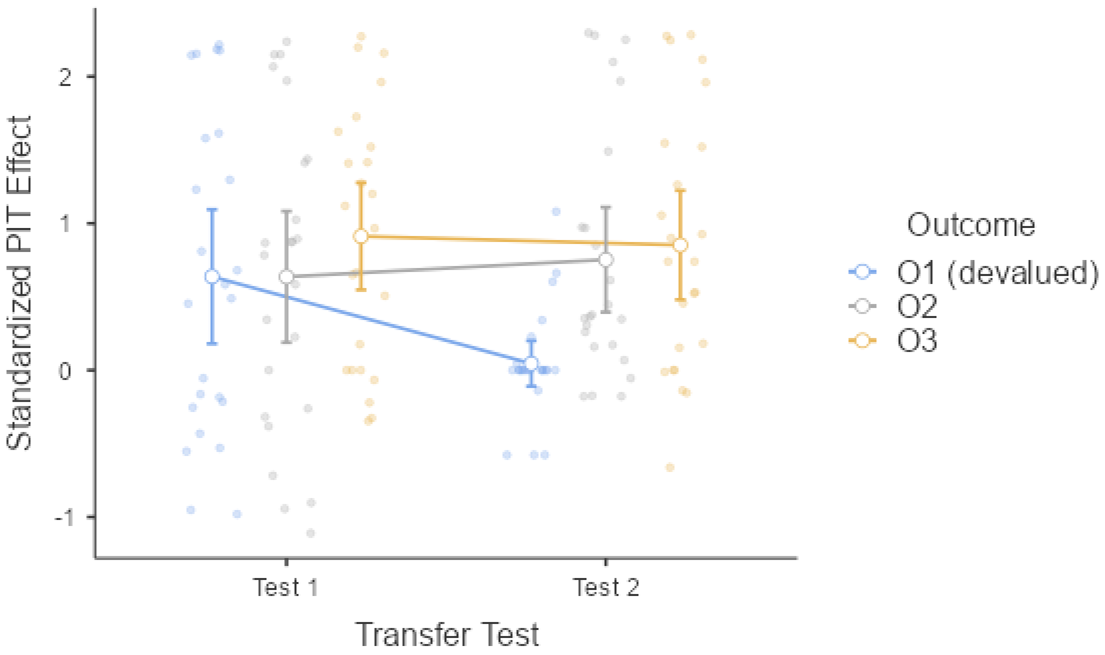
We designed task procedures that should minimize risks of (cost-intensive) dropouts. Critical for the interpretation of the results are manipulation checks of the effectiveness of the learning phases and of the devaluation treatment (see also Table 2 in the revised paper for an updated summary of these hypotheses. In the following, more specific explanations:

- Currency ratings should show the trivial result that ppt give a currency a lower rating after devaluation of that outcome. This could serve as an additional manipulation check for the effectiveness of the devaluation treatment. Note, however, that finding no effect on this measure would not threaten the internal validity of our (devaluation) procedure, as long as the behavioural devaluation effect specified in statistical hypothesis 1 is obtained. Hence, we do not plan to replace ppts who do not show a difference in the expected direction on this measure, and we will perform the planned statistical analyses irrespective of the results obtained on this measure.
- Devaluation effect (hypothesis 1): this is a critical effect for the interpretation of the results. If we do not obtain an effect on the aggregated level, something was really wrong with the study. In this case, we will not publish our study in a journal due to a grotesquely defect study procedure. Individuals who do not show an effect in the expected direction (= lower R1 rate in Test 2 than Test 1) are still included to avoid selective attrition of the sample.
- Pavlovian training: Learning is repeated until perfect knowledge of the Pavlovian relation. Theoretically, endless repetition is possible but extremely unlikely. We will terminate data recording if the ppt cannot remember the six Pavlovian relations correctly after one testing hour.
- Instrumental training: The experiment terminates in the unlikely case that the ppt cannot remember the three R-O contingencies after training and retraining.
- There are no performance dropouts in the transfer tests.

We had no exclusions on the participant level in our pilot research. However, we agree that our research plan should prepare for this possibility. We will aim for a valid sample of n = 38 and potential dropouts (caused by technical failures, large head movement, etc.) will be replaced. This is now also mentioned in the revised research proposal.

- Please report information on data distribution / individual datapoints in the figures (e.g. Fig 2)

Figure 1 shows hypothetical data. We revised this figure so that it matches more closely the paradigm used for the pilot study. Error bars in Figure 2 showed the 0.95 confidence interval. Below the same figure with observed individual data points (z-transformed PIT effects). This figure is now also shown in the revised version of the manuscript.



- Abstract: Too much jargon and too complex/long sentences (took me several reading to understand the last sentence), please simplify or develop to make it accessible to a larger readership

- I found the introduction well written and clear, but too long (2200 words!). I think a more focused review of current literature on the very specific functional hypothesis would be more appropriate for an experimental study article, and more in line with usual journal guidelines. Most notably, I'm not sure it's worth describing in so much details the animal literature in the context of the present human fMRI study.

We revised the abstract and we also made efforts to remove redundancies and shorten the report to increase the readability of the research proposal. To this aim, we also moved large parts of the results section (pilot study) to a supplementary information file. The discussion of rodent research is relevant for the present research question because of rodent and human homologies in action control (Balleine & O'Doherty, 2010). Therefore, we decided to keep these sections in the revised manuscript.

Reviewer 2

I cannot recommend this study because it lacks a proper control. The fundamental argument in the introduction is that the study offers a way to assess the neural bases of habits without the confound of extensive training. Specific PIT is touted as a means to this end, presumably because the often-demonstrated devaluation insensitivity of specific PIT aligns with habitual performance. So far so good; however, it then turns out that specific PIT is usually sensitive to devaluation (i.e., goal-directed) in human subjects, as it is in the reported pilot data, and this is attributed to the severity of the outcome devaluation triggering cognitive control processes, predicted to be implemented by the dorsal ACC. Here's the problem – the study doesn't have a condition in which the outcome devaluation procedure is successful, and yet not severe enough to recruit control processes: in other words, there is no demonstration of "habitual" specific PIT. Without such behavioral pilot data, the study seems meaningless to me. Involvement of the dACC in outcome devaluation sensitivity may well reflect an attentional control signal, but it doesn't say anything

about habits, unless outcome devaluation insensitive behavior is firmly established in a contrasting condition.

Thank you for this critique. The reviewer’s comment can be decomposed into three parts: (1) our study procedures do not establish cue-motivated, aka “habitual”, action tendencies; (2) comparisons between devalued and non-devalued transfer test conditions are not diagnostic with respect to the hypothesized engagement of control processes; (3) the study lacks a proper control condition with a weak devaluation treatment.

Ad 1:

Cue-motivated action tendencies are demonstrated in transfer tests in which Pavlovian cues should modulate performance based on specific features that differentiate between the reinforcers (here: African currencies). Corresponding outcome-selective PIT effects were already empirically demonstrated in our pilot study and by other source studies (Allman et al., 2010; Eder & Dignath, 2016). These outcome-specific PIT effects constitute the explanandum of our research.

Researchers using structurally identical PIT designs have argued in the past that outcome-selective PIT effects represent ‘habitual action tendencies’ because they interpreted statistical null effects of devaluation treatments on PIT effects in humans as evidence for a goal-independency (e.g., de Wit et al., 2009; Hogarth & Chase, 2011; van Steenbergen et al., 2017; see also Allman et al's 2010 discussion of their finding as a "surprising" result). This definition via a particular empirical effect is fundamentally flawed because the null effect could be caused by other factors (e.g., residual value after devaluation, lack of statistical power, etc.). Another problem is that the empirical demonstration of the independence from one particular goal (i.e., specific rewards selected by a researcher) does not imply independency from other goals that could have motivated performance in PIT tasks (for evidence see De Houwer et al., 2018). In short: it is not meaningful to identify “habits” by particular empirical demonstrations of a “goal-independency” (statistical null effect of a [weak] devaluation treatment) or by making assumptions on underlying mental processes (e.g., S-R-O associations) (for a detailed discussion see De Houwer, 2019; Hogarth, 2018; Hommel, 2019; Watson & de Wit, 2018). Rather, habits should be defined procedurally (e.g., as cue-dependent action tendencies in PIT tasks; see Eder & Dignath, 2019), leaving the explanation of the procedurally defined phenomenon open to scientific inquiry (i.e., a procedurally defined habitual action tendency could turn out to be dependent on particular values and a procedurally defined goal-dependent behavior could turn out to be independent of particular outcomes).

Ad 2:

Our research design includes several controls (see Statistical Hypotheses 1-3 in our response to R1 above and in the revised manuscript pp. 21-22):

First, an outcome-selective PIT effect should be observed in the first transfer test before devaluation of the outcome (Statistical Hypothesis 2). This test serves as a control that the task procedure was effective in generating a cue-dependent action tendency (see our procedural definition of ‘habit’ above).

Second, R1 responding for the now-devalued outcome in Transfer Test 2 should be drastically lowered in comparison to Transfer Test 1 (Statistical Hypothesis 1). This test serves as a manipulation check of the effectiveness of the devaluation treatment.

Third, cue-instigated action tendencies (PIT effects) should still be observed for responses associated with non-devalued outcomes (R2 & R3). This serves as a control that the intermittent retraining and devaluation procedures did not generally disrupt transfer of behaviour control (Statistical Hypothesis 3).

The critical empirical test with respect to the Expected Value of Control (EVC) theory is the comparison of dACC activation before and after devaluation of the associated outcome. EVC theory makes the specific prediction that dACC activation should be stronger after compared to before devaluation. We are not aware of any other theory or account of PIT effects that would predict this neurophysiological effect—hence we view it highly diagnostic with respect to this specific causal theory. The reviewer suggests that "involvement of the dACC in outcome devaluation sensitivity may well reflect an attentional control signal" – but what does "attentional control" exactly mean in this context and which theory would predict this attentional effect without making references to control processes similar to the ones proposed by EVC theory? The cognitive control perspective put forward by EVC theory is well prepared to explain attentional effects using the concept of "executive attention" (Botvinick et al., 2001). The reviewer is treating our hypothesis of increased dACC activation as if it was a commonplace assumption shared with other influential theories of PIT—which it is clearly not.

Ad 3:

"the study doesn't have a condition in which the outcome devaluation procedure is successful, and yet not severe enough to recruit control processes"

Here, the reviewer eventually misunderstood our research objective. We do not propose a test of the account with ineffective/weak devaluation treatment. In fact, we mentioned numerous explanations for spared PIT effects after devaluation, with incomplete devaluation being only one of them (see pp. 9-10 in the revised MS and the quote below in our reply to the last reviewer question).

Our research question is whether cognitive control processes (or more precisely, their neural implementation as hypothesized by EVC theory) become engaged when the expected payoff of engaging in cognitive control is high. What is consequently needed is a study condition that motivates cognitive control of cue-influenced action tendencies, which is arguably better realized with a strong/complete devaluation procedure. This condition (Transfer Test 2) is compared with a condition in which motivation for control is not present (before devaluation; Transfer Test 1). In short: it is not necessary to have conditions that compare strong versus weak devaluation treatments. This could be an interesting research avenue for future studies, but it is not necessary for our research objective!

"An everyday example is continuation of snacking although having reached a state of satiety." This is not a good example, because the outcome is being delivered/consumed contingent on the response, which should modulate habitual S-R associations. In other words, the example describes

a "reinforced test" well known to disrupt habitual performance and reinstate devaluation sensitivity (see e.g., Figure 1b, Adams, 1982).

The integrity of our research plan does not hinge on an everyday example given in the introduction, so we deleted this sentence from the text. Of note, it is difficult to find an everyday example for "habitual action" that is performed in complete extinction (i.e., without active pursuit of a reward or goal), which also threatens the ecological validity of this type of laboratory research (for a discussion of this problematic issue see Lovibond & Colagiuri, 2013).

"Hence, an explanation for the motivational insensitivity in previous studies could be that the treatment was simply not strong enough to induce a motivation to control the cue-motivated response tendency."
A more obvious explanation is that studies showing devaluation insensitivity of specific PIT were done in rodents, and those showing sensitivity to devaluation were done with humans (Colwill & Rescorla, 1990, is not a proper comparison because those were SDs, not Pavlovian cues).

Agreed. We now mention this possibility (among others) on pp. 9-10 in the revised manuscript and we removed our reference to Collwill & Rescorla (1990).

> "It should be highlighted that weak and/or incomplete outcome devaluation is not the only explanation for spared PIT tendencies in previous studies. Other possible explanations are (i) species-specific processes differing between humans and rodents (but see Balleine & O'Doherty, 2010); (ii) systematic differences in baseline responding (Seabrooke et al., 2019), (iii) residual beliefs about the informativeness of the Pavlovian cues with respect to the availability of outcomes (Seabrooke et al., 2017), (iv) and the operation of additional goals during the PIT test (De Houwer et al., 2018; Hommel, 2019). Latter explanations concur in the present argument that the motivational insensitivity observed in human PIT studies was the result of a goal-dependent process—and not a design feature of a 'habitual action controller'."

References

Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 402–408. https://doi.org/10.1037/a0017876

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Colwill, R. M., & Rescorla, R. A. (1990). Effect of reinforcer devaluation on discriminative control of instrumental behavior. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*(1), 40–47. https://doi.org/10.1037/0097-7403.16.1.40

De Houwer, J. (2019). On how definitions of habits can complicate habit research. *Frontiers in Psychology*, *10*, 2642. https://doi.org/10.3389/fpsyg.2019.02642

De Houwer, J., Tanaka, A., Moors, A., & Tibboel, H. (2018). Kicking the habit: Why evidence for habits in humans might be overestimated. *Motivation Science*, *4*(1), 50–59. https://doi.org/10.1037/mot0000065

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., & Fletcher, P. C. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *The Journal of Neuroscience*, *29*(36), 11330–11338. https://doi.org/10.1523/JNEUROSCI.1639-09.2009

Eder, A. B., & Dignath, D. (2016). Asymmetrical effects of posttraining outcome revaluation on outcome-selective Pavlovian-to-instrumental transfer of control in human adults. *Learning and Motivation*, *54*, 12–21. https://doi.org/10.1016/j.lmot.2016.05.002

Eder, A. B., & Dignath, D. (2019). Expected value of control and the motivational control of habitual action. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01812

Hogarth, L. (2018). A critical review of habit theory of drug dependence. In B. Verplanken (Hrsg.), *The psychology of habit: Theory, mechanisms, change, and contexts* (S. 325–341). Springer International Publishing. https://doi.org/10.1007/978-3-319-97529-0_18

Hogarth, L., & Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: Implications for dependence vulnerability. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(3), 261–276.

Hommel, B. (2019). Binary theorizing does not account for action control. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02542

Lovibond, P. F., & Colagiuri, B. (2013). Facilitation of voluntary goal-directed action by reward cues. *Psychological Science*, *24*(10), 2030–2037. https://doi.org/10.1177/0956797613484043

Mumford, J. A. (2012). A power calculation guide for fMRI studies. *Social Cognitive and Affective Neuroscience*, *7*(6), 738–742. https://doi.org/10.1093/scan/nss059

Seabrooke, T., Hogarth, L., Edmunds, C. E. R., Link to external site, this link will open in a new window, & Mitchell, C. J. (2019). Goal-directed control in Pavlovian-instrumental transfer. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(1), 95–101. http://dx.doi.org/10.1037/xan0000191

Seabrooke, T., Le Pelley, M. E., Hogarth, L., & Mitchell, C. J. (2017). Evidence of a goal-directed process in human Pavlovian-instrumental transfer. *Journal of Experimental Psychology: Animal Learning and Cognition*, *43*(4), 377–387. https://doi.org/10.1037/xan0000147

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

van Steenbergen, H., Watson, P., Wiers, R. W., Hommel, B., & de Wit, S. (2017). Dissociable corticostriatal circuits underlie goal-directed vs. cue-elicited habitual food seeking after satiation: Evidence from a multimodal MRI study. *European Journal of Neuroscience*, *46*(2), 1815–1827. https://doi.org/10.1111/ejn.13586

Watson, P., & de Wit, S. (2018). Current limits of experimental research into habits and future directions. *Current Opinion in Behavioral Sciences*, *20*, 33–39. https://doi.org/10.1016/j.cobeha.2017.09.012