POINTS OF SIGNIFICANCE

The standardization fallacy

"We demand rigidly defined areas of doubt and uncertainty!" —D. Adams

Bernhard Voelkl, Hanno Würbel, Martin Krzywinski and Naomi Altman

A popular notion about experiments is that it is beneficial to reduce subjects' biological and environmental variability to mitigate the influence of confounding factors on the response. The argument is that by keeping the levels of such factors fixed — a process called standardization — we increase precision by limiting the component of response variance that is not due to the experimental treatment. Unfortunately, although standardization increases power, it can also induce such unrealistically low variability that the results do not generalize to the population of interest and may thus be irreproducible — the so-called "standardization fallacy"[1]. This month, we show how to avoid this fallacy by balancing standardization, which increases power to detect an effect but reduces external validity, with controlled heterogenization, which may reduce power but increases external validity.

Suppose we wish to test the effect of a treatment factor X_1 (for example, a drug) on some physiological response of an organism (for example, a mouse) in the presence of two other factors X_2 and X_3 that interact with X_1 and whose effects are not of primary interest but should not be ignored. We'll write the response as $R = dX_1 + X_2 + X_3 + X_1X_2 + \beta X_1X_3 + \varepsilon$, where d is the treatment effect and ε is random error. To account for the X_1X_2 interaction[2] in the analysis, subjects will be grouped into blocks, each with a fixed level of X_2, and assigned randomly to control and treatment within a block[3].

Let's assume that X_3 is continuous, as is the case for most confounding variables. Relatively few are discrete, and those that are, such as lab or batch of reagent, often represent a proxy for a large number of unmeasured (often unknown) continuous covariates. How can we deal with this factor and its potential interaction with X_1 (when $\beta \neq 0$)?

First, we can use an agnostic design (AGN) that simply ignores X_3 and relies on randomization to balance its effect across treatment and blocks (Fig. 1b). Here, X_3 will vary unpredictably and its effect will be part of random unexplained variation, which is now the sum of variance of X_3, βX_1X_3 and ε. This increase in variance reduces power[4] to detect a treatment effect (Fig. 2a).

Second, we can standardize X_3 by keeping its level constant (Fig. 1c) so that it no longer contributes to response variation (Fig. 2a). Power will be increased but the variation may be artificially low. For example, observations from mice of a given strain or age or kept under specific housing conditions may not apply to all mice. The design may still be internally valid but is no longer externally valid. If we ignore this in favor of increased power, we risk falling victim to the standardization fallacy.

Critically, this standardized (STD) design does not allow us to determine whether X_3 interacts with X_1. In the absence of an interaction, the choice of $X_3 = k$ does not affect power or the observed effect. However, if interaction is present, the observed effect $\varphi = d + \beta k$ will be a biased estimate of d and depend on both β and k (Fig. 2b). Because STD removes variability within experiments but does nothing to mitigate its effects across experiments (or labs), reproducibility of the observed effect is poor.

Third, we can use one of two heterogenized designs in which X_3 is systematically varied: randomized complete block (RCB) or fully crossed factorial (FCF) design[2]. In RCB (Fig. 1d), two levels of X_3 are selected and aligned with X_2 blocks. Statistical significance of the treatment effect is determined by comparing the variation between treatment and control groups to within-group variation within blocks. Because RCB accounts for variance of X_2 and X_3, it has higher power than AGN.

However, RCB cannot untangle the effects of the X_1X_2 and X_1X_3 interaction terms because the levels of X_2 and X_3 are not sampled independently. This can be achieved in a FCF design (Fig. 1e), which can decompose variance into components attributed to each factor as well as any interactions and, given replicates within a block, can isolate and measure unexplained residual variation. Because each block requires more subjects, contributing to a reduction in power unless sample size is increased[4], the number of factors that can be controlled by using FCF is usually limited by economic, ethical and logistic costs.

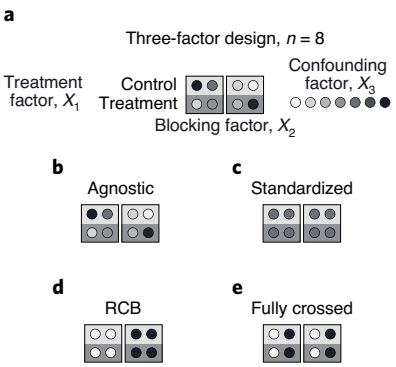

Fig. 1 | A confounding factor can be ignored, standardized or heterogenized. a, In an experimental design that compares the response to a treatment factor X_1 in control and treatment groups (rows) in the presence of a blocking factor X_2 (columns), a third confounding factor X_3 with continuous levels (circle fill) can be either **b**, ignored and subject to random variation (agnostic), **c**, standardized to a fixed laboratory-specific level, or heterogenized with two (or more) levels either **d**, aligned with X_2 (randomized complete block, RCB) or **e**, fully crossed (independently sampled) with X_2.

We will illustrate the implications of these designs and the consequences ofthe standardization fallacy by exploring how power and reproducibility vary with design and treatment effect size and how the observed effect φ relates to the underlying treatment effect d. We will explore scenarios with $d > 0$ and use one-sided tests that require $\varphi > 0$ to avoid considering opposite observed effects (negative) as significant. However, since in a real experiment we would not know the direction of the true effect, we will also keep track of φ^{ns}, $\varphi_-{}^\star$ and $\varphi_+{}^\star$ to indicate effect estimates that are non-significant, significant and negative (opposite to d) or significant and positive, respectively.

Let's simulate 25,000 experiments with sample size $n = 8$ across a range of effect sizes $d = 0$–5 in the absence ($\beta = 0$) and presence ($\beta = -2$) of an interaction between

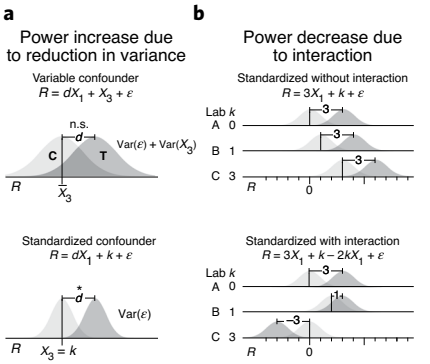

Fig. 2 | Standardization increases power but leads to higher variation between labs if an interaction between treatment and confounder is present. a, When X_3 with variance Var(X_3) is fixed ($X_3 = k$), power to detect main effect d is increased because variance of control (C) and treatment (T) response (R) is reduced. **b**, In the absence of an X_1X_3 interaction ($\beta = 0$), power is unchanged across labs that standardize X_3 differently (for example, $X_3 = k = 0, 1$ or 3) and all three labs observe the true treatment effect $d = 3$. With interaction ($\beta = -2$), the observed effect will vary on average by $\beta X_1X_3 = -2k$, so the observed effect may now be negative. Differences in k result in inconsistent power and decreased reproducibility.

X_1 and X_3. We use $X_1 = -0.5$ and 0.5 for control and treatment, $X_2 = -0.5$ or 0.5 for blocks; this maintains a zero mean and unit difference between levels. In AGN, X_3 is sampled from a standard normal distribution. In STD, a fixed value of X_3 is selected randomly from one of the values in AGN. In RCB, the two levels of X_3 are determined by the minimum and maximum values in AGN. Because FCF performs similarly to RCB, albeit with lower power, we will not consider it further. Finally, random error ε is sampled from the standard normal distribution.

We will use linear regression (analysis of variance, or ANOVA) to fit R and determine the significance ($\alpha = 0.05$) and the magnitude and direction of the estimate φ of the treatment effect. Power, P, will indicate the probability of a significant positive estimate (φ_+^*), which corresponds to rejecting the null hypothesis that there is no positive treatment effect. Reproducibility will be the probability of making the same inference (rejecting or not rejecting the null) twice.

Without interaction between treatment X_1 and confounder X_3 ($\beta = 0$), the STD and RCB designs perform equivalently because the impact of X_3 on the control and treatment groups is the same and cancels out (Fig. 3a). As expected, the AGN design overall has the lowest power because variation in X_3 is not controlled and adds to unexplained residual variation (Fig. 3a).

In the presence of an interaction, differences in power to detect $\varphi > 0$ arise because now the effect of X_3 no longer cancels out (it depends on the treatment and control via βX_1X_3). The larger the magnitude of β (or variance of X_3), the greater the impact of the interaction on power. As before, AGN has the lowest power for all d in our range, but now STD has higher power than RCB — but only for $d < 1.8$. This can be explained by elimination (in STD) or reduction (in RCB) of variance due to X_3. For insight into why STD has higher power than RCB for low d but lower power at high d, let's look at the distributions of the observed effect, φ, for each design (Fig. 3b).

At $d = 0$, significant estimates of either sign (φ_-^*, φ_+^*) are false positives. STD has the highest power ($P = 0.20$) because it has the largest variance in φ, making it more likely to cross the significance threshold by grossly under- or overestimating the actual effects — thereby illustrating the standardization fallacy. It produces the most inflated effect size estimates, with an average φ_+^* of 2.82.

For a small treatment effect, $d = 1$, STD still has the highest power ($P = 0.35$) and continues to misestimate the effect size and its direction, since negative values of φ still occur at substantial rates. Although AGN and RCB have lower power, they almost never yield $\varphi < 0$, so the directions of the estimated effects are consistent with d among experiments.

As the treatment effect size d increases, the distributions of φ continue to shift and narrow. Simultaneously, the bias in the estimate of d decreases: the average of φ_+^* gets closer to true d. However, both the width and the bias of the φ_+^* distribution decrease very slowly in STD, making it perform poorly. Note that even at large $d = 4$, STD results in a strong right skew in φ_+^*.

Of the three designs, in the presence of interaction, RCB has an excellent balance of high power and low bias (the average of φ_+^* is closest to true d). Importantly, the reproducibility is better with RCB than STD

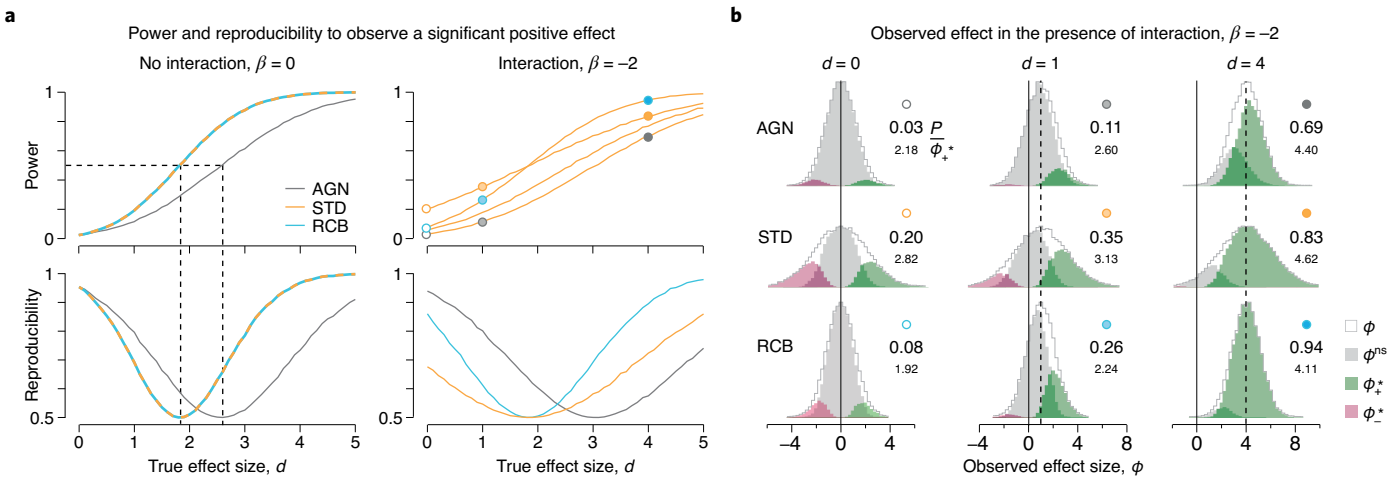

Fig. 3 | Confounder interaction can increase power to detect a treatment effect but reduce reproducibility. a, Power and reproducibility profiles for AGN, STD and RCB designs for true treatment effects $d = 0$–5 and sample size $n = 8$. **b**, Distributions of the observed effect φ (gray line) for true effect $d = 0$, 1 and 4 for each design in the presence of interaction. Histograms with solid fill color indicate distributions of non-significant (φ^{ns}, gray), significant and negative (φ_-^*, magenta) and significant and positive (φ_+^*, green) estimates. Also shown are power P (ratio of areas of φ_+^* and φ) and the mean of φ_+^*.

for all d (Fig. 3a). If we increase our sample size to $n = 16$, FCF's power catches up to RCB's and FCF becomes a viable alternative if we wish to measure the interaction. However, because a larger sample will yield higher power, we expect to see more false positives at $d = 0$.

In summary, by incorporating variation of confounding factors through controlled heterogenization, we can avoid the standardization fallacy and improve reproducibility. The magnitude of confounding effects can be analyzed with fractional factorial[5] or fully crossed designs[2]. However, as more confounders are added, the number of blocks (and hence subjects) increases quickly, practically restricting this approach to scenarios with only a few confounders. In the presence of an interaction between a confounder and treatment, heterogenization (RCB or FCF) is more effective than lab-specific standardization in detecting small treatment effects. Furthermore, when the experiment is already run as a RCB (for example, in batches), further heterogenization factors can be introduced with no need to increase to sample size. Even when heterogenization requires more blocking and larger samples, the higher external validity and improved reproducibility will often outweigh the extra effort and costs of introducing more heterogenization factors[6]. ❐

Bernhard Voelkl[1], Hanno Würbel[1], Martin Krzywinski[2] ✉ and Naomi Altman[3]

[1]Veterinary Public Health Institute, University of Bern, Bern, Switzerland. [2]Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada. [3]Department of Statistics, The Pennsylvania State University, State College, PA, USA.

✉*e-mail: martink@bcgsc.ca*

Published online: 6 January 2021
https://doi.org/10.1038/s41592-020-01036-9

References

1. Würbel, H. *Nat. Genet.* **26**, 263 (2000).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 1187–1188 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
5. Smucker, B., Krzywinski, M. & Altman, N. *Nat. Methods* **16**, 211–212 (2019).
6. Voelkl, B. et al. *Nat. Rev. Neurosci.* **21**, 348–393 (2020).

Competing interests

The authors declare no competing interests.