I have reviewed the document “Decline Effects Proposed Analyses” as well as version 2.2 of the operations manual for the project. Following are my recommendations and suggestions.

1. If the pre-registration document is meant to be for public consumption or comment, I would recommend adding to it an initial section that describes the design of the overall project and operational procedures. This should include details such as:
   1. the broad structure of the study, with four labs, each of which puts forward four original studies; assignment of blinded/unblinded status to individual studies;
   2. assignment of participants to half-samples for confirmation and replication studies (right now these details are there, but in the middle of the analytic plan);
   3. assignment of analysis order;
   4. assignment of replication order to labs (i.e., using a latin square);
   5. the criteria for a study to qualify as part of the project (i.e., effect of interest must be a between-subjects contrast between two groups; pilot study result must be statistically significant).
2. Include clear statements of the predicted effects, including direction if relevant. This is similar to what you have in the very first paragraph, but that text is a bit ambiguous.
   1. I would also suggest expunging the references to statistical significance in this text because significance is only an artifice of the experiment rather than a theoretically meaningful characteristic of an effect.
   2. If you include directional hypotheses, one-tailed hypothesis tests would also be appropriate. In either case, one- or two-tailed hypothesis tests should be specified in the pre-registration.
3. I would also strongly recommend specifying what effect size metric will be used for the analysis (presumably the standardized mean difference) and how this will be calculated for each study. For studies that do not use interval-scale outcome measures, also specify how effects will be converted to standardized mean differences.
4. A related consideration is that the effects will need to be coded so that positive effect sizes correspond to effects in the hypothesized direction. Consider what would happen if this weren’t done and several of the studies tested manipulations that led to a large *reduction* (negative effect) in the outcome. I assume that the decline effect theory would then predict that confirmation/replication studies should observe *smaller reductions*—that is, *less negative* effects—than the original studies. Averaging these together with studies where the initial results were positive could result in null changes, on average, because the less negative effects would cancel out the less positive ones. If all studies were coded so that positive effect sizes correspond to the hypothesized direction of effect, then the decline effect would predict that changes from original to confirmation or from confirmation to replication should all be negative.
5. Requiring statistical significance for each pilot study that moves forward to the confirmation stage has several implications, I think.
   1. First, I think that this requirement will cause the initial effect size estimates to be inflated. Thus, I would predict that there will be decline effects from pilot to confirmation for this reason alone. (Is this the mechanism that you intend to test here?)
   2. Furthermore, the magnitude of declines from pilot to confirmation should be moderated by the size of the pilot study. Unfortunately, if all of the labs stuck to the recommendation in the operational manual of using N = 200 for each of six pilots, then there won’t be any variation in this dimension and so this moderator effect can’t be tested. However, you might be able to examine this mechanism if you have data from *all* of the pilot studies conducted by each lab. The number of studies, sample sizes, and magnitude of effect size estimates observed in the pilots could then be used to predict the expected magnitude of decline from pilot to replication.
   3. Finally, I would argue that this requirement creates a constraint on generalization of the findings from the project because the set of effects selected for confirmation and replication are likely to be large (so that it is possible to obtain statistically significant results with a between-subjects design and N = 200). It seems likely that, given the parameters of the task, each lab will select to explore effects that they expect to be large—or at least to be adequately powered given the budget/sample size constraints. Thus, the effects whose replicability is tested would not be representative of the full set of effects that are of theoretical interest in the relevant domains (or even to the originating labs).
6. The analyses that you’ve laid seem very reasonable for the most part, but I have some suggestions for how you could simplify them and thus reduce the possibility of having to modify the analytic plan after beginning to work with the data (and also reduce the possibility of inadvertent garden-of-forking paths problems). My main suggestion here is to use univariate meta-analysis methods when possible. Thus, in the first analysis of first/second splits, rather than modeling the *joint* distribution of the first and second half effect size estimates as you’ve proposed, use as your effect size metric the *difference* in effect size estimates between the first and second splits. To be precise here, let me lay down some notation.   
     
   Let be the effect size estimate and let be the corresponding standard error, both from half *h* = 1,2 of experiment *i* = 0,…,4 in study j = 1,…,4 from lab *k* = 1,…,4. Take *i* = 0 for the original confirmation experiment and *i* = 1,…,4 for the subsequent replications in chronological order. Then to study the decline effects from the first to the second split, calculate , with standard error

Under the null of no decline effect, we would expect the differences to have mean zero and very small variance across experiments, studies, and labs. The meta-analytic model here would then be

where is the overall average change from first to second half split; is a lab effect (about which more below); is a study-specific random effect; is an experiment-specific random effect, and is the sampling error with mean 0 and known standard deviation . This model can also be elaborated to test for differences in analysis order and the analysis-order-by-split-half interaction by defining covariates = 1 when the first half is analyzed first, when the second half is analyzed first, when the first half is analyzed first, and otherwise. Including these in the model above provides tests for analysis order effects and the interaction effect, respectively.   
  
In the second analysis, comparing decline effects from original experiments to confirmation experiments, you can follow the same approach of calculating *difference* between the confirmation study effect size estimate and the original effect size estimate. The main advantage of this approach is that it simplifies the meta-analysis model and thus requires weaker assumptions than fitting a model for the joint distribution of the effect size estimates. If the multivariate model is reasonable, then corresponding univariate model for the difference between effect size estimates should give very similar results. But if the multivariate model is mis-specified (e.g., by assuming homoscedasticity of the random effects across the original and confirmation studies), then inferences based upon it will be less credible than the simpler univariate model.

1. Your proposed analytic models include lab-specific random effects, but I would suggest that you might want to consider instead treating the labs as fixed, for both pragmatic and theoretical reasons. On a pragmatic level, it is very difficult to estimate a between-lab variance component with only four units. On a theoretical level, use of random effects models is appropriate when one is interested in generalizing to a population of units, from which the random effects are sampled (cf. Hedges & Vevea, 1998; Rice, Lumley, & Higgins, 2016). In this case, the target population would be the overall average decline effect across a population of *labs*. It might be difficult to argue that the four labs participating in the study are in any sense representative of a population of labs conducting experimental psychology studies. More reasonable would be to limit the scope of generalizations to the populations of studies like those conducted by the four participating labs.
2. A potential challenge in setting up the analysis of these effect sizes (particularly for the third analysis of decline effects over subsequent replications) is the possibility that some or all of the variance components will be very close to zero. If you use restricted maximum likelihood estimation, zero variance components will lead to non-convergence with some software (including the metafor package in R). The analytic plan should therefore pre-specify how to proceed in the event that the pre-specified model is non-convergent. Or, alternatively, the central hypotheses could be tested by estimating meta-regression models using fixed effects weighting in conjunction with robust variance estimation methods (Hedges, Tipton, & Johnson, 2010) to account for the possibility of additional random variation at the study- or experiment level. This approach avoids the need to estimate the random effects variance components (so again, it simplifies the analysis), but can still be interpreted as inference to the population of studies that may vary in their underlying average effects.
3. If you use random effects models, I would recommend using the Knapp-Hartung (2003) corrections to standard errors and hypothesis tests, which lead to more accurate inferences when the number of highest-level random effects is limited. If you use fixed effects weighting with robust standard errors, the analogous small-sample correction is to use Satterthwaite degrees of freedom (Tipton, 2014; Tipton & Pustejovsky, 2015).

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. http://doi.org/10.1002/jrsm.5

Hedges, L. V, & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. <http://doi.org/10.1037/1082-989X.3.4.486>

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693–710. <http://doi.org/10.1002/sim.1482>

Rice, K., Higgins, J. P. T., & Lumley, T. (2017). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. <http://doi.org/10.1111/rssa.12275>

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*. <http://doi.org/10.1037/met0000011>

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. <http://doi.org/10.3102/1076998615606099>