Advantages of within-subjects designs

E-mail distributed on 12-09-2023

Dear all,

When designing a study, one of the most important decisions to make is how to assign the study's subjects to conditions. Generally we distinguish two types of designs, between-subjects and within-subjects designs. In the former, every subject is assigned to only one condition, in the latter, every subject is assigned to all conditions.

For observational studies, this assignment may be forced by the sampling method. That is, in the absence of experimental conditions, most data collected observationally are cross-sectional at the between-subjects level (e.g., gender, age, traits), with the values belonging to exclusive levels (e.g., one cannot be a smoker and non-smoker simultaneously). Observational data can also be longitudinal, however, where each subject is measured repeatedly, and hence some variables may vary within-subjects (e.g., season, monthly performance).

For experimental studies, the researcher typically has the choice between the two designs. In this case, **within-subject designs should almost always be preferred over between-subjects designs**, because **(a)** they allow stronger causal conclusions, **(b)** they have more statistical power, **(c)** the resulting data are more likely to satisfy distributional assumptions (Maxwell & Delaney, 2004). Today I will discuss these advantages one by one, as well as some limitations of within-subject designs.

Causality

In traditional between-subjects experiments, subjects are assigned randomly to one of the conditions or groups (e.g., Drug A, Drug B, or Placebo). **Randomized condition assignment (RCA)** guarantees that, as the groups grow very large, they should on average have the same characteristics (e.g., mean age, gender balance). Hence any difference that is subsequently observed between the groups, should be a consequence exclusively of the difference between conditions. In other words, RCA prevents subject-level confounding (but not other types of confounding, such as treatment-level!), and allows causal inference of the experimental manipulation on the outcome.

A drawback of the between-subjects approach is that, even for larger sample sizes, it is unrealistic that RCA can match the groups perfectly on all possible confounding characteristics. In fact, significant confounding is often observed in experiments, especially in small samples, necessitating statistical controls and/or limiting causal conclusions.

A within-subjects design improves on this limitation, by holding the entire group fixed, and changing *only* the conditions. This procedure has a natural compatibility with the so-called *counterfactual* interpretation of causality (Pearl, 2000). That is, when we say X causes Y, we imply that, when holding all other factors constant, if X had *not* happened, Y would not have happened. A within-subjects design allows counterfactuals to be observed directly, by assigning the subjects to all conditions. With the groups remaining exactly the same between conditions, the observed differences can only be a consequence of the experimental manipulation.

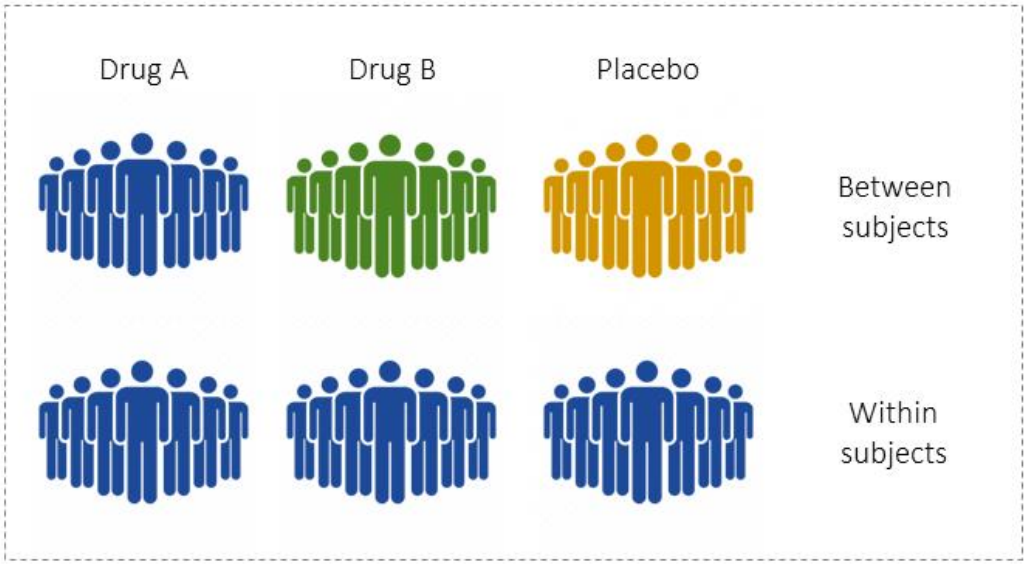

Figure 1. Within-subjects designs change condition while holding the group fixed.

That said, a number of cautions are attached to this:

1. Like between-subjects designs, within-subjects designs aim to exclude subject-level confounding, but cannot exclude treatment level confounding (i.e., confounds in the manipulation).
2. Within-subjects designs are vulnerable to confounds of order, time, and exposure. That is, receiving Drug A before Placebo may have a different effect than receiving Placebo before Drug A. Therefore, within-subjects designs should take care of **randomized order assignment (ROA)**.
3. Even under ROA, presenting certain conditions early may irreversibly contaminate measurement in later ones (e.g., by revealing the purpose of the experiment). If ignorance of the other conditions is essential to the effect of the manipulation, within-subjects designs are not recommended.
4. Neither between-subjects designs (with RCA) or within-subjects designs (ROA) guarantee that the samples are representative of their respective populations. Representativeness and generalizability are linked to the sampling procedure, which should aim to ensure balanced representation regardless of the experimental design under consideration.

Power, reliability, and multiple trials

Another reason to prefer within-subjects designs over between-subjects designs is improved power and reliability. That is, we can detect the same effect as a between-subjects design using a smaller sample. For example, for a standardized difference of means of 0.5 (Cohen's *d,* medium effect), a power level of 80% and significance level of 5%, an independent *t*-test requires 128 subjects total (64 per group), while a paired *t*-test requires only 34, so 4 times more efficient! This is because, by measuring subjects repeatedly, we can estimate within-subject variability, and hence remove this variability from the measurement error in our analysis model. This also makes the resulting estimates (e.g., mean differences in the outcome measure) more precise and hence more reliable.

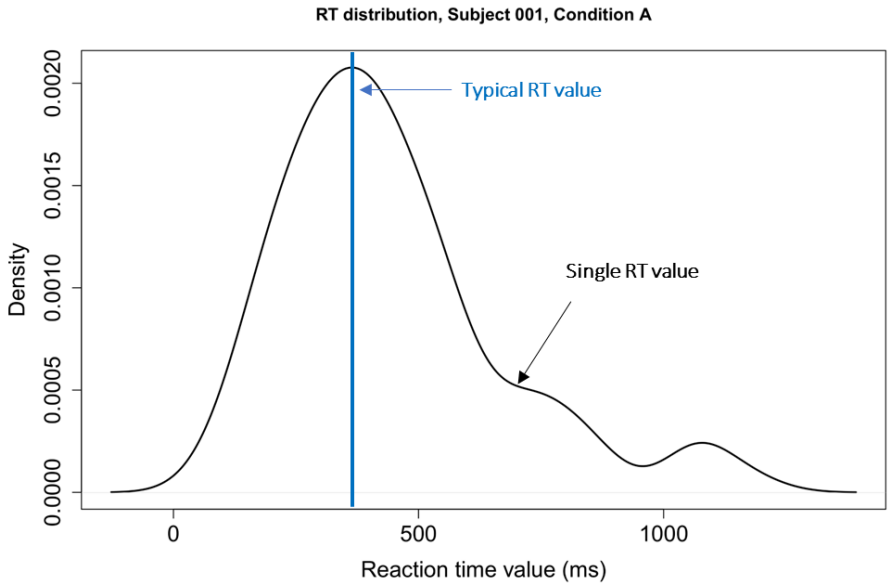

Figure 2. Distribution of reaction times for a hypothetical subject in one condition.

The efficiency of within-subjects design can be extended to **multiple trial designs**. That is, instead of merely assigning every subject to every condition once, the conditions are presented repeatedly (e.g.. 30 trials per condition), and the outcome is measured each time. For the resulting data, it is then common to average the outcome measurements across trials within each condition, yielding a good "typical" estimate for that subject in that condition. This is attractive especially for measurements that have high variance (low precision), such as reaction times, where a single trial would yield a very unrepresentative outcome value, while an average across trials may coincide with a representative typical outcome value (see Fig. 2 for an example).

Alternatively, one can run multiple trials and, instead of averaging the values, use them directly in a multilevel regression. This has other advantages, such as maximizing the available information, taking into account trial-variability (which is otherwise lost with averaging), and the ability to estimate individual differences in experimental effects. A question that comes up in this context is how many trials are needed for such estimation, or how the number of trials influences the power of the overall analysis. Although it is difficult to give precise guidelines for this, between 5 and 10 trials is a good start, and more desirable. For purposes of power and generalizability, however, one should always

strive to have many subjects rather than having many trials, although such balances partly depend on the phenomenon being measured (see my workshop on multilevel regression for more discussion).

A final caution is that not all tasks benefit from multiple trials. For example, subjects become desensitized to emotional stimuli after repeated exposure, making the resulting average across trials less typical/representative, than after one or two exposures.

Distributional assumptions

When we analyze data from a within-subjects design, it is done most often with methods like the paired *t*-test, or repeated measures (M)ANOVA. These methods technically analyze difference scores (e.g., $Y_{T1} - Y_{T2}$ values). According to the central limit theorem in statistics, a sum/difference of independent and identically distributed values will be normally distributed, as the number of values being summed/differenced increases to infinity. However, the normal approximation will be quite good even for low such numbers (e.g., 20, 30).

A difference of just 2 values would seem insufficient to guarantee this approximation, but it is a start, and again this can be improved in a multiple-trial design. If each condition has 30 trials, the resulting averages across trials will likely be normally distributed, and hence subsequent mean differences between two conditions will become a difference of two normally distributed values. Such differences are themselves likely to be normally distributed.

Conclusions

In sum, within-subjects designs have numerous advantages over between-subjects designs, such as stronger causal inference, improved power, and improved normality convergence. When combined with a multiple-trial approach, these designs can be especially effective at reliable estimation of effects. However, they do have limitations and cautions (e.g., order effects) and the choice against a between-subjects design should be decided on a case by case basis.

Best,
Ben

References

- Maxwell, S.E., & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd Edition)*. Mahwah, NJ: Larwence Erlbaum Associates.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

--

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