Common pitfalls in experimental design

E-mail distributed on 06-11-2023

Dear all,

For today's stat support, I would like to share advices regarding experimental design, specifically which pitfalls to avoid that will complicate data analysis. The following is a non-exhaustive list of common mistakes that I have encountered over the years. None of these will make your analysis impossible but may complicate it and/or necessitate a model that violates important assumptions. Certainly a design that combines several of the problems below can run into a lot of trouble!

1. Not consulting a statistician

The most common pitfall in experimental design is simply not consulting a statistician. This is always a good idea, and the best guarantee to avoid painful discoveries later. Even a simple one-way ANOVA design may have unforeseen complications, such as peculiar outcome data, missing data, or known assumption violations (see further). Instead of consulting an expert, another good strategy is to write up your analysis plan in advance, even if only to reflect how the data will be analyzed.

2. Non-factorial designs

For factorial designs of the type e.g., A×B×C, ensure that **no condition is missing by design**! Missing conditions will prevent a full interactional model from being fitted (e.g., no A×B×C three-way interaction test). However, in some studies missing conditions may not be avoidable. For example, a study may present emotion stimuli with information from facial, vocal and bodily modalities that is either absent or present (2×2×2). The condition where all modalities are absent will necessarily be missing, since it would require presenting an empty stimulus. Only 7 out of 8 conditions are therefore observed. Some solutions in this case include:

- Present the empty stimuli anyway and measure the response, no matter how nonsensical the values, if one can assume that the resulting data are uninformative noise and will not have 0 variance!
- Concatenate the 7 observable conditions into a single condition factor, which will allow all possible pairwise comparison between the 7 levels but not a breakdown in terms of A, B, and C contribution to the outcome variance.
- Restrict the analysis model to lower-order interactions, e.g., (A+B+C)^2, or to full-factorial subsets, e.g., A*B, or B*C, which would run for the above example.

Table 1. Observed means for non-factorial design in an emotion perception study

		Body	
Face	**Voice**	*0*	*1*
0	*0*	NA	0.82
	1	3.40	4.34
1	*0*	5.12	4.27
	1	3.08	2.87

Some designs are so large that not all conditions can be presented to participants. In this case, a subset of conditions is sometimes presented, ideally corresponding to a fully observed factorial subset, or selected to be orthogonal to a specified interaction order (e.g., main effects only). In R, the package `ExpertChoice` offers tools to construct orthogonal combinations of condition levels to satisfy such requirements. Be warned however that higher-order interactions may still not be estimable in such designs!

Finally, note that missing conditions in the analysis of frequency tables (e.g., chi-square analysis) can sometimes be accommodated as so-called **"structural zeroes"**, by adding dummy indicators for the missing conditions in the appropriate log-linear model. It is a unique solution, however, that does not extend to factorial ANOVA for continuous outcomes. Although solutions have been proposed for this scenario—most famously Type IV ANOVA by SAS—none are ideal.

3. Non-factorial baseline

A variation of the second pitfall is designs where the baseline condition is not part of the main factorial design, but included as a separate neutral "back-up" condition. The question arises how to integrate this condition with the factorial design, with some options **(a)** conducting only simple pairwise comparisons between the baseline and the factorial conditions, or **(b)** subtracting the mean baseline response from all factorial condition responses. For the latter, the interpretation of the outcome changes from an absolute response to one relative to the baseline!

4. Partial within-subjects designs

For various reasons, a design variable may be assigned neither fully between-subjects nor fully within-subjects, for example because **(a)** conditions are observational and it is not guaranteed that every participant completes all of them, **(b)** there are too many conditions to be presented to a single participant, or **(c)** because ignorance of some conditions is required to observe an unbiased effect in others (e.g., participants receive placebo and drug A or B, but never drug A *and* B). The consequence of a partial within-subjects design is that, for every participant, at least one of the within-subjects conditions will be missing, hence methods that rely on complete-case analysis in wide-format data will have 0 cases to analyze! Repeated measures MANOVA cannot be used. Partial within-subjects designs **necessitate the use of multilevel ANOVA**. Although this is not a problem in itself, be sure that you understand this analysis before using it.

5. Repeated categorical outcome

Most of us are familiar with repeated measures designs for continuous outcomes, which are commonly analyzed with repeated measures (M)ANOVA or multilevel regression. While these methods introduce assumptions that should not be trivialized, they are relatively straightforward to apply. This is no longer true when the outcome concerns a repeated categorical outcome, including binary and multiclass outcomes. As a general recommendation, **such outcomes should be avoided**! Repeated binary outcomes may necessitate the use of logistic multilevel GLMs (so-called GLMMs or GLMERs), which introduce complexities that are not well-understood by most researchers. Repeated multiclass outcomes would require a multinomial multilevel GLM but, practically speaking, this model does not currently exist.

Some solutions for repeated multiclass outcomes include **(a)** breaking down the multiclass levels into binary pairwise comparisons with logistic GLMMs, **(b)** using multinomial models of the kind found in "discrete choice" literature, which require super-long-format data with additional parameter constraints on the model, **(c)** removing levels of the outcome to remove the linear dependence among all levels, **(d)** aggregating multiclass choices to continuous frequencies and proceed with continuous models, or **(e)** applying chi-square analyses that will necessarily violate the independence assumption of the underlying data.

6. Conditional outcomes

In some studies, the outcome to-be-analyzed is not guaranteed to be observed, or depends on some criteria to qualify as a "bona fide" or usable response. This can include, for example, **(a)** manifesting certain symptoms of a phenomenon for the phenomenon to be analyzable, **(b)** reaching a required threshold for the response to qualify as a "meaningful" response (e.g., skin conductance response), or **(c)** a secondary outcome whose observation is conditioned on a primary outcome. The primary problem with such variables is that they will subset the data to only the "usable" cases, which does not guarantee that the factorial design remains preserved, or all its conditions equally represented. In fact, if the outcome strongly depends on the manipulations, then it is likely that the outcome is never observed for certain combinations of design levels.

For example, a researcher may have constructed jokes to elicit amusement according to an A×B design, but the combination A1B1 is essentially always not funny, and therefore fails to produce an analyzable amusement response. Any subsequent analysis that wishes to examine the properties of the amusement responses (e.g., physio, facial expression) based on the A×B effects will therefore work with an incomplete design, leading to the aforementioned problems of non-factorial designs. Even if a factorial design can be preserved, the design after subsetting will most likely be highly imbalanced and **non-orthogonal**, which poses challenges for reliable estimation, inferential testing, and causal interpretation of effects.

Although solutions depend on the precise design, one option is to avoid conditioning or thresholding, when possible, for example by treating "missing" responses as 0 rather than missing. For the skin conductance response example, one could choose to analyze all SCRs, regardless of whether they reach the required threshold to be a "genuine" response, so that no cases become missing and all have a continuous non-zero value.

7. Expected problems and violations

Another important pitfall in experimental design is to ignore problems and statistical violations that could have been foreseen during study development. Previously I touched upon this issue in my Stat Support on sample size. For example, longitudinal studies typically suffer from dropout, while randomized clinical trials typically suffer from heteroscedasticity (e.g., placebo condition has smaller outcome variance than treatment conditions). This poses challenges for the sample size that was calculated *a priori* to be sufficient to infer the target effect. Recall that heteroscedastic tests (e.g., Welch *t*-test), shrink the degrees-of-freedom (DF) of the reference distribution, and therefore have lower power than originally intended. This should be taken into account at the design/planning stage.

Planning for dropout can be done by recruiting additional participants to compensate for the fraction of expected dropout. Planning for heteroscedasticity should be done by checking past studies for the amount of DF reduction in heteroscedastic tests and compensated accordingly with additional participants.

Best,
Ben

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79