STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

Designing an intervention study

E-mail distributed on 09-12-2024

Dear all,

Today I want to give some practical tips for designing an intervention study, considering primarily the classical 2×2 design of pre-post measurement (T0, T1) on two independent groups (Intervention, Control), with randomized group assignment (e.g., psychotherapy, mindfulness training, video games, diets, etc.). For the study to conclude that there was a causal effect of the intervention on one or more outcome measures, several control features need to be considered, ideally all the ones mentioned below. Firstly, I discuss design considerations, secondly how to analyze such data to correctly infer the causal effect.

1. Design considerations

1.1 Groups

Traditionally, intervention studies contrast an experimental group against a control group, with both groups measured in parallel, and participants assigned randomly. The control group is typically a *passive control group*, and is measured without experiencing the intervention. The purpose of the passive control group is to exclude the causal confound of "mere time", that is, an effect of the intervention simply due to an amount of time passing between T0 (pre-intervention) and T1 (post-intervention).

Better yet for an intervention study is to also include an *active control group*, which is usually a kind of placebo group. These participants receive an intervention that is unrelated to the experimental intervention, or a version of the experimental intervention that removes its main therapeutic ingredient. The purpose of the active control group is to exclude the causal confound of "mere intervention", that is, an effect of the intervention simply by participating in the study, and the belief in the participant that the activity will impact them (e.g., improve well-being). To this, some researchers add further experimental groups. This can be useful when an intervention contains multiple therapeutic ingredients, and the researchers wish to isolate the effect of each (e.g., medication and therapy versus therapy alone).

Sometimes the passive control group consists of a “waiting list” group, and is also measured in parallel with the experimental group but receives the intervention at a later time (e.g., T1-T2). For the T0-T1 comparison, they still serve as a conventional control group against the experimental group. This design is sometimes implemented out of ethical considerations. That is, if there is a possibility of benefiting from the intervention, it is ethical to provide it to all participants, especially in vulnerable or clinical populations. Moreover, if the primary T0-T1 analysis demonstrates the effect of the intervention, both groups can be pooled to pre-post times to estimate a more reliable intervention effect.

1.2. Time

Typically, an intervention study has a simple pre-post design with two measurement points, one before and one after the intervention (T0, T1). To this, researchers often add a follow-up measurement (T2), after a duration much longer than the T0-T1 interval, to show persistence of the intervention effect. The precise spacing of these measurement points will depend on the intervention and the particular field of study, and could be as short as seconds or minutes, and as long as days and months.

To exclude the confound of mere time, it is not sufficient to have a passive control group present. One should also ensure that the groups do not differ (significantly) in interval duration between measurement times, and that no seasonal influences affected the study. To control duration, researchers should fix the T0-T1 interval as strictly as possible, and establish at data analysis that the groups do not differ in the average interval. To control seasonality, the best strategy is to recruit participants from both groups throughout the entire study window. Combined with randomized group assignment, this will generally ensure that no group is more likely to have been sampled in any given part of the study window.

An alternative to random time sampling is fixed time sampling, where participants are all measured at a common T0 and T1. This could happen in studies where the availability of the participants is somehow constrained, or mass testing is more efficient than individual sessions (e.g., in schools or companies). Whereas fixed time sampling has the benefit of holding both the T0-T1 duration and the seasonality effect constant, it absolutely requires the presence of a passive control group to exclude on the time confound.

1.4. Outcomes

Intervention studies typically have primary outcomes and secondary outcomes, reflecting foremost the interest and priorities of the researchers. However, it may be useful to classify outcomes according to criteria other than their importance:

- **Proximal versus distal:** Outcomes where short-term change is expected (e.g., anxiety), versus long-term change (e.g., school performance). Distal outcomes are usually measured at a later point in the study.
- **States versus traits:** Outcomes that reflect the participant's current state at measurement (e.g., state anxiety, behavioral task performance), versus stable traits and dispositions (e.g., trait anxiety).
- **Experimental versus control:** Outcomes that are expected to be impacted by the intervention, versus outcomes that are not expected to be impacted. For example, an intervention on emotion regulation may wish to show an effect on affective outcomes but not cognitive outcomes. Note that active control groups may have their own experimental outcomes that serve as a control outcome for the experimental group (e.g., when the active control group is an Italian class, we do not expect the experimental group to learn Italian).
- **Mediators versus outcomes:** Some outcome measures of the intervention may in fact mediate its effect on others. This may be especially true if the primary outcome of interest is more distal

than the outcome which is directly targeted by the intervention (e.g., increased emotion regulation skill mediates increased well-being).

Ideally, an intervention study has multiple outcomes, representing a mix of some or all of the above, so that the impact of the intervention can be evaluated comprehensively, including the exact therapeutic mechanism responsible for changing the primary outcome. Likewise, if the primary outcome of a study is for example, well-being, then it is advised to measure it with more than one questionnaire, to capture diverse facets of well-being (e.g., depression, anxiety, quality of life).

1.4. Adherence, commitment, and quality

The design aspects discussed in the previous section are important controls to estimate the effect of the intervention. However, these too will fail if the quality of the intervention protocol is degraded in some way. This can happen in several ways:

- **Dropout:** Participants abandon the study, especially in the experimental groups.
- **Non-adherence:** Participants fail to adhere properly to the intervention (e.g., attend few sessions).
- **Lack of commitment:** Participants adhere but have low (psychological) commitment to the intervention, possibly due to the belief that the intervention does not work.
- **Non-response:** Participants adhere and commit to the intervention but do not show the expected effect.
- **Technical issues:** Some participants receive a different version of the intervention due to technical problems (e.g., equipment failure, replacement instructors).
- **Seasonal issues:** Some participants receive a different version of the intervention due to a significant seasonal issue (e.g., a pandemic forces the intervention to take place via video conferencing).

Generally, one should ensure that the intervention protocol is as similar as possible for all participants to exclude these confounds. As a further safeguard, all issues should be logged, including the use of explicit questionnaires on commitment and satisfaction with the intervention. During analyses, such measures can be useful to find out when the intervention failed to work.

Even when perfect protocol standards are observed, intervention studies will suffer from dropout and non-response. If these rates are known in advance, they should factor into power and sample size calculations for the projected design. For example, if 15% total are expected to drop out over the course of the study, and the intervention is expected to not be effective for 20% of intervention participants, the planned sample size should compensate upwards for these expected losses.

2. Analysis considerations

2.1. Primary analysis

Traditional 2×2 intervention designs (two groups, two time points) have been analyzed in two dominant ways, by repeated measures (M)ANOVA—sometimes called mixed ANOVA—and by ANCOVA. The mixed ANOVA consists of a Group × Time interaction effect on the outcome measures, with Group as a between-subjects factor and Time as a within-subjects factor. It estimates an unconditional intervention effect across the whole population. The ANCOVA consists of a Group effect on the outcome measures at T1, with the outcome measures at T0 serving as a control-covariate. It estimates a conditional intervention effect, across participants that started with equal T0 scores.

The difference and appropriateness of these two approaches has been much debated in the literature. In the worst case, their conclusions can contradict each other, a scenario known as Lord's paradox. However, for interventions with randomized condition assignment, they will likely produce very similar results. To some researchers, the mixed ANOVA appears to be inappropriate when there are baseline differences between groups (e.g., the intervention group is more depressed at T0). While such differences will generally be avoided by random condition assignment, having them present makes no difference for the validity of the Group × Time interaction test, as this test evaluates *relative change* between the two groups.

The mixed ANOVA can also be implemented in a multilevel model, where one again analyzes the Group × Time interaction effect on the outcome measures, but controlling for within-subject correlation of repeated measures by the use of a random subject intercept. This model has the added benefit that it can include time-varying covariates, which is not possible in the mixed ANOVA.

For longitudinal designs with more than two time points, a repeated measures MANOVA is preferred over either a repeated measures ANOVA or multilevel ANOVA. This is because the latter two assume a constant correlation between time points. This is not realistic as correlation between measurements typically decreases for larger time spacing, especially when one of the times occurs much later than the others. By contrast, the repeated measures MANOVA allows an arbitrary correlation pattern between time points.

2.2. Follow-up analyses

Once the primary intervention has—or has not—been established, there are a number of follow-up analyses that can be conducted in intervention studies.

Pooling. If the passive control group was a waiting list group, it could be pooled with the intervention group to common pre-post times (e.g., T1 becomes T0, and T2 becomes T1 for the waiting list), and a more reliable intervention effect can be estimated. Note however that this pooling is only permitted if the primary intervention effect is present! Moreover, the pre-post comparison may be sensitive to seasonal influences, since the waiting list group received the intervention in the later part of the study's time window.

Clustering. Clustering can help separate responders from non-responders. For all participants and outcome measures, one can calculate the T0-T1 change scores, and then submit these change scores to a cluster method (e.g., hierarchical clustering). The idea is to identify qualitative groups with similar change patterns. Although these groups ideally align with the randomized groups, the intervention group is typically more heterogeneous than control groups, containing a mix of responders and non-responders, as well as subtypes among these. Less frequently, control groups also contain a certain number of "responders", which may be participants that changed incidentally (e.g., major life-event). Clustering enables one to identify all these subtypes, including groups that may merely represent data artefacts (e.g., outliers). Once the clusters are identified, simple ANOVA *F*-tests and pairwise *t*-tests can further reveal on which change scores the clusters differ.

A next important step when clustering intervention data is to try and predict cluster membership, ideally using other measures than the change scores that determined the clusters initially, e.g., demographical characteristics, personal history, traits and dispositions, that may give more insight into *why* certain people respond to the intervention and others do not. Such analyses could be done in a univariate fashion (e.g., *t*-tests and chi-square tests) or in a multivariate, model-based fashion (e.g., multinomial regression, discriminant analysis).

In a similar fashion, clustering also enables one to analyze the impact of the intervention quality (e.g., adherence, commitment, satisfaction). Since such measures are often not available for control groups, they cannot be introduced as moderators in the main intervention analysis. However, within the intervention group, one can analyze whether intervention quality predicts being a responder or non-responder.

Moderation. A type of follow-up analysis that is closely linked to clustering is moderation analysis. Instead of empirically determining qualitative groups in the data, the groups are instead linked directly to measured variables. This can be achieved by introducing moderator variables to the main intervention analysis, for example a Gender × Group × Time interaction in a mixed ANOVA. Once again the goal is to find out which characteristics increase or reduce the intervention effect. Here, it is important to note that absence of confounding does not exclude the possibility of moderation! For example, men and women may be perfectly balanced in the experimental and control group (hence no confounding), but gender may still moderate the intervention effect.

Mediation. Finally, researchers may be interested in finding out which processes mediate the intervention effect causally. This can be achieved with various methods for causal inference, such as the method by Baron and Kenny (1986), the Rubin causal model (propensity score matching), or Structural Equation Modelling (SEM). However, doing so requires stringent criteria, such as **(a)** the model containing all important sources of confounding, **(b)** a plausible temporal contingency between the intervention, mediator changes, and outcome changes (each must precede the next), and **(c)** a temporal contingency in measurement of the intervention, mediator changes and outcome changes. Apart from these, I caution that mediation analyses require strong a-priori hypotheses. When these are absent, or some of the preceding criteria are in doubt, it may be safer to run only moderation follow-ups.

3. Conclusion

In sum, a good intervention study should incorporate a number of important design characteristics, such as randomized group assignment, passive and active control groups, fixed pre-post duration, control outcomes, and contingency plans against foreseeable problems. For data analysis, it is recommended to follow up the main intervention analysis with cluster and moderation analyses to gain more insight into qualitative groups of participants and their characteristics.

Many more technical aspects of good intervention design can be discussed which I have not covered in today's newsletter. When designing an intervention, it is therefore advised to inform yourself thoroughly of good practices (possibly consulting with a statistician). Likewise, I note that there exist standardized guidelines for the reporting of intervention studies (CONSORT guidelines) for many different types of designs (not just randomized trials).

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