Experimental design: The joys and caveats of setting up a proper study framework

WEINKAUF, Manuel

Experimental design: The joys and caveats of setting up a proper study framework

Manuel F. G. Weinkauf

1Department of Earth Sciences, Université de Genève, Geneva, Switzerland

Experimental design is an important, but often neglected topic in micropalaeontology. Fortunately, quite a lot of literature exists to introduce this topic properly to natural scientists (Federer 1955, Hurlbert 1984, Quinn and Keough 2002, Dytham 2011). But why is experimental design in all its facets important? For starters, having a proper experimental design makes all ensuing quantitative analyses much more easy, because one does not have to tweak statistical methods for an ill-chosen or complicated setup that introduces unnecessary random terms and bias into the analyses. But the problem does not end there, because an inappropriate experimental design can lead to wrong results, incomparable studies, experiments that cannot be replicated, and many other problems. In this article, I will try to share the very basics of experimental design and show some of the problems and pitfalls that come with it.

Mensurative vs. manipulative experiments

When thinking about experiments, many people tend to think only of manipulative experiments, i.e. experiments where something (e.g. a population of benthic Foraminifera) is manipulated in the laboratory (e.g. cultured in water with different salinities). I believe this is part of the reason why many palaeontologists tend to think that experimental design need not concern them, since they only work with material sampled in the field without actively manipulating anything. Un-
Fortunately, this is wrong (Hurlbert 1984)! We indeed distinguish two types of experiments, and experimental design is important for both of them.

The first type, as already mentioned, are manipulative experiments. Those are all experiments which are performed by actively manipulating some of the variables involved, so that one has an actively chosen treatment. Often, manipulative experiments take place in laboratories, where organisms can be cultured in several vessels that are exposed to different environmental parameters (e.g. culturing populations at a range of different temperatures, salinities, light levels, etc.) (see Marschner et al. 2012, Schmidt et al. 2014, for examples of well-designed manipulative experiments). They can, however, also be performed by manipulating populations directly in situ (e.g. Nomaki et al. 2005).

Mensurative experiments, on the other hand, are experiments where samples are taken in nature across space and/or time and then analysed for similarities and differences. The ‘treatment’ is space or time, i.e. it is known that some environmental parameters changed in a spatial or temporal pattern, and assemblages or populations are investigated concerning their reaction to that change. This type of experiment is what the majority of micropalaeontological analyses entails (e.g. Aldridge et al. 2012, Weinkauf et al. 2014, Knappebusch 2016). Mensurative experiments are a bit more complicated to analyse, since they do not have a control treatment in the strict sense (as manipulative experiments do). They also have a larger error associated with the treatment variable because it is only measured (often with a non-negligible error due to temporal or spatial integration) instead of actively chosen by the experimenter. Therefore, although this is often forgotten, mensurative experiments often call for more robust statistical analyses than manipulative experiments to be meaningfully analysed (Dytham 2011).

**Sampling schemes**

The sampling scheme is the most integral part of any design, because it provides the raw material for manipulative experiments and comprises the entire dataset of mensurative experiments. In order not to make
things overly complicated, I will present the three most prevalent end-members here, which are equally useful for mensurative and manipulative experiments and for sampling in space (area) and time (sediment column) (Fig. 12).

**Random sampling scheme**

The random sampling scheme is the easiest and on first glance the most robust, because it takes samples at completely random locations. When the environment is very homogeneous, for instance a sediment column that comprises only one lithology or a mudflat area without changes in environmental parameters or sedimentology, this is the easiest design to follow. Sampling locations are depicted by a random generator and are thus ensured to not be biased by any pre-existing assumptions of the experimenter.

**Stratified random sampling scheme**

It can happen that previous knowledge of the sediment column or area to be sampled requires to modify the sampling scheme. For instance, one can imagine a sediment column that mainly consists of limestones but has thin intercalations of sandstone, or a mudflat that is crossed by a river which delivers coarser sediment and more nutrients to the area in its vicinity. In both cases there is a dominating environment (limestone/mudflat) and a rare environment (sandstone/river area). A completely random design runs the risk of by chance only or disproportionately often sampling the dominating environment and neglecting the rare environment. This can give a biased picture of the development of e.g. the microfossil community over time or space, because it can be assumed that the under-represented rare environment has a different assemblage than the dominating environment. The solution for this is the stratified random sampling. Here, it is made sure that each stratum (e.g. dominating environment and rare environment) is represented with the same number of samples but samples within each stratum are still distributed randomly in space and time.

**Systematic sampling scheme**

Systematic sampling is most often applied in micropalaeontology, because it is easily designed and fol-
Figure 12: Examples of three prevalent sampling schemes, shown on the example of sampling a geological sediment column. **Random sampling:** For a pure limestone assemblage, where the environment was stable over the entire time, three randomly distributed samples (black triangles) may be enough for analyses and do not bias the results. **Stratified random sampling:** When there are intercalations of sandstone in the limestone, the three original random samples (black triangles) would by chance all fall in the limestone facies, obscuring the short-term environmental changes. A stratified random sampling (black and red triangles combined) with two random samples per stratum gives a better picture of the true development over time by sampling equally in both lithologies. **Systematic sampling:** A systematic sampling every 1 m (black triangles) could lead to the conclusion that the sediment column represents only one environmental gradient, from sandy to calcareous, because each sample is a little bit more calcareous than the one before. In reality, there are three limestone–sandstone cycles, which are obscured because the sampling interval (1 m) shows interference with a naturally occurring pattern (c.1.25 m-cyclicity). A random sampling design (red triangles) would give a less biased picture here, but a systematic sampling with a much finer resolution may be the best choice to fully understand this sediment column.
allowed in the field. In systematic sampling, samples are taken in pre-defined distances to each other (e.g. every 1 m along a sediment column or every 1 by 1 m across a mud-flat). Under many circumstances, this scheme is not problematic and eliminates bias by the experimenter as effectively as the random schemes. However, the systematic sampling can lead to a large bias when the chosen sample distance interferes with a naturally underlying pattern. In Fig. 12, the sampling interval of 1 m would create a set of samples which show gradually more calcareous components and fewer sand content. This is because there is a real cyclicity with three limestone–sandstone cycles of c. 1.25 m thickness. In effect, each of the three systematic samples comes from another cycle, but because the cycles are a little bit thicker than the sampling interval, each sample is located a little bit further down in its cycle than the previous sample was in the previous cycle. In absence of further knowledge, only based on these three systematic samples, one may interpret the sediment column as one large sandstone–limestone trend (e.g. due to slow subsidence), when in effect it is composed of three limestone–sandstone cycles (e.g. indicating intervals of slow uplift phases terminated by rapid subsidence). Systematic sampling on that scale would thus lead to a completely wrong observation and interpretation of the real situation, where random sampling would at least show you that the pattern is rather complex and may warrant further investigation. In a spatial sampling, an equivalent example may be sampling in a mudflat in a 20 by 20 m grid, where there are waste-water exhausts at the coast in 20 m intervals. When by chance samples are taken in the drainage channels, assemblages observed there will indicate much more eutrophic conditions in the mudflat than is really the case.

**Laboratory setup**

There is a vast variety of more or less established experimental setups (for which analytical methods are well understood) for manipulative experiments, which are partly derived from the sampling schemes discussed above (Dytham 2011, Cornwall and Hurd 2015). The important thing about a proper laboratory setup
is first and foremost an unbiased distribution of treatment and control units. There are different types of controls, but the two most important are (1) the procedural control and (2) the experimental control (Dytham 2011).

The experimental control is what many scientists consider the only control in an experimental setup. It is a population, that is not exposed to the treatment (e.g. increased salinity) of the experimental group, but is instead cultured at natural conditions. This allows to evaluate, if the treatment had any effect on the investigated parameter, or if the control group shows the same patterns as the experimental group.

The procedural control is another type of control group that may be worthwhile when the treatment involves a lot of disturbances (e.g. specimens have to be taken out of the aquaria each day for measurements and then put back). In such cases, the stress from the manipulation alone can have an effect, independent of the actual treatment. A procedural control is a control group that is disturbed in the same way as the experimental group (e.g. taken out of the aquarium into measurement vessels each day) but otherwise cultured under the same conditions as the experimental control group.

When designing the setup of control and experimental groups, it is important to make sure that there are no confounding factors. For instance, having all the experimental groups stand closer to the window and all the control groups further in the room closer to the air conditioning in an experiment studying the effect of salinity levels can lead to an unobserved influence of light attenuation and temperature. It can thus lead to an observed difference between control and experimental group, that has nothing to do with salinity but is misinterpreted as such, because salinity was the only factor that was actively manipulated. The two most common base forms of proper experimental setup are the Latin square design and the randomized block design (Hurlbert 1984, Dytham 2011). In the Latin square design, all treatments are applied within one block, but the different treatments have different positions in each row to avoid confounding of treatments. In the randomized block design, treatments are equally distributed across blocks (e.g. incubat-
ors) but their position per block is randomized (Fig. 13).

**Latin square**

\[
\begin{array}{cccc}
A & B & C & D \\
B & C & D & A \\
C & D & A & B \\
D & A & B & C \\
\end{array}
\]

**Randomized block**

\[
\begin{array}{cccc}
A & B & C & D \\
B & C & D & A \\
C & D & A & B \\
D & A & B & C \\
\end{array}
\]

**Block 1**

\[
\begin{array}{cccc}
A & C & B & A \\
D & B & C & D \\
C & C & B & A \\
B & A & C & B \\
\end{array}
\]

**Block 2**

\[
\begin{array}{cccc}
A & B & C & D \\
B & C & D & A \\
C & D & A & B \\
D & A & B & C \\
\end{array}
\]

**Block 3**

\[
\begin{array}{cccc}
A & C & B & A \\
D & B & C & D \\
C & C & B & A \\
B & A & C & B \\
\end{array}
\]

**Block 4**

\[
\begin{array}{cccc}
A & B & C & D \\
B & C & D & A \\
C & D & A & B \\
D & A & B & C \\
\end{array}
\]

**Figure 13**: Examples of the two most common experimental setups. In the Latin square design all treatments (including controls) are set up in one block, and each row and each column contains exactly one of each treatment. The shown arrangement is indeed not optimal, because treatment ‘D’ occurs in the corner twice (in the first and last row). In the randomized block design the experiment is spatially segregated into blocks, and each block contains one of each treatment (the arrangement of treatments is different in each block).

**Replication**

Replication is a very controversial matter, and far too complex to be discussed here. Ever since Hurlbert (1984) published his devastating conclusion about extensive pseudoreplication in scientific experiments, experimental design was the most dreaded part of work for many scientists. Indeed, replication as expected by Hurlbert (1984) requires elaborate design, and Heffner et al. (1996) and Cornwall and Hurd (2015) indicate that the situation in scientific practice had not changed for the better since. There are two things to say about this here.

The first is, that Hurlbert (1984) is right in demanding proper replication of experiments. Some of the experimental setups he criticises as pseudoreplication (Hurlbert 1984, fig. 1), for instance simple or clumped segregation or no replication, can easily be avoided by just sticking to the guidelines outlined in this article and the literature here cited. Others may be harder to grasp for the reader. For instance, the fact that having only one incubator or aquarium per treatment—so that all control groups are in Incubator 1 and all experimental groups in Incubator 2—is isolative segregation (Hurlbert 1984, fig. 1) and not replication may be hard to understand. The simple reason is, that even modern technology does not work perfectly, and one of the incubators may not work properly, break entirely, or have in the past accidentally been contaminated with a substance that influences the experiment. This can
lead to erroneous results, and it is not only in the interest of proper replication but also of saving time in the case of catastrophic events (e.g. one incubator breaks down entirely, destroying the entire experiment if there is no replication in another incubator) to avoid this.

The second thing is, that there are those who argue that Hurlbert (1984) might be a bit too restrictive in his approach (Hargrove and Pickering 1992, Oksanen 2001). Especially when dealing with mensurative and manipulative field experiments (natural experiments), the high standards of replication demanded by Hurlbert (1984) can hardly be met, or if so only by strict spatial or temporal limitation of the study. In such cases, a large range of treatments that offers properly robust results takes priority over replication, if replication would mean that the potential range of treatments is so small that any potentially observable effect is ambiguous.

To make this clear, one should try to properly replicate ones experiments as far as possible and avoid pseudoreplication whenever possible. But the wish to replicate must never limit the experiment in a way that makes its results ambiguous. As Oksanen (2001, p. 33) put it: ‘Let us face it. If the concept pseudoreplication is used in the broader sense, including compound treatments, then all experiments are pseudoreplicated, though we do not always have enough information to understand how.’

References


