On point 2:

I'm happy to agree to including scatterplots (and commenting on them where necessary). And on the reviewer's minor point, I have amended the passage in question in accordance with her recommendation. Accordingly, the final paragraph of the minimum effect size and sample size determination section on p. 15 now reads:

Data from all participants will be included unless there has been technological failure in data recording or the experimental session was not completed (e.g. because of participant withdrawal). If such events occur, further participants will be run until the target number is reached. There is no clear objective criterion for making decisions about excluding participants on grounds of data quality, so all will be included. In the absence of objective criteria for exclusion of participants there is a danger that some participants might not engage fully with the task and therefore might not differentiate between the rating scales. To check on the frequency with which this occurs, scatterplots depicting relationships between responses on the different rating scales will be included for each experiment. Power analysis is problematic because designs vary between the planned experiments. A sample of studies using launching stimuli and published since 2000 revealed considerable variation in numbers of participants. Several studies reported between 8 and 20 participants (Guski & Troje, 2003; Kim et al., 2013; Kominsky et al., 2017; Mitsumatsu, 2013; Parovel & Casco, 2006; Ryu and Oh, 2018; Scholl & Nakayama, 2002; Vicovaro & Burigana, 2014; Vicovaro, Battaglini, & Parovel, 2020; Zhou, Huang, Jin, Liang, Shui, & Shen, 2012). A few ran more than 20 but had different dependent measures as a between-subject variable, with numbers varying from 14 to 16 for each dependent variable (Hubbard & Ruppel, 2013, 2017; Sanborn, Mansinghka, & Griffiths, 2013). Of the remainder, in ascending order of numbers, Umemura (2017) ran 27; Vicovaro (2018) ran 40; Young, Rogers, and Beckmann (2005) ran 44; Wang, Chen, and Yan (2020) ran 57 with 32 on a causal judgment measure and 25 on a force judgment measure; Young and Falmier (2008) ran 58; Falmier and Young ran 67 in a four-way mixed ANOVA design; Schlottmann et al. (2006) ran 72 in a study where the measure was free verbal reports; Mayrhofer and Waldmann (2016) ran 934 in an online study with 233 or 234 participants allocated to each of four between-subject conditions. Reliability is a major issue in a replication study and there are indications of substantial inter-individual variability in responses (e.g. Schlottmann et al., 2006; Straube & Chatterjee, 2010). For those reasons it was decided to run a sample towards the higher end of the range in recent research, 50 participants. It is anticipated that data from all participants will be included, barring unforeseen events such as technological failure in data recording. If such events occur, further participants will be run until the target number is reached.

On point 6, experiment 1:

Apologies, I misunderstood what the reviewer was asking for. Yes, I think the reviewer is right that the linear trends test the hypothesis, but you can have linear trends without a transition from one impression to the other, so I would say that both kinds of evidence should be tested for. I have amended the manuscript to say this (p. 26), and also the design plan.

The reviewer asks, "would higher ratings on the launching than passing scale... for any width other than the narrowest be sufficient?". Yes, that would be sufficient, though I would be quite surprised if that happened. So, to make clear, both ANOVA and correlations need to be reported, as is now stated in the design plan.

The review says that part of the entry in the design plan was cut off. I had great difficulty fitting the text into the design plan with the portrait format, especially with track changes activated. So, while I have retained the design plan as Table 4 in the manuscript, I have created a separate document for the design plan in landscape format, which should be more readable. I hope this will be useful.

On point 6, experiment 2:

The design plan has been amended with the phrasing suggested by the reviewer. Re the t-tests, I have added a note saying that ANOVA was selected for consistency with the other tests.

On point 6, experiment 7:

O.K., I understand now. I have amended the design plan for experiments 7 and 9 to specify t tests involving comparisons with the scale mid-point.

On point 7, I have added a paragraph to the participants section (pp. 12 - 13) explaining the situation. Apologies for the uncertainty but it is hard to judge the duration of these experiments so I won't be able to make final decisions until piloting has settled the practicalities of running them.

Power analysis is problematic because designs vary between the planned experiments. A sample of studies using launching stimuli and published since 2000 revealed considerable variation in numbers of participants. Several studies reported between 8 and 20 participants (Guski & Troje, 2003; Kim et al., 2013; Kominsky et al., 2017; Mitsumatsu, 2013; Parovel & Casco, 2006; Ryu and Oh, 2018; Scholl & Nakayama, 2002; Vicovaro & Burigana, 2014; Vicovaro, Battaglini, & Parovel, 2020; Zhou, Huang, Jin, Liang, Shui, & Shen, 2012). A few ran more than 20 but had different dependent measures as a between-subject variable, with numbers varying from 14 to 16 for each dependent variable (Hubbard & Ruppel, 2013, 2017; Sanborn, Mansinghka, & Griffiths, 2013). Of the remainder, in ascending order of numbers, Umemura (2017) ran 27; Vicovaro (2018) ran 40; Young, Rogers, and Beckmann (2005) ran 44; Wang, Chen, and Yan (2020) ran 57 with 32 on a causal judgment measure and 25 on a force judgment measure; Young and Falmier (2008) ran 58; Falmier and Young ran 67 in a four-way mixed ANOVA design; Schlottmann et al. (2006) ran 72 in a study where the measure was free verbal reports; Mayrhofer and Waldmann (2016) ran 934 in an online study with 233 or 234 participants allocated to each of four between-subject conditions. Reliability is a major issue in a replication study and there are indications of substantial inter-individual variability in responses (e.g. Schlottmann et al., 2006; Straube & Chatterjee, 2010). For those reasons it was decided to run a sample towards the higher end of the range in recent research, 50 participants. It is anticipated that data from all participants will be included, barring unforeseen events such as technological failure in data recording. If such events occur, further participants will be run until the target number is reached.