After reading your explanation of the illusion scale I think this part can probably stay as it is (although adding similar clarification to the manuscript probably wouldn't hurt either).

There has been an addition to the text regarding the normalised scale to show the clarification for the 50-point neutral stance:

"The normalised (baseline corrected) data will be used for analyses, with a new scale from -100 to +100 with 100 indicating strongly agree, 50 indicating a neutral opinion, and sco res below 0 indicating strongly disagree with the statements on the questionnaire. 50 is maintained as a neutral opinion so that the normalised data still adhere to the thresholds that the participants are presented with during the experiment."

As I wrote previously, it is crucial for a RR Stage 1 to lay out the hypothesis carefully and define the approach used to test it. And as per guidelines all hypotheses require a power analysis. This is to ensure that the risk of obtaining inconclusive findings is minimised. When you work out the power only for the omnibus test, you clearly risk the possibility of non-significant findings in your posthoc tests, especially if there is a large number of pairwise comparisons.

Practically what this means in your case is that Hypothesis 1 must be powered for the individual pairwise comparisons at alpha=0.0125. The test as it stands is for an unspecified difference between the four experimental conditions. But the hypothesis you defined is whether there is a significant difference between MS and NI as well as NIT conditions. To achieve that you need to provide the power of your posthoc tests. Since this should be a strong effect this is unlikely to affect the overall required sample size. There are other solutions (changing the hypothesis to main effect, using two-way ANOVA or planned comparisons) but what I suggested here is the simplest change.

There have been explicit post hoc tests mentioned within hypothesis 1 and these now have their own power analyses included as can be seen below. The alpha for the post hoc tests has been set at 0.025 due to there being 2 comparisons, one between MS and NI, and one between MS and NIT – this has been corrected in the design table which previously inaccurately listed 4 comparisons.

Updated Hypothesis: "(1 – Positive Control) There will be a greater illusory experience, measured via a subjective illusory experience questionnaire, in the (1a) MS condition compared to the NI condition and in the (1b) MS condition compared to the NIT condition."

*Updated Power analyses "Hypothesis 1a and 1b: A priori power analysis using G*Power shows that for a one-tailed difference between 2 means (pairwise) t test, with an effect size of dz = 1.4, alpha of 0.025, power at 80%, a total sample size of 7 participants is needed."*

These post hoc tests have also been added to the design planner.

I see you have included power for the posthoc tests in Hypotheses 2 now. But now these are no longer Bonferroni corrected. Given that these are the specific contrasts of interest you want to ensure that your experiment is sensitive enough to detect these differences but also provides adequate control for false positives. In this context, it isn't clear why you now use 80% power for these tests but 90% for the ANOVAs.

The alphas for the post hoc tests for Hypothesis 2 have now been corrected for 3 comparisons (MS Vs NI, UV Vs NI, NIT Vs NI) and can been seen updated below. 80% power has been used for these since practical constraints limit the sample size that can be tested – all other hypotheses have now been updated to be at 80% power for consistency.

Update to Hypothesis 2 section 2.4.2.2: "Given significant findings in the ANOVA, post hoc comparisons of every condition will be conducted at a new alpha of .016 (corrected for 3 comparisons (MS Vs NI, MS Vs UV, UV Vs NI))."

*Update to Hypothesis 2 power analyses: "Hypotheses 2a – 2c: A priori power analysis using G*Power shows that for a two-tailed difference between 2 means (pairwise) t test, with an effect size of dz = .5, alpha of 0.016, power at 80%, a total sample size of 46 participants is needed."*

Hypothesis 2c has been included in the same power analyses as this has been correct to be a two-tailed comparison considering comments below.

The sample size has been updated within the manuscript based on the larger samples size indicated from these power analyses.

Also, why does the necessary sample size differ between the three tests when they all use the same parameters?

After rereading the comments I have sent you, I figured out the reason for the different sample sizes in your pairwise t-tests for Hypothesis 2. In my previous reading I missed that you are using a one-tailed test for 2C but two-tailed tests for 2A and 2B. I apologise for this oversight - this was an attentional lapse on my part although this also arose from the description of your hypotheses. (Personally this is why I prefer less verbose Design Tables that simply state the comparisons - but I believe I'm probably in a minority on that point).

Anyways, this is yet another example of why the planned statistics must match the hypothesis they are supposed to test. It is not clear why you decided to use a one-tailed test for this comparison. Specifically, the hypothesis reads:
"There will no significant difference in SSEP response across the electrodes of interest (F1 & FC1) when comparing the NIT condition to the NI condition."

This is not a one-tailed comparison. The appropriate hypothesis here would be *"The SSEP response across electrodes of interest (F1 & FC1) will be larger for the NIT than the NI condition."* Of course it could also be the other way around - it would be crucial to define the direction. But critically, you could have a pronounced difference in the opposite direction, which would then be a non-significant one-tailed effect! It is not clear why you would posit a directional effect here, especially considering that you are hypothesising there is no difference between these two control conditions.

This brings us to another issue, which is that frequentists statistics cannot support the null hypothesis directly. Personally I find the best way to deal with this is using Bayesian tests although you could use other approaches (predefined confidence interval, equivalence test). But I wouldn't suggest adding this at this stage, as again this should then really be sent back out to reviewers. Instead, I would suggest following the advice I already gave and use stringent control of false positives for your posthoc tests. This ensures you have adequate power for comparing MS and UV, respectively, to NI. But keep in mind that if you use the same effect size for Hypothesis 2C (Cohen's dz=0.5) you will not have sufficient power to detect smaller differences.

Therefore you should ask what is the minimum difference between NI and NIT that you would consider as evidence against your hypothesis. You may need to adjust this which would require a larger sample size.

Hypothesis 2c has now been changed to a two-tailed test with a minimum effect size of interest being a Cohen's d of 0.5. An effect size of 0.5 would be the minimum difference between the NI and NIT conditions that we would consider as evidence against our hypothesis. If an effect smaller than 0.5 is detected, we would consider this to show no difference between the conditions due to limits of sample size. We will make sure to mention that future research with greater samples sizes would be beneficial to consolidate this finding, however it is beyond the capacities of this study to recruit a larger sample size.

Please note that many RRs use simple preplanned pairwise comparisons for their hypotheses, without omnibus tests. This is an acceptable solution and actually the sensible thing to do here. Looking back at the review history we discussed this previously. Although this would be a substantial change from your current design, which I'd not be very comfortable to make at this point without sending it back out for review to be honest, especially since the ANOVA was added after comments by reviewers. But you could drop the ANOVAs, and specify the necessary pairwise tests, with strict correction for multiple comparisons. (The ANOVAs and any exploration of main effects could still be added as exploratory analyses in Stage 2).

Given the ANOVA's being suggested by a reviewer at an earlier stage, we have maintained these within the manuscript so that it does not need to be sent out for another round of review.