In general, the correct and full reporting of the Bayes Factors addresses my primary concerns about this study. Many of my comments below are intended only to clarify my original statements for discussion purposes.

Serje Robidoux (initial)	**Rebuttal**	**Reply**
3.1 The term "automatic" is not consistently defined across studies (see Logan, 1988, and more recently Reynolds & Besner, 2006). I think it's important that researchers in the area are explicit about what they mean by the term, and which features of automaticity they are concerned with. I believe a discussion of this problem is warranted.	The paper is not concerned with the disparate nature of definitions of automaticity across studies and theories. This would be appropriate in a review article, or a paper that tried to differentiate between different definitions by using different operational definitions to produce different predicted behaviours. The aim of our study was to test one view of automaticity. We used a definition that stemmed from a particular theory (Logan's Instance theory) and an experimental paradigm – the dot counting task – that has been used to test various aspects of that theory. The definition of automaticity in that theory holds that performance is automatic when performance is determined by retrieval of a response from memory rather than the application of an algorithm to generate a response. On the basis of this definition, the dot counting task allows for an operational definition of automaticity – automaticity is attained when RT on	Upon rereading the article, the second paragraph provides the statement I was looking for.

	repeated dot stimuli is independent of the number of dots in each stimulus.	
3.2 Data is shared, but only in aggregate form making it impossible to replicate the analysis from scratch, or conduct alternative analyses with different decisions around outlier removal etc. I would prefer to see the original trial-level data included as well (anonymized, of course). General Statement on Open Science I think that open science principles are critical to improving research, and have adopted a policy of promoting them in my reviews. To this end, I would invite the authors to provide the data, materials, and analysis files/scripts in a public repository such as http://osf.io so that others can verify and re-examine claims, use the results to inform their own study designs, and ensure a more complete record of experiments to counteract publication bias in meta-analyses. Alternately, manuscripts should include a clear statement justifying the decision not to provide these	Raw data is now available at http://osf.io/cnr5z	Addressed.

materials. (See opennessinitiative.org for more on open science practices)		
3.3 The research question could be clarified with a clearer statement of how automaticity is defined and operationalized here.	See response to 3.1, and lines 112-116.	Addressed.
3.4 The design is an extension of a paradigm used by Lassaline & Logan (1993). The authors test more subjects (17 per condition rather than 4) but with far fewer trials (540 vs. 6000+) per subject, and conducted in a single session rather than across 13 sessions. I think a discussion of these differences is warranted (it's worth noting that Speelman and Townsend (2015) used a similar design, and observed the same pattern of a reduction in the slopes that Lassaline and Logan observed).	It is not clear what it is about the difference between the studies that the reviewer is concerned about, and so it is not clear how we should respond to this request. We were interested in the transition to automaticity, and so it makes sense to only focus on the relative early stages of practice. Given that all but two participants in our study made the transition to automaticity, it is clear that 540 trials was a sufficient number of trials to observe. More trials of the same would not have addressed the research questions we were focussed on.	I should have phrased this as a more direct "The authors should explain why they adopted a much shorter experiment than Lassaline and Logan, since, a priori, this might have led to a weaker manipulation, reducing their power to detect effects or interactions." Pointing to Speelman and Townsend (2015) would provide the necessary support for that design choice.
3.5 1. If I understood correctly, the index of "time to automaticity" is the first block of 18 trials in which the subject achieves a slope of 100ms or less.	We agree with the reviewer that this measure of automaticity is far from perfect. It was, however, the measure used by Lassaline and Logan. The reviewer's	My concern here is not about bias in the measure, but about its precision: A highly variable measure will increase the variance in both groups, reducing the

However, it's quite clear from the data that in many cases the participants do not *remain* below 100ms from this point on, which raises the question of whether the task can really be considered to have become automated at that point, or if perhaps some index that takes into account the stability of that slope going forward is needed. The slope for each block is, after all, based on a maximum of 18 trials (3 for each number of dots). This is likely to be highly volatile.	suggestion of some index of slope stability might work better, however, again some arbitrary value for this index that represents 'stability' would need to be defined. Ultimately, though, it does not really matter which index of automaticity was used, as the same index was used for both groups. Thus, any shortcomings with the index would apply to both groups. The reviewer does not indicate how this might have contributed to the conclusions we made.	power of the experiment to detect differences between those groups.
3.6 2. Relatedly, the authors have adopted this 100ms/dot criterion as the index of automaticity. However, many of the subjects seem not to have achieved that measure consistently during Phase 5 (given the error bars there). How do the authors reconcile these with their conclusion that automaticity is achieved? Perhaps Speelman and Townsend's findings that not all subjects achieve automaticity in this task would be worth discussing here.	This point is largely dealt with by our response to 3.5. In addition, the fact that some people performed sub-automatically in the last phase of practice, after previously performing automatically is most likely an example of the fatigue and/or lack of attention that participants exhibit at the end of this sort of practice experiment.	This point is not made in the article.
3.7		

On reporting and interpreting Bayes factors: 3. I applaud the authors for using both Bayesian and traditional Null Hypothesis techniques to address the question, however the conclusions they draw from the BFs are misleading. Bayes factors around 1 are equivocal: providing little evidence in either direction. Statements such as "A similar conclusion is suggested by a a Bayes .... (Bayes Factor (.05) = 1.63)" (line 301-302) are thus misleading. The Bayes factor of 1.63 suggests no ability to draw conclusions of any kind.	This has been amended to include a more appropriate interpretation of the Bayes Factor result (last line of Results section lines 335-336)	Addressed
3.8 4. The key hypothesis here was that subjects in the aware condition might achieve automaticity faster than subjects in the control condition. This would be indicated by the presence of a group x phase interaction for the slopes such that the aware group's slopes dropped faster than the control group. The ANOVA analysis considers this interaction directly, but the Bayes Factors are reported only for the main effect of group. Why are Bayes	The Bayesian t tests have been replaced with Bayesian repeated measures ANOVAs. These include tests of the interactions between practice and group. These reveal there is no evidence for the existence of interactions in the RT or the slopes data (see lines 278-285 and lines 300-304).	Addressed

Factors for the critical interaction not considered?		
3.9 5. I'm unfamiliar with the notation "Bayes Factor (.05)". What is the "(.05)" here? Similarly, no indication is given of which model the Bayes factor favours. BF=1.63 (line 302) may indicate slightly more evidence for the null, or for the alternative. (Based on the t statistic, I think the BF likely favours the null here.) This ambiguity needs clarification throughout.	The notation has been amended. In all cases, BF_{10} is used, to signify a comparison of the alternative hypothesis with the null hypothesis.	Addressed
3.10 6. The BFs all being around 1 also speaks to the statements in lines 326-330. BFs in that range should be taken to imply that more data is needed to discriminate the two hypothetical models (null, vs. alternative). They certainly can not be taken to support the view that since some effects are significant, there likely isn't a power problem. However, I reiterate that the BFs reported do not appear to directly address the critical interaction. (It's also worth noting that the	See response to 3.8.	Addressed

significant effects are all main effects of phase, while the critical, non-significant results are interactions - ANOVA is more powerful for main effects, so the presence of those effects does not offer much defence against power concerns.)		
3.11 The article raises an interesting question (can awareness of relevant features of stimuli speed the process of moving from "slow and effortful" to "fast and automatic"). The design of the study is a simple extension of one used before, however I wonder at whether the design is well-enough powered to conclude there is no effect. The use of Bayesian analysis is great, but as applied here does not help to address the most important feature of the data.	No response required.	
3.12 There is a significant benefit to the decision to use Bayes Factors. If the BFs for the interaction are equivocal as they are for the main effects of group, the the authors are welcome to simply collect more data until the	We disagree with this suggestion. This is an extreme example of the practice criticised as “p-hacking” in frequentist analyses. We also reiterate the response made to point 2.3. Furthermore, the Bayesian ANOVAs revealed fairly clear	The theoretical and philosophical principles that distinguish Bayesian analysis from NHST have very different consequences for sample size choice. Dienes (2011) provides a very good primer on how the two differ and why

Bayes Factor is able to discriminate the base model (main effects of Phase and Group) from a model with the addition of the interaction.	evidence against the existence of interactions in the data.	topping up sample sizes are p-hacking in NHST, but not a problem for Bayes Factors.
3.13 I also worry about the way that "achieving automaticity" is operationalized here since subjects seem to go from the "automatic" 100ms/dot to "effortful" 101+ms/dot from block to block (though it's hard to be sure how much this occurs, since the data provided are aggregated into phases of 6 blocks.) This suggests to me that there is a high level of volatility in the measure as an index of automaticity.	See response to point 3.5	Though I still think the index is likely to suffer from poor precision, given that it is the index of choice in the literature and that I don't have a clear alternative to suggest, I don't think it should prevent publication.