BNL-221600-2021-JAAM

The importance of independent replication of treatments in plant science

A. Rogers

To be published in "Journal of Experimental Botany"

June 2021

Environmental and Climate Sciences Department

Brookhaven National Laboratory

U.S. Department of Energy

BNL Program Development

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

eXtra Botany

Editorial

The importance of independent replication of treatments in plant science

Alistair Rogers[1] Karl-Josef Dietz[2], Miriam Gifford[3], John E. Lunn[4]

[1] Environmental Sciences Department, Brookhaven National Laboratory, Upton NY 11973-5000, USA

[2] Biochemistry and Physiology of Plants – W5-134, Bielefeld University, 33501 Bielefeld, Germany

[3]School of Life Sciences and Warwick Integrative Synthetic Biology Centre, Gibbet Hill Road, University of Warwick, Coventry, CV4 7AL, UK

[4]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

Correspondence: arogers@bnl.gov

Name	ORCID
Alistair Rogers	0000-0001-9262-7430
Karl-Josef Dietz	0000-0003-0311-2182
Miriam Gifford	0000-0002-4005-2513
John Lunn	0000-0001-8533-3004

Running header: independent replication

Key words: experimental design, interspersion, power, pseudoreplication, randomization

Word count: 2,159

Among other foci, the *Journal of Experimental Botany* aims to advance understanding of plant environment interactions, including abiotic stress, mineral nutrition, and the response of plants to global change. Advancing understanding in these areas often requires manipulation of the growth environment. Experiments range in scale and include: growth chambers (Yiotis *et al.*, 2020), glasshouses (Rho *et al.*, 2019), whole plant chambers (Sharwood *et al.*, 2017), and open-air field experiments (Ruiz-Vera *et al.*, 2020). The application of environmental manipulations can be constrained by available resources, operational costs, and time. In designing experiments, researchers attempt to balance logistical challenges with the desire to address scientific questions and maximise statistical power.

Unfortunately, by design or default, pseudoreplication (Hurlbert, 1984) remains a problem in plant science. The *Journal of Experimental Botany* expects independent replication of experimental treatments, randomized experimental designs, and use of appropriate analytical approaches. Here we highlight the issue for our readers, reviewers, and authors; and remind authors that a full description of the experimental design and statistical methods should be included in research papers. This is now aided by the journal's new policy of excluding the materials and methods section from the word count of research papers.

What is pseudoreplication and why is it a problem?

Pseudoreplication is the collection of what the researcher considers to be independent samples from different treatments that they wish to compare, when in fact the samples are not independent because they come from a single experimental unit (Davies and Gray, 2015). Pseudoreplication is a problem because: (1) random effects or events affecting just one treatment can lead to incorrect conclusions and, (2) the treatment of subunits as independent experimental units artificially inflates the statistical significance of numerical differences (Hurlbert, 1984; Lindroth and Raffa, 2017; Silk *et al.*, 2020).

Identifying the experimental unit

The experimental unit is the unit to which the treatment has been applied (Box 1). Correctly identifying the experimental unit is critical, but it is not always intuitive (Lindroth and Raffa, 2017). The nature of the experimental unit can change with scale or experimental design. To illustrate this point, we provide three examples, where the experimental unit is either a field plot (Fig. 1A), a growth chamber (Fig. 1B) or a plant (Fig. 1C). Terms such as replicate, repeats, technical replicates, biological replicates, semi-biological replicates, treatment replicates, measurements, or samples are often used to describe experimental design but have ambiguous meanings (Cumming *et al.*, 2007; Vaux *et al.*, 2012; Zar, 1999). For example, in Fig. 1B the biological replicate is a sub-unit, but in Fig. 1C it is the experimental unit. The statistical evaluation of a treatment effect, and the presentation of error bars associated with a treatment, should be at the level of the experimental unit (Cumming *et al.*, 2007; Hurlbert, 1984; Vaux *et al.*, 2012).

Random effects and events

At the simplest level there are two basic statistical requirements for a good experiment, replication and randomization, or more accurately, interspersion (Hurlbert, 1984; Mead *et al.*, 2003). The hypothetical validity of using unreplicated treatments is dependent upon the experimental units being identical at the time of manipulation and remaining identical throughout the duration of the experiment, except for differences resulting from the manipulation (Hurlbert, 1984). Meeting these criteria is necessary to ensure that measured effects in an unreplicated experiment can be attributed to the manipulation, and not to random variation between experimental units. For this reason, such validation is practically impossible if n=1 and a treatment effect is indistinguishable from random effects (Cottenie and De Meester, 2003; Potvin and Tardif, 1988; Ramage *et al.*, 2013; Vaux *et al.*, 2012). Replication of the experimental unit, and interspersion—the distribution of experimental units in space—provide the best insurance against chance events producing spurious effects that could be attributed to the treatment or manipulation (Hurlbert, 1984).

Consider a field experiment (Fig. 1A) with replicated (n=5) but undispersed experimental units (plots). In this example all the plots receiving treatment level 1 are located at the north end of the field and all those receiving treatment levels 2 or 3 are located in the middle and the south end of the field, respectively. The problem with this experimental design is that unknown differences in land use history or drainage between the north and the south end of the field could lead to a spurious treatment effect. Even if the field had no pre-existing variation, events occurring during the experiment could impact the treatments unevenly. For example, overspray of irrigation or a pesticide from an adjacent field to the north will affect plots used for treatment 1 to a much greater extent than those used for treatments 2 and 3. The problems with this hypothetical experimental design are well recognized and easily mitigated with alternative designs, e.g. a randomized block (Mead *et al.*, 2003). Fortunately, the experimental design depicted in Fig 1A is rarely encountered in field trials. However, an extreme form of this kind of simple segregation often arises when plants are grown in controlled environments where logistical constraints begin to exert an influence on experimental design.

Consider a growth chamber experiment (Fig. 1B) where the effect of elevated carbon dioxide concentration ($[CO_2]$) is investigated using plant growth chambers set at three different $[CO_2]$s. The experimental design described in Fig. 1B is an example of what Hurlbert referred to as isolative segregation (Hurlbert, 1984), and is unfortunately common in plant science. In this example the researchers have taken care to set the environmental conditions of the chambers to be the same, except for the CO_2 treatment, but the potential for spurious effects resulting from uncontrolled or unmeasured effects remains. As with the research field example (Fig. 1A), spurious effects can manifest as pre-existing differences between chambers; e.g. prior to this experiment chamber 1 was used by another researcher for a plant pathogen experiment and not properly cleaned after use, the air intake for chamber 2 is adjacent to a piece of equipment that generates ozone, a researcher installed different bulbs in chamber 3 that do not have the same spectral qualities as those installed in chamber 1 or 2. Differences can also emerge during an experiment, e.g. a reduction in irradiance in chamber 1 due to a bulb failing, or contamination of

a CO_2 tank with ethylene in chamber 2 (Morison and Gifford, 1984). Since all the plants for a given treatment are located in one chamber, pre-existing or emergent effects that are unique to one chamber are guaranteed to affect all of the plants experiencing the treatment assigned to that chamber, i.e. isolating and segregating the treatments exacerbates the dangers of simple segregation (Hurlbert, 1984). Furthermore, isolation in a single chamber increases the risk of Type I errors (Box 1) because isolation promotes uniformity among the plants within the chamber, reducing variance and increasing statistical power.

Note that the principal of interspersion is also important to consider when evaluating genetic modifications. Segregation and isolation of transgenic plants from their wildtype controls would confound the ability to confidently attribute an observed phenotype to the genetic manipulation, and it is therefore critical to grow transgenic plants in the same chamber as their wildtype control plants.

Statistical considerations

In statistics a Type I error occurs when a test falsely identifies a significant difference between treatments when there is no real difference. The specified significance level (α) determines the probability of committing a Type I error, i.e. when $\alpha=0.05$ we are accepting the probability of incorrectly identifying a difference between treatments 5% of the time. A Type II error occurs when a test fails to identify a difference between treatments when it actually exists; the probability of committing a Type II error is ß (Box 1). Higher probabilities of committing a Type I error are associated with lower probabilities of committing a Type II error and *vice versa* (Mead *et al.*, 2003; Zar, 1999).

The statistical power (1-ß) is the probability of correctly identifying a difference between treatments for a given α. Power is determined by the following parameters; the difference between the treatment means, the pooled variance of the treatments, the number of levels of the treatment (k), and the number of experimental units (n) (Cohen, 1988; Cottingham *et al.*, 2005; Zar, 1999). The resulting degrees of freedom associated with the treatment, commonly referred to as the group error degrees of freedom (v_1, k-1), and the experimental units, commonly referred to as the error degrees of freedom (v_2, $k(n$-1)) also influence statistical power. Statistical power increases notably when the magnitude of the treatment effect increases, or when variance decreases, and importantly as n and the associated v_2 increase. This was presented graphically by Pearson & Hartley (1951) and reproduced by Zarr (1991). Therefore, incorrectly treating sub-units as experimental units falsely inflates statistical power and dramatically increases the likelihood of Type I errors. For these reasons, authors are expected to clearly identify the experimental unit, and report n, and the degrees of freedom associated with their experimental design and statistical analysis.

In our growth chamber example (Fig. 1B) the plants within each chamber are not independently subjected to the CO_2 treatment and for the purposes of statistical analysis cannot be considered as true experimental replicates (Box 1). In this example the unit of replication is the chamber and

n=1 for each level of the treatment (Cumming *et al.*, 2007; Johnson *et al.*, 2016; Lindroth and Raffa, 2017; Wernberg *et al.*, 2012).

Common approaches to address issues of replication and interspersion

It is common for treatments to be rotated between growth chambers during an experiment as a means to reduce the potential for spurious chamber effects (Drag *et al.*, 2020; Yiotis *et al.*, 2020). For example, researchers running the experiment described in Fig. 1B may attempt to equalize unintended chamber effects across the three CO_2 treatments by rotating the plants, and their CO_2 treatment, among the three chambers during the experiment. This action may indeed reduce the possibility for spurious chamber effects, but it does not eliminate it as the effect of variation among chambers may have a temporal component (Potvin and Tardif, 1988). This could be addressed by staggering the treatments so that, in our example, all treatments are in each chamber for the same stage of growth (Johnson *et al.*, 2016). However, there is still no control for what Hurlbert terms ephemeral events or demonic intrusion (Hurlbert, 1984). However, whilst chamber swapping may serve to minimize the chance of chamber effects impacting the results, the practice in no way guarantees it and does not address the central problem that the treatments are not independently replicated (Hurlbert, 1984; Johnson *et al.*, 2016).

Faced with limited resources for plant growth, repeating experiments in time and randomly assigning the treatment to growth chambers is one approach that can be used to generate independent replication (Johnson *et al.*, 2016). When replication of treatments in time or space is not possible, alternative designs can be considered. A regression approach can be used to analyse the response to continuous factors (Cottingham *et al.*, 2005). For example, Drag et al. (2020) used a regression approach to analyse the effect of an unreplicated gradient of [CO_2] spanning eight treatment levels. Mixed effect models can account for pseuodoreplication by the use of random effects that take into account the hierarchical nature of the experimental design (Harrison *et al.*, 2018), however mixed effect models are not a cure for pseudoreplication resulting from poor experimental design (Silk *et al.*, 2020) and can only help when a treatment has been repeated in more than one experimental unit (Johnson *et al.*, 2016). In short, you cannot analyze your way out of a poorly designed experiment.

Is pseudoreplication a pseudo issue?

There has been considerable debate over the of dogmatic identification of pseudoreplication and editorial treatment of work where it is identified (Cottenie and De Meester, 2003; Davies and Gray, 2015; Hurlbert, 2004; Oksanen, 2001, 2004). This debate has been focused on opportunistic experiments resulting from natural disturbance, or landscape scale phenomena. In their contribution to the debate, Davies & Gray (2015) emphasize the importance and value of publishing such papers, but advocate for a cautious approach when reporting results from unreplicated natural experiments and for open acknowledgement by the authors when pseudoreplication is an issue (e.g. Eastman et al 2021). However, they also highlight the genuine

problem of simple pseudoreplication – such as that presented in Fig. 1B – in designed experiments (Davies and Gray, 2015).

Key conclusion

The credibility of a study is fundamentally determined by the experimental design (Christie et al. 2020). For that reason *JXB* expects independent replication of experimental treatments and also requires a full description of the experimental design and the statistical approach so that reviewers and readers can properly understand how an experiment was replicated and analysed.

Acknowledgements

AR was supported by the United States Department of Energy contract No. DE-SC0012704 to Brookhaven National Laboratory. MG was supported by a Biotechnology and Biological Sciences Research Council (BB/T015357/1) grant. JEL was supported by the Max Planck Society.

References

Cohen J. 1988. *Statistical power analysis for the behavioural sciences*. Hilldale, New Jersey: Lawrence Erlbaum Associates.

Cottenie K, De Meester L. 2003. Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). Oikos **100**, 394-396.

Cottingham KL, Lennon JT, Brown BL. 2005. Knowing when to draw the line: designing more informative ecological experiments. Frontiers in Ecology and the Environment **3**, 145-152.

Cumming G, Fidler F, Vaux DL. 2007. Error bars in experimental biology. Journal of Cell Biology **177**, 7-11.

Cristie AP, Abecasis D, AdjeroudM, Alonso JC, Amano T, Anton A, Baldigo BP, Barientos R, Bicknell JE, Buhk DA et al. 2020. Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. *Nature Communications*. **11,** 6377.

Davies GM, Gray A. 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). Ecology and Evolution **5**, 5295-5304.

Drag DW, Slattery R, Siebers M, DeLucia EH, Ort DR, Bernacchi CJ. 2020. Soybean photosynthetic and biomass responses to carbon dioxide concentrations ranging from pre-industrial to the distant future. Journal of Experimental Botany **71**, 3690-3700.

Eastman BA, Adams MB, Brzostek ER, Burnham MB, Carrara JE, Kelly C, McNeil BE, Walter CA, Peterjohn WT. 2021. Altered plant carbon partitioning enhanced forest ecosystem carbon storage after 35 years of nitrogen additions. *New Phytologist*. doi 10.1111/nph.17256

Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CE, Robinson BS, Hodgson DJ, Inger R. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. Peerj **6**, 32.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological fiedl experiments. Ecological Monographs **54**, 187-211.

Hurlbert SH. 2004. On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. Oikos **104**, 591-597.

Johnson SN, Gherlenda AN, Frew A, Ryalls JMW. 2016. The Importance of Testing Multiple Environmental Factors in Legume-Insect Research: Replication, Reviewers, and Rebuttal. Frontiers in Plant Science **7**, 4.

Lindroth RL, Raffa KF. 2017. Experimental approaches for assessing invertebrate responses to global Oxford: John Wiley & Sons, Ltd., 30-45.

Mead R, Curnow RN, Hasted AM. 2003. *Statistical methods in agriculture and experimental biology*. London: Chapman & Hall/CRC.

Morison JIL, Gifford RM. 1984. Ethylene contamination of CO_2 cylinders - effects on plant growth in CO_2 enrichment studies. Plant Physiology **75**, 275-277.

Oksanen L. 2001. Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos **94**, 27-38.

Oksanen L. 2004. The devil lies in details: reply to Stuart Hurlbert. Oikos **104**, 598-605.

Pearson ES, Hartley HO. 1951. Charts for the power function for analysis of variance tests, derived from the non-central *F*-distribution. *Biometrika*. **38,** 112-130.

Potvin C, Tardif S. 1988. Sources of variability and experimental designs in growth chambers. Functional Ecology **2**, 123-130.

Ramage BS, Sheil D, Salim HMW, Fletcher C, Mustafa NZA, Luruthusamay JC, Harrison RD, Butod E, Dzulkiply AD, Kassim AR, Potts MD. 2013. Pseudoreplication in Tropical Forests and the Resulting Effects on Biodiversity Conservation. Conservation Biology **27**, 364-372.

Rho H, Doty SL, Kim S-H. 2019. Endophytes alleviate the elevated CO2-dependent decrease in photosynthesis in rice, particularly under nitrogen limitation. Journal of Experimental Botany **71**, 707-718.

Ruiz-Vera UM, De Souza AP, Ament MR, Gleadow RM, Ort DR. 2020. High sink strength prevents photosynthetic down-regulation in cassava grown at elevated CO2 concentration. Journal of Experimental Botany. **72,** 542-560.

Sharwood RE, Crous KY, Whitney SM, Ellsworth DS, Ghannoum O. 2017. Linking photosynthesis and leaf N allocation under future elevated CO2 and climate warming in Eucalyptus globulus. Journal of Experimental Botany **68**, 1157-1167.

Silk MJ, Harrison XA, Hodgson DJ. 2020. Perils and pitfalls of mixed-effects regression models in biology. Peerj **8**, 20.

Vaux DL, Fidler F, Cumming G. 2012. Replicates and repeats-what is the difference and is it significant? A brief discussion of statistics and experimental design. Embo Reports **13**, 291-296.

Wernberg T, Smale DA, Thomsen MS. 2012. A decade of climate change experiments on marine organisms: procedures, patterns and problems. Global Change Biology **18**, 1491-1498.

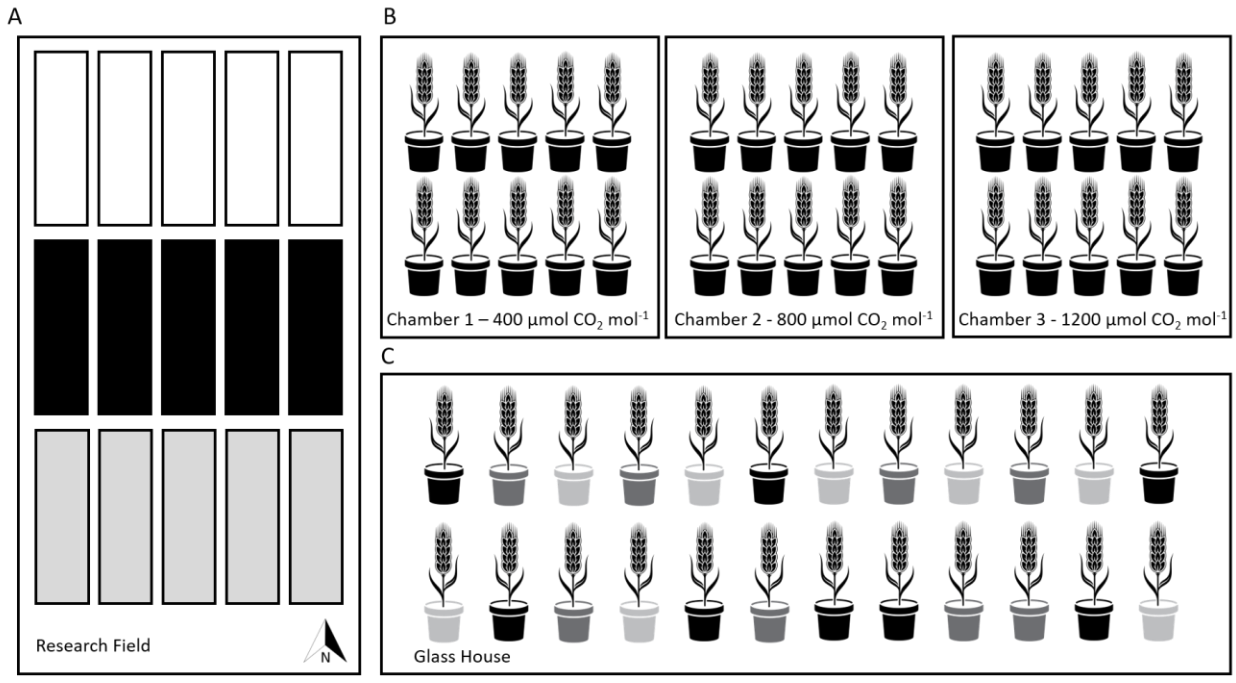
Yiotis C, McElwain JC, Osborne BA. 2020. Increased productivity of ryegrass grown at elevated CO2 is dependent on tillering and leaf area development rather than leaf-level photosynthesis. Journal of Experimental Botany.

Zar JH. 1999. *Biostatistical Analysis*. Upper Saddle River, New Jersey: Prentice-Hall Inc.

Figure legends

Figure 1 Experimental design

(A) shows a research field containing fifteen plots, each 5 x 15m plot contains 160 individual plants. Three levels of a treatment (1=white, 2=black & 3=grey) were independently applied to five experimental plots (n=5 experimental units). In this experimental design there is independent replication of the levels of the treatment, but no interspersion of those treatments throughout the research field e.g. all five of the experimental units associated with treatment level 1 are located in the north end of the research field. This experimental design has three levels of one treatment, five independently replicated experimental units per treatment and 160 sub-units (plants) in each experimental unit. (B) shows an experiment where three growth chambers, each with ten individual plants, are used to investigate the effect of elevated carbon dioxide concentration ($[CO_2]$) on growth. Growth conditions of irradiance, vapor pressure deficit and temperature have the same set points, potted seeds were randomly assigned to chamber 1, 2 or 3, and all plants were managed using an identical protocol. The treatment, $[CO_2]$, was the only planned difference between chamber 1 (400 µmol mol^{-1}), chamber 2 (800 µmol mol^{-1}) and chamber 3 (1200 µmol mol^{-1}). This experimental design has three levels of one treatment, there is one experimental unit (the chamber) per treatment level, and ten sub-units (plants) within each experimental unit. There is no independent replication of the treatment and no interspersion. **(C)** shows a glasshouse experiment to investigate the effect of three levels of a micronutrient, indicated by black, grey, and light grey pots. The researcher was careful to avoid segregation and randomized the location of the treatments within the glasshouse. This experimental design has three levels of one treatment, there are eight independently replicated experimental units (plants) and no sub-units. Note, there is still a risk of inter-dependent pseudoreplication in this design because although the eight replicates are randomly distributed, they may share a common nutrient solution (see example B4 in Hurlbert, 1984).



Box 1

Key definitions	
Experimental unit	The unit to which the experimental treatment has been applied
Experimental sub-units	Sub-units contained within the experimental unit that were not independently subjected to the experimental treatment.
Degrees of freedom	The number of independent pieces of information that are used to evaluate an estimate of a parameter
Type I error	Incorrectly identifying a difference between treatments when no real difference exists
Type II error	Failure to detect a difference between treatments when there is in fact a difference
Statistical power	The probability that a statistical test will correctly reject a null hypothesis when it is false
n	The number of experimental units to which the treatment has been independently applied

(Lindroth and Raffa, 2017; Mead *et al.*, 2003; Zar, 1999)