Dear Zoltan,

Thank you for your insightful review. We address each of the comments in detail below, and hope that you find our revision acceptable for publication.

Kind regards,
Gustaw and Sanne

---

---

---

| | |
|---|---|
| 1) | But first row 1. The final column of the first row is no longer accurate; you will not be testing Hull's theory. |

Apologies for missing this, we have now deleted Hull's theory from the table.

| | |
|---|---|
| 2) | In terms of "interpretation" in the first row, you can leave it as it is. But whether the best estimate is really exactly 66 is not really the point. I would think if the overall confidence interval included any values in say the range 55 - 75, the original estimate was really pretty good. (Imagine you ran a million subjects and no CIs included 66, but they were all tightly formed around 61 days. One would still conclude that the original study had done a fine job of estimating the time taken.) So you mghit want to rephrase along these lines. |

That's a good suggestion and we have changed the approach in line with your reasoning. We will now conclude the finding is replicated if at least three out of four research sites yield 95% confidence intervals containing values between 56-76.

| | |
|---|---|
| 3) | Second row. Three tests are suggested here. Will you conclude consistent performance is important if any one of them are significant? If so you should use a bonferroni or other familywise error correction. If not specify how you will draw conclusions - will you infer consistency is important only if all three tests are significant, for example? You do not justify a minimally interesting effect size i.e. give reasons why there is minimally interesting effect relevant to this scientific problem. Thus, a non-significant result does not count against any theory. Thus, the inference in the final column is incorrect. |

We have now specified that we will correct for multiple testing using Bonferroni. We will use the lowest effect size we are powered to detect (i.e., 0.09) as our minimally interesting effect size. If any of the lower bounds of the three confidence intervals around the effect sizes are above the minimal interesting effect, we will conclude that performing the behaviour consistently is important for the automatization of a routine. After analyzing the data, we will convert the minimal interesting effect size to a raw unit difference.

| |
|---|
| 4) Third row. Be clear about which family you're correcting for: Are you correcting for the fact you are looking at 5 DVs? Or for pairwise comparisons for each of those ANOVAs? State how you will correct for each. The same issue mentioned for row 2 also arises here, you find it hard to give reasons for why some effect is of minimal interest. That is not surprising, but it does mean a non-significant result is not support for any conclusion. What if an effect were interesting that were smaller than even the small ones you are powered to detect? Then you haven't got evidence against any theory. |

We have now specified that we will correct for looking at 5 dependent variables with the Bonferroni method. For the pairwise comparisons in each of the ANOVAs we will use the Tukey-Kramer correction. We will use the lowest effect size we are powered to detect (i.e., 0.12) as our minimally interesting effect size.