1 Response to Editor

1.1

Give the paper a much more self-contained and concise format as a Research Report. Your writing must reflect the fact that the study is a conceptual and empirical addition to Crossley et al. (2013), and thus it should make explicit all the arguments needed to understand the model and its predictions, but it should also exclude all the extra information that is not essential to understand the motivation of the study (such as the figures corresponding to already published results).

The revised manuscript is not greater than 5000 words, and is therefore suitable for consideration as a research report. We have made a sincere effort to quickly and clearly indicate that this paper is a conceptual and empirical addition to Crossley et al. (2013). We've followed the suggestions of the reviewers by including the relevant background from this earlier work in an explicitly labeled subsection. We have also removed almost everything that reviewers thought irrelevant. One notable exception to this is that we have opted to keep the figure showing the data from Crossley et al. (2013). Our feeling is that this figure makes the digestion of this background material considerably easier. If, after reading the revision, the consensus is still that this figure is unnecessary, then we will happily remove it.

1.2

Argue convincingly, if it is at all possible, about the extent to which the present study is consistent with your own predictions. If I understand something about the rationale of your study, it is aimed at understanding why relearning a procedural task after having experienced a period of extinction proceeds faster than the first learning process. However, if the results observed in this procedure just fail to show those savings, then, no matter how eloquently you could speculate about the role of fatigue in dual-task conditions, I guess that this just shows that the procedure has failed. My recommendation in this case would be to go back to the lab, remove the sources of this contaminating fatigue, and conduct better experiments.

Our previous analyses computed savings by considering all 150 reacquisition trials against the first 150 trials of acquisition. This choice was motivated by the original Crossley et al. (2013) study, which computed savings by considering all 300 trials of acquisition against all 300 trials of reacquisition. However, the choice of the number of trials with which to compute savings is ultimately arbitrary. For instance, in the visuomotor adaptation literature, it is common to use as few as 8 trials to compute savings. The rationale for this is that since savings is about *relearning* faster than initial *learning*, differences on early trials are the most diagnostic. In this spirit, our revision considers only the first 50 trials of acquisition and reacquisition. This method makes the key point of

our result – that savings is present in the no dual-task control but not in the dual-task conditions – much clearer.

1.3

Revise procedure and analyses: In addition to facing the most serious concern about the lack of savings and its potential causes, there is also another couple of problems concerning the dual-task procedure and the statistical analyses. As for the dual task, I'm worried that you might have included in your sample a large series of participants who performed the Stroop task only slightly better than expected by chance. As judging from the histograms, there were many participants in your samples who performed below 80% of accuracy, which could be taken as a reasonable limit if participants are really complying with the described instructions, and there is still a number of participants who performed at chance. This is plainly unacceptable if you want to assume that participants are performing this secondary task.

Our revised manuscript and corresponding new analyses exclude participants who failed to perform the dual-task at an average accuracy of 80% or greater. Performing this exclusion had a modest effect on our results. In particular, the overall accuracy reached during the acquisition phase is now somewhat different between some conditions. This raises the possibility that our between-condition savings comparisons might be contaminated by differences in acquisition accuracy. Fortunately, this was true of few enough conditions that our overall inferences are left intact. We are nevertheless careful to explicitly address this facet of our data.

1.4

Finally, you must exhaustively revise the statistical analyses reported in the study, especially with respect to the reported degrees of freedom. I failed to understand how you conducted your statistical analyses, but my computation of the degrees of freedom corresponding to the reported designs is very different from those reported in the manuscript. I'm afraid that your inferences may be unwarranted if your computations are mistaken.

The unusually high degrees of freedom reported in many of our original statistics come from two sources. First, the t-tests treated each block from each participant as an independent sample. This is certainly a mistake on our end, and we thank the reviewers for noticing. The second source is that the t-tests use the Welch-Satterthwaite approximation to the degrees of freedom appropriate when homogeneity of variance is violated. This method is now explicitly mentioned in a methods subsection titled "Statistical Analyses."

2 Response to R1

2.1

One main concern is that at present the logic in this paper is hard to follow. I say this, being very familiar with much of this group's earlier work on Rule-Based and Information-Integration learning, though I had not read the series of Crossley studies on which the current paper builds. In fact, after reading the current manuscript, I went back to look at the Crossley JEP:General paper to help me grasp this one. We can't expect readers to do this, though, so the logic needs to be spelled out early here in a way that is self-standing from the original.

We heavily revised and rewrote the introduction to convey our prior work, and the logic of the present study more cogently and concisely. Please see response 1.1 in "Response to the Editor" for details.

2.2

The key concept of feedback contingency is not defined clearly or consistently. The abstract states: "feedback contingency, defined as the degree of non-randomness between response and outcome, plays a vital role in controlling a gate that normally prevents knowledge from being modified during interventions". As I read this for the first time, I wondered whether this meant actual feedback contingency or detected feedback contingency. The title suggests the latter, which seems to be what is actually meant, but that distinction is not stated clearly in the abstract, nor the first paragraph of the intro, nor, I think, through most of the text. In fact, adding to my confusion, feedback contingency is defined in two different ways in this paper, e.g., in the abstract "the degree of non-randomness between response and outcome" vs. in the discussion (p. 15) "the correlation between response confidence and response valence." These don't seem equivalent.

We now consistently refer to feedback contingency as "the correlation between response confidence and outcome."

2.3

Some statements seemed very general and vague and felt like double negatives. For example, I had trouble with the abstract's statement that increasing the cognitive load, ""should disrupt the accurate estimation of contingency, and thereby prevent the gate on procedural learning from closing." It seems it would be more direct to end with "thereby keep the learning gate open, so that the original habit can be unlearned."

We changed this sentence to "If feedback contingency is estimated by executive mechanisms, then increasing cognitive load during the intervention phase (by requiring participants to simultaneously perform a dual task) should impair the ability of participants to detect random feedback. This should in turn cause the TANs gate to remain open, thereby allowing RF to modify the procedural knowledge that was acquired during initial learning."

2.4

Much of the introduction describes the Crossley paper, and yet it is not clearly stated where that is the case (i.e., where the Crossley paper is being described vs the present experiment). Just one example of this is at the beginning of the last paragraph on the first page of the introduction: on first reading I wasn't sure that this paragraph was still referring to Crossley et al. And Figure 3 shows Crossley data, but the Figure caption doesn't indicate this. Perhaps subheadings would help to differentiate the Crossley paper findings from the current study.

Thank you for the helpful suggestion. We have added subheadings to the introduction and believe that they greatly aid the readability of the paper. We have also added a Crossley et al. (2013) reference to the figure caption.

2.5

In addition, right now the second paragraph moves into details of Crossley's evidence without making the broader picture clear. The description of how the model evolved (in the first few pages of the intro) is difficult to follow, because the first-time reader doesn't know where we are headed or why. I think the paper needs to begin with a clear, concise, specific statement of the underlying hypotheses about unlearning (or overwriting) that are being tested here, before going into the details of the evidence in the earlier Crossley work that led to these hypotheses.

The introduction has been heavily revised to address this and other similar comments. For instance, our opening paragraph now reads:

"Relapse often occurs when an addict returns to the original context of their drug use (Higgins et al., 1995). This may occur because interventions given in clinics do not modify addiction-driving stimulus-response (SR) associations, but instead cause the learning of new clinic-specific associations. Returning to the original context of drug abuse then reactivates the preserved addiction-driving SR associations, causing relapse. If true, then this hypothesis means that the brain has a gating mechanism to protect learning obtained in old contexts from being modified. Our prior work, which was focused on understanding this gating mechanism (Crossley et al., 2013), found that feedback contingency defined as the correlation between response confidence and outcome is a principle driver of this gate. The present study is an extension of this earlier work, asking

whether the estimation of feedback contingency depends on executive mechanisms. The rest of the introduction proceeds with a brief summary of the key findings reported by (Crossley et al., 2013), followed by the logic of the current study."

What follows is marked by subheadings indicating whether we are talking about Crossley et al. (2013) or the present study. We hope you agree that the revised version is considerably easier to digest.

2.6

Another general concern is that the present paper, while offering intriguing data, seems to be missing some key components that would provide clearer evidence for the conclusions that are being drawn. For example, a key assumption underlying the conclusions drawn here is that the original category learning is not itself declarative. There is certainly evidence from the earlier work of this group that that is the case, but providing direct evidence for that here (using the current dual task) would strengthen the paper. Thus, it seems important to include a condition in which the dual task occurs throughout training; the dual task should not affect the original category learning, even though it is influencing contingency detection during the intervention phase. As another example of missing conditions, we don't have some of the clearest evidence regarding the extent to which unlearning has occurred because the present study lacks New Learning groups (of the sort included in the Crossley 2013 paper). Thus, the present experiments are more like an addendum to the Crossley 2013 paper, than a freestanding study, and so they provide a more modest increment in evidence than is typically expected for a JEP paper. All of this would, perhaps, be one thing if the present paper could be framed as a brief report, but given the complexity of the argument, at least as presented in this manuscript, this doesn't seem feasible.

We agree with the reviewers assessment of our work as an addendum to our previous work. Our revision is written as a brief report (less than 5000 words), and makes explicit that it is an extension of our earlier work.

The reviewer also makes an interesting point about whether or not the underlying learning is procedural or declarative, but the proposed experiment — to include a condition with dual-task present throughout — has a few problems. First, the experiment is already very long, and including an additional 250 trials with dual-task performance would increase this length substantially. Second, while true that II category learning is much more immune to dual-task interference than RB category learning, this is still subject to overall difficulty constraints. Starting both the dual-task and the category learning task right from the beginning would significantly increase the number of trials needed to reach adequate categorization accuracy. Third, there is already 20+ years of existing research indicating that II categories like those used here are reliable tools to tap into procedural systems, and that applying a dual-task does not change this mapping (in particular, see Crossley et al. 2015 listed below).

New learning conditions would be useful, but (1) to include them for all tested conditions would mean another 150+ participants (2) we don't believe they are necessary here because the results obtained by Crossley et al. (2013) showed perfect convergence when drawing inferences from New Learning vs Relearning conditions.

Crossley, M. J., Paul, E. J., Roeder, J. L., & Ashby, F. G. (2016). Declarative strategies persist under increased cognitive load. Psychonomic bulletin & review, 23(1), 213-222.

2.7

Related to point 2 above, as the manuscript indicates (page 14), some of these results are unanticipated in light of this group's earlier work, e.g. the lack of robust savings in any group here. That is, contrary to the idea (and their earlier findings) that random feedback during intervention leads to no unlearning (i.e., to savings of the original learning), in fact none of the present groups, including the control, showed such savings. Some plausible explanations for these unanticipated results are offered, but this situation seems to make those missing conditions mentioned in 2 more problematic.

Please see section 1.2 of "Response to the Editor" for our response to this point.

2.8

On page 6, is the paragraph concerning the mixed-feedback really necessary here? It seems the goal here should be to present only as much of the detail of the original Crossley paper as is needed to set up the present study.

The entire introduction has been heavily revised. However, we do think that some discussion of mixed feedback is necessary to properly motivate the importance of feedback contingency. We have tried to make this as cogent as possible.

2.9

On page 9, end of middle paragraph, I think that the trial numbers indicated here for condition 4 (400-650) seem to be inconsistent with what was said in the middle of page 7?

Thank you for catching this error. The correct trial numbers are 350 to 600, which is consistent with the prior description. The revised text fixes this.

2.10

Bottom of page 11: It is stated that in Fig 5, the red lines are always below the blue line during reacquisition. This does not seem to me to be true of Fig 5D.

Thanks for the catch. This passage no longer appears.

2.11

Page 14, summary: Contrary to what the third sentence here says, this paper does not investigate the "neural mechanisms" in any direct way, so I think better to not overstate.

We agree that the evidence for neural mechanism is indirect. We removed these claims, except for our final remarks, where we were careful to properly qualify them.

3 Response to R2

3.1

The introduction is very confusing. There is a lot of material that is explained in a very short amount of space. I am not sure why the authors go into so much details of Crossley et al. (2013) and why they even include figures of past results. I would suggest removing the figures and many details. Also, breaking the introduction into subsections might help to streamline the presentation. The authors might also cover work that was not performed in their lab.

We heavily revised our introduction. The current revision is explicitly framed as an extension of Crossley et al. (2013), so the time spent describing these earlier results should make more sense. We hope our new use of subsections in the introduction also increases our cogency.

3.2

In Fig. 2, what is SPN?

Thanks for the catch. SPN stands for Striatal Projection Neuron. This neuron is also called an MSN, for Medium Spiny Neuron. The inconsistency has been corrected.

3.3

p. 6 "More specifically, when the feedback is random the correlation between response confidence and feedback valence is zero." This is

true, but there are other cases when the correlation between confidence and feedback is zero. For example, when one begins a task and has no idea how to proceed. Confidence is near 0, and some of the responses are correct (randomly). Why are the TANs not closing the gate?

This is a very astute observation. The model built and tested by Crossley et al. (2013) dealt with this by assuming a non-zero prior for the feedback contingency estimation.

3.4

p. 12: The number of df in the analyses are uncharacteristically large. Did the authors use each trial as a df instead of averaging the data by subject (e.g., t-tests with over 900 df)?f

Please see response 1.4 in "Response to the editor" for our response to this point.

3.5

p. 12 "The Condition Block interaction was also significant [F (4, 2598) = 14.64, p ¡ 0.001, 2 = 0.02], reflecting the slower change in performance in the dual-task conditions relative to the no dual-task control." The authors would need to decompose the interaction in order to make that claim.

Our statistical results have been heavily revised and this passage no longer appears.

3.6

p. 14 "...whereas savings in the other dual-task conditions were all marginally less than in the control condition..." The p values are all ¿= .1, and some even ¿ .2. It is misleading to consider them as "marginally" less. As the authors mentioned in Footnote 1, these tests should also all be corrected for multiple comparison – so your significant result might also not be significant.

Our statistical results have been heavily revised and this passage no longer appears.

3.7

p. 14: I would suggest removing the regression model since it is not bringing a new understanding of the data and further increases the number of statistical tests on the same data.

We have removed the regression model.

3.8

p. 15: "Specifically, our goal was to determine whether prefrontal-based declarative memory mechanisms mediate contingency estimation." Similar claims are also made on p. 19: "Our results indicate that prefrontal networks likely do play an important role in controlling the estimation of feedback contingency, and therefore may provide an accessible cortical target for electrical or magnetic intervention." The experiment supports the hypothesis that estimating contingencies depends on executive functions (or a declarative mechanism), but there is no direct test of the role of the PFC in the estimation of contingencies so I suggest removing these claims from the paper.

We have removed these claims except from our final remark, where we have carefully qualified it.

4 Response to R3

4.1

I did not have significant concerns with the experiment. The primary manipulation (use of a numerical Stroop task) is appropriate. The statistical analyses are appropriate and complete.

We thank the reviewer for the optimistic comment. None of our revisions should influence this comment to the negative.

4.2

I thought the Introduction was particularly well written: the authors successfully laid out the logic of the study, which required integration across neural, computational, and prior behavioral work.

Again, we thank the reviewer for the remark. We have made improvements including the use of subsections to delineate when we are referring to our previous work and when we are referring to the current work. We hope these changes help.

4.3

My main concerns are with the Discussion. It lacked a clear conclusion and "take-home" message for the reader. In addition most of the discussion was taken up by two somewhat tangential topics (lack of savings and the procedural nature of category learning). The discussion should be broadened, and the treatment of these two topics condensed.

Our revision uses new analyses that make the presence of savings much more obvious, and therefore removes the need for the regression model and discussion surrounding its interpretation. We did not, however, broaden the scope of our discussion. We made this decision primarily to adhere to the word limits given our new goal to meet the requirements for a research report.