Affirmation of the Classical Terminology for Experimental Design via a Critique of Casella's *Statistical Design*

Stuart H. Hurlbert*

ABSTRACT

In many disciplines, basic and applied, a high frequency of errors of statistical analysis has been documented in numerous reviews over the decades. One insufficiently appreciated source of this has been the failure of statisticians, individually and collectively, to provide clear definitions for many of the terms they use—and failure to adhere to those definitions across time and across disciplines. The field of experimental design is one area where such problems have become acute. This essay documents that phenomenon via analysis of the terminology used in a recent text in that field, *Statistical Design* by G. Casella, but the problems identified are widespread and of ancient lineage. There exists a clearer, more consistent terminology, most of it well established more than half a century ago. Key issues are the tripartite structure of the design of an experiment, the need for experimental units to be physically independent of each other, the definition of *pseudoreplication*, and confusion about the meaning of split-unit designs. The problems identified seem to reflect a long-standing conflict between the classical, experiment-focused approach to design and the model-focused approach to the topic. Proponents of the latter have tended to stray from the classical terminology of experimental design, redefining terms in a somewhat casual fashion and thereby considerably confusing non-statisticians in particular. Wider understanding of these matters should lead to better textbooks, better teaching, and better statistical practice.

It is convenient to introduce a standard terminology.
–COX (1958, P. 2)

The users of statistics encounter a frustrating problem: statisticians seem inconsistent in the definitions they attach to certain words and in their use of symbols.
–URQUHART (1981)

Is the subject of statistics to lead to different terminologies in different areas of application? This reviewer suggests not. If this be accepted then the onus is on the latter-day workers, e.g., in psychology, to read the prior literature and try to follow usage or at the very least, give also the nomenclature that is standard to the statistics profession.
–KEMPTHORNE (1982)

Conceptual and inferential errors may arise because of vague and imprecise definitions and formulations.
–FEDERER (1993)

Unfortunately, the terminology for error reduction designs using the split-unit principle is not quite uniform.
–HINKELMANN AND KEMPTHORNE (2008)

In conclusion, reform and standardization of terminology in statistics, experimental design and sampling design is badly needed, is possible, and would improve statistical practice.
–HURLBERT (2009)

S. Hurlbert, Dep. of Biology, San Diego State Univ., San Diego CA 92182. Received 10 Oct. 2012. *Corresponding author (shurlbert@sunstroke.sdsu.edu).

Published in Agron. J. 105:412–418 (2013)
doi:10.2134/agronj2012.0392

Terminology in statistics, as in other fields, cannot be irrevocably fixed and evolves with the subject. Nevertheless, where there is a firmly established set of concepts and definitions encapsulating important concepts, it is regrettable to see these definitions abandoned without good reason. This seems to have happened frequently in treatises on the design of experiments, and again in the recent book *Statistical Design* (Casella, 2008).

This book rejects some of the widely used terms of experimental design, redefines others inappropriately or incompletely, and accepts other confusing usages long adopted by others. The purpose of this essay is to identify these conflicts and call attention to existing, more suitable terminology. The essay might serve as a supplement to *Statistical Design* for researchers and students who wish to use it but who are discomfited or confused by its terminology. Because *Statistical Design* is just one example from the large number of texts with similar terminological problems, this essay should have wider value as well.

Over the years a few statisticians have called for improving the clarity with which statistical concepts and methodologies are communicated to researchers and students. The aim has been to improve statistical practice. In particular, there are good grounds for defending the core classical terminology of experimental design and for arguing that this terminology is suitable for adoption by all disciplines in which experimental work is carried out.

On the other hand, many statisticians and scientists have not been much concerned with the lack of standardized terms and definitions for even core concepts. Statisticians have sometimes been heard to say, for example, that a common terminology is not important as long as researchers make clear how they define the terms they use or as long as they select the right model. This unkindness toward, especially, non-statistician users of the statistics literature is not often put into print. At the beginning of his classic monograph on factorial experiments, Yates (1935), however, did so in particular reference to the term *experimental unit*:

> "It has been thought better to retain the terminology of agricultural field experiments, rather than create a more generalized terminology which might be applicable to all experimental material. This course recommends itself the more in that workers in any field will in practice refer to their experimental unit by their appropriate names, so that some transposition of terms when passing from one field to another will always be necessary." (p. 183)

This "necessity" has never been demonstrated, and the sad consequences of nonstandardization are now widely evident in the high frequency of erroneous statistical analyses in many disciplines (for recent reviews see Hurlbert, 2009, 2013; Hurlbert and Lombardi, 2009). Perhaps Yates felt that acceptance of his ideas on factorial experiments outside of the agricultural sciences might be facilitated by positing a libertine spirit on terminology.

At least three reviews of *Statistical Design* have been published (Puntanen, 2008; Vahl, 2008; Verkuilen, 2010). These are brief, largely complimentary, and summarized elsewhere (Supplementary Material, Appendix 1). Five pages of errata for the book and a PowerPoint presentation for his short course based on *Statistical Design* may be found on Casella's website (http://www.stat.ufl.edu/~casella/).

THE PROBLEMS

Rather than use a more complex structure for my discussion, I have simply used the terms on which I comment as section titles.

Experimental Design vs. Statistical Design

The definition given by Fisher (1935, p. 2) of an experimental design as "the logical structure of the experiment," the sum of what we now call its treatment, design, and response structures (see below), has long served us well.

Casella follows his mentor (Federer, 1975, 1984, 1993; Federer and King, 2007) and other close colleagues (e.g., McCulloch et al., 1986) in using *statistical design* as an alternative label for experimental design. Something different might be implied, however, by his statement (p. 2): "Statistical design is about understanding where the variance comes from, and making sure that is where the replication is."

No good purpose is served by renaming the discipline of experimental design, and the original justifications given by Federer (1984) for terming it *statistical design* have not been widely accepted. Moreover, statistical design connotes the design of observational studies as well as experimental ones. Indeed, Casella presents (p. 61), as an example of an experiment, an *observational* study designed to look at the correlation of education (three levels), home environment (four levels), and country of origin (seven countries) with performance on a citizenship test. Casella calls this a "social science experiment." This example closely parallels an observational study that Raktoe et al. (1981, p. 8) presented as another example of a "factorial experiment."

Experiment

Casella provides no definition of *experiment*, which is a cause of some concern given the topic of his book and the examples above. One possible statement of the classical concept is the following, from Hurlbert (2004):

> "A manipulative experiment is an exercise designed to determine the effects of one or more experimenter-manipulated variables (= experimental variables or treatment factors) on one or more characteristics (= response variables) of some particular type of system (= the experimental unit). Its primary defining features are: (1) that the experimenter can assign treatments or levels of each experimental variable at random to the available experimental units; and (2) that there are two or more levels established for each experimental variable used."

Comparative experiment is a synonym for *manipulative experiment* and is the older term. *Comparative experiment* can be misleading, however (Hurlbert, 1984). Many observational or correlational studies have comparisons as their objective, as in Casella's "social science experiment" or in a comparison of the fish assemblages in three lakes. On the other hand, comparative observational studies do not involve "manipulations" of treatment factors, although they do make use of models, ANOVAs, and other statistical methods of every degree of complexity.

Treatment Structure, Design Structure, and Response Structure

Casella states that "there are two pieces to a design which we separate into Treatment Design and Experiment Design" (p. 18). The first he defines as "the manner in which the levels of treatments are arranged in an experiment." The second he defines as "the manner in which the randomization of experimental units to treatments is performed and how the data are actually collected" (p. 22). This terminology apparently was first put forward by Federer (1955, 1973) (Table 1).

Finney (1955) and Urquhart (1981), however, recognized a tripartite structure for the subject matter, distinguishing what Urquhart (1981) labeled *treatment design*, *experimental design*, and *response design*. Finney characterized the three aspects but without labeling them. *Response design* refers not to the data collection itself but to the formal plan for its collection. As Urquhart noted, "*response design* usually gets relegated to a secondary position in much of agriculture and biology," The claim is valid for most other disciplines as well. The matter is reflected well in the last column of Table 1. Milliken and Johnson (1984) altered the terminology to *treatment structure* and *design structure* but omitted to coin *response structure* or explicitly treat it as the crucial "third leg" of design that it is. Casella has followed that custom of lumping design structure and response structure together.

A formal definition of *response structure* is (Hurlbert and Lombardi, 2004; Hurlbert 2009; after Finney, 1955, and Urquhart, 1981)

> "the list of response variables to be measured and the sampling plan that specifies when, where, and on what components of the experimental unit one will make and record their observations and measurements."

Formalization of these three distinct and independent aspects of an experimental design should make clear that each requires explicit description, using well-defined terms, in the Methods section of any report of an experimental study. Each of the

Table 1. Concordance for terms used to designate the three aspects of the structure of an experimental design, compared with the terms used here (after Hurlbert and Lombardi, 2004; Hurlbert 2009).

Reference	Treatment structure	Design structure	Response structure
Finney's (1955) definitions	the set of treatments selected for comparison	the rules by which the treatments are to be allocated to experimental units	the specification of the measurements or other records to be made on each unit
Federer (1955, 1973)	treatment design	experimental design	(unlabeled, little discussed)
Urquhart (1981))	treatment design	experimental design	response design
Mead and Curnow (1983)	treatment structure	structure of the experimental units	(treated independently but without label)
Milliken and Johnson (1984, 2009)	treatment structure	design structure	(treated as a part of design structure)
Mead (1988)	treatment structure	experimental design	(treated independently but without label)
Hinkelmann and Kempthorne (1994, 2008)	treatment design	error-control design	observation design
Valiela (2001)	design of treatments	design of layout	design of response
Mead et al. (2003)	treatment structure	(no label assigned)	(treated independently but without label)
Federer and King (2007)	treatment design	experiment design	(treated as part of experiment design)
Casella (2008)	treatment design	experiment design	(treated as part of experiment design)

classical terms of experimental design carries strong implications for the types of models that could validly be used for statistical analysis of a data set. Recognition of the distinctness of the three aspects also aids resolution of some of the other terminological—and ultimately statistical—issues discussed below.

Experimental Unit

Casella aptly notes that "[p]erhaps the most important concept in statistical design is the experimental unit," and then defines this as "the unit (subject, plant, pot, animal) that is randomly assigned to a treatment" (p. 3). This definition is incomplete on two counts.

First, sometimes the experimenter assigns treatments in some manner other than random, e.g., via the systematic procedures like those of the early British agricultural experimenters. Doing so calls for caution in interpreting any subsequent statistical analyses, but the units remain properly designated as experimental units.

More importantly, Casella's definition neglects that the experimental unit is defined de facto not just by how the experiment is initiated but also by how it is managed and maintained for the duration of the experiment (Fisher, 1935; Cox, 1958; Hurlbert, 1984, 2009; Mead, 1988). Basing themselves on a discussion of the matter by Cox (1958, p. 2, 155), Kozlov and Hurlbert (2006) offered this definition of experimental unit:

> "The smallest system or unit of experimental material to which a single treatment (or treatment combination) is assigned by the experimenter **and** which is dealt with independently of other such systems under that treatment at all stages in the experiment at which important variation may enter. By 'independently' is meant that, aside from both receiving the same treatment, two systems or experimental units assigned to the same treatment will not be subject to conditions or procedures that are, on average, more similar than are the conditions or procedures to which two systems each assigned to a different treatment are subject."

They acknowledged that "[t]his may seem overly lengthy, but on the evidence of dozens of textbooks, a shorter definition seems incapable of making explicit the key critical elements of the concept." Following Cox (1958, p. 19–21), Mead (1988, p. 120–121), and Hurlbert (2009), and with a slight risk of redundancy, we might even add the following sentence to that definition: *Independence also requires that the experimental units be bounded or physically defined in such a way that what transpires on or in one experimental unit can have no effect on other experimental units*.

In an example concerning an experiment testing the effects of food type on the growth of fish maintained in groups in tanks (Casella, 2008, p. 4), Casella correctly points out that "the experimental unit is the tank, as the treatment is applied to the tank, not to the fish." But he then states that "if the experimenter had taken [each individual] fish in hand, and placed the food in the fish's mouth, then the fish would have been the experimental unit..."

To the contrary, the tank would still be the experimental unit, as classically conceived. This would be true regardless of whether all the fish in the tank were individually given the same food type or whether half the fish in the tank were individually fed one food type and half fed a second type. Likewise, if the experiment concerned the effects of an injected hormone on fish growth and fish were individually injected but then housed in groups in tanks, the tank would still be the experimental unit, whether all fish in the tank were injected with the same hormone or not. The physical conduct of the whole experiment defines the experimental unit and determines the model, not vice versa. Both biologist and statistician should know that strong interactions among fish housed in the same tank are likely. In that case, the individual fish cannot be treated as independent experimental units, whole or sub, regardless of the "independent" procedures or operations involved in early stages of the experiment.

The need to avoid physical interaction among experimental units if they are to be considered independent replicates of a treatment and if biased estimates of treatment effects are to be avoided has been repeatedly clarified in the literature (e.g., Cox, 1958, p. 19–21, 1961; Federer 1975, 1993, p. 280; Mead, 1988, p. 12, 119–122; Wiley, 2003; Hurlbert, 2009). Those authors do not agree with Casella. Consider three examples of that disagreement.

Cox (1958, p. 19) discusses "the requirement that the observation on one [experimental] unit should be unaffected by the particular assignment of treatments to other units, i.e., that there is no "interference" between different units... [I]f

different units are in physical contact [e.g., organisms interacting within a tank or field plot], difficulties can arise and these will now be illustrated by some examples."

Mead (1988, p.120) gives an example involving pigs grouped in pens, where individual pigs were assigned different hormone treatments. Mead then pointed out that analyzing this as a randomized block design would be "inappropriate" given the high potential for interactions among the pigs.

Federer (1993, p. 280) likewise notes that "[o]n statistical grounds, statistical analyses are developed for independence between e.u.'s [experimental units]. Some results have [been] available, and are becoming available, to handle correlated responses. In general, though, statistical procedures require independence. Competition between e.u.'s would result in dependence of observations, perhaps in a complex manner."

The requirement of physical independence for experimental units is not negated by the fact that when weak designs create high potential for "neighbor effects" or "carryover effects" of experimental units on each other, there are procedures whereby one can attempt to correct for such effects *if* the effects are assumed to be of a simple sort—as, of course, they rarely would be, at least with biological material. Good design is not compatible with heavy reliance on weak assumptions.

The pervasive misunderstanding of these basic matters within the statistics profession was well illustrated by comments received when an earlier version of this essay was submitted to and rejected by *The American Statistician,* as discussed in Supplementary Material, Appendix 2.

One root cause of confusion on the definition of experimental unit may be that many statisticians think that the need for physical independence of experimental units is such an obvious prerequisite for error control that it does not need explicit stating. Too much faith is thus put in the defective intuitions of the rest of us. Too little weight is given to the negative influence of the abundance of misanalyzed experiments in statistics books and the disciplinary literature. One anonymous reviewer of this essay noted as another root cause the fact that "many agricultural students are not being trained in the fundamentals [of experimental design] anymore... defective intuitions is only a part of the problem." That comment undoubtedly applies to the statistical curricula of most universities and to students in other experimental disciplines.

Sampling Unit, Observational Unit, and Evaluation Unit

Casella defines *sampling unit* as "the object that is measured in an experiment." This term and *observational unit* (e.g., Kempthorne, 1952, p. 163) are standard terms used by textbooks to draw the critical distinction between the experimental unit and the samples or measurements, often multiple, that may be taken from or made on the experimental unit. So in using *sampling unit* for the concept, Casella is hewing to a common convention.

Recognizing that both *sampling unit* and *observational unit* have long had very general connotations, are equally applicable to observational studies, and thus are problematic as labels for a specific concept in experimental design, Urquhart (1981) proposed the term *evaluation unit* for "the unit of research material on which a response [to a treatment] is evaluated." Adoption of that term was urged by Hurlbert (1990), and a modified definition of it as "that element of an experimental unit on which an individual measurement is made" was offered by Hurlbert and White (1993). On all grounds, *evaluation unit* seems the clearest, most specific, most useful label.

When this essay was under consideration by another journal (*Journal of Animal Ecology*), its reviewers implied that my weak understanding of these matters might be remedied by looking at "Milliken and Johnson" or "the most recent edition of a best-seller like Montgomery," a "classic text." So I looked at them. Both texts abandon the classical terminology as much as does Casella. Milliken and Johnson (2009, p. 114) stated, for example, that "[h]ierarchical designs are often used in the social sciences where groups of individuals form the larger size of experimental unit and the individuals within the group are *the smaller size of experimental unit*" (my emphasis) and give as an example different teaching methods being applied to entire classes of students. That reflects severe confusion between the concepts of experimental unit and evaluation unit. Montgomery (2009) is even more confusing, making no use whatsoever of the terms *experimental unit*, *evaluation unit*, *sampling unit*, or *observation unit*. The labels "recent," "classical," or "best-seller" cannot be taken as prima facie evidence of superior quality. One is reminded of Federer's (1986) comment,

> "[M]y view of what constitutes important developments in Statistics ... is also colored by my eight year tenure as Book Reviews Editor for *Biometrics*. The majority of text and reference books published were disappointing. The profession and investigators in other fields would have been better off without many of these books." (p. 213)

Replication

Casella defines *replication* as "the repetition of the experimental situation by replicating the experimental unit" (p. 4). Later on (p. 22), he uses *true replication* as the label for this concept, distinguishing it from *technical replication*, which he defines as "where the experimental unit is subsampled." But why not stick with the old standard, *treatment replication*, instead of opting for a new label like *true replication*?

In experiments, one can have replicate experimental units, replicate samples or evaluation units, replicate subsamples, replicate blocks, replicate measurements, etc., etc. There are no grounds for considering any of these types of replication more *true* or more *technical* than any others. Furthermore, the concept of replication is equally important to sampling design, where Casella's definitions would be irrelevant. It does not seem useful to create new terms or definitions that "work" or can be understood clearly only within the confines of one's own writing or subdisciplinary network. Similar confusion over "replication" terminology can be found in other books on experimental design, e.g., Milliken and Johnson (1984) or Mead (1988), and does not originate with *Statistical Design*.

Clarity can be increased simply by never using the terms *replication* or *replicate* without explicit indication of the type of entity to which one is referring. In formal writing, *replication* always needs to be preceded by an adjective, and *replicate* is always best followed by a noun. For example: *treatment replication was fivefold, and 12 replicate core samples of soil were taken from each of the five plots under each treatment*.

Pseudoreplication

Casella uses the term *pseudoreplication*, although without defining it directly or giving a literature reference to the concept. In discussing the fish experiment mentioned above, he states (p. 5), "Replicating the fish [i.e., having more than one per tank] is *subsampling* or *pseudoreplication*, and does not affect the main test." To equate pseudoreplication with subsampling is flatly incorrect. Pseudoreplication is an error of statistical analysis and interpretation, not an error or aspect of design. Also, the main test *is* "affected" by having multiple fish per tank even though the error degrees of freedom available for the test of food type effect are not.

Later, in a section titled *Pseudoreplication,* Casella describes (p. 25) an improper analysis of a randomized block 3 × 2 factorial experiment and states, "This analysis again treats subsamples, or technical replicates, as true replications." This comes close to a correct definition of pseudoreplication. It is unclear whether it is intended as such. Pseudoreplication is a category of error that has been the subject of numerous critiques and reviews since the 1980s (Hurlbert 1984, 2009; Hurlbert and White 1993; and references therein). Various types of pseudoreplication have been recognized (simple, temporal, sacrificial, test-qualified sacrificial). With respect to experimental studies, it is broadly defined as (Hurlbert and Lombardi, 2004; Hurlbert, 2009)

> "a serious type of statistical error that…occurs when measurements made on multiple evaluation units, or multiple times on a single evaluation unit, in each experimental unit are treated statistically as if each represented an independent experimental unit… The usual [but not universal] consequence of pseudoreplication is exaggeration of both the strength of the evidence for a true difference between treatments and of the precision with which any difference that does exist has been estimated."

This definition naturally will be unclear to persons who are unfamiliar with the classical definition and concept of the experimental unit.

Split-Plot or Split-Unit Designs

Casella defines split-plot designs rather vaguely as designs "in which there is more than one type of experimental unit" (p. 171). He states that split-unit design is a "more accurate" albeit less "popular" label. I agree. Two of the clearest primers on experimental design (Cox, 1958, p. 142–151; Mead, 1988, p. 382–421) uses the split-unit label. *Plot*, unlike experimental unit, is not a statistical concept, and the split-plot label seems unlikely ever to gain wide transdisciplinary acceptance (think of psychology, medicine, industry, etc.). In the long view, *split unit* thus seems the label best promulgated.

Casella's chapter on split-unit designs opens with this sentence: "Split plot experiments are the workhorse of statistical design" (p. 171). That statement contrasts sharply with the opinion of many other statisticians, as exemplified by Mead (1988, p. 390–393): "[T]he disadvantages of the split unit design are many and four disadvantages are discussed here in detail… The only sound advice I can offer is to avoid split unit designs except when they are essential for practical reasons…"

The contrast is explained by the fact that Casella has rejected the classical use of *split unit* as a label only for a particular type of design structure (sensu Urquhart, 1981; Hurlbert and Lombardi, 2004). Casella uses *split unit* also as a label for a variety of different types of response structures (e.g., repeated monitoring dates, multiple response variables, subsamples, etc.). That conglomeration of concepts also explains why the chapter on split-unit designs is the longest one (71 pages) in his book. On this subject matter Casella (p. ix) was following his mentor "Walt Federer at Cornell, who…made me really understand split-plot designs…" Although in his first book, Federer (1955) seems to have adhered to the classical definition of split-plot designs, he was abandoning that definition by 1977 (Federer, 1977), and by the time of his final book (Federer and King, 2007) was as "libertine" in his use of *split unit* and *experimental unit*, as is *Statistical Design*. Casella was a departmental colleague of Federer's at Cornell University from 1981 to 2000.

Casella (p. 171) presents as one example of a split-unit design a study where the effects of four different diets on the blood pressure of three subjects per diet type were measured once in the morning and once in the evening 2 wk after the diet regimes were imposed. But this is *not* a split-unit design structure. The study has a unifactorial treatment structure, a completely randomized design structure, and a repeated-measures response structure. "Morning" and "evening" cannot be regarded as subplots or subunits. They are merely labels for the successive points in time when the repeated measurements were made.

Relevant to this example are the comments of Yates (1982) on another study that involved a repeated-measures response structure and was claimed, on that basis, to have a split-plot design. Yates stated: "This is incorrect terminology. In a split-plot experiment different treatments are applied to the subplots constituting a whole plot; if the experiment is properly randomized these treatments are *randomly* assigned to the subplots. In an experiment with repeated measurements…the measurements have a temporal sequence…" Rowell and Walters (1976), de Klerk (1986), Finney (1990), Mead et al. (2003), Hinkelmann and Kempthorne (1994, 2008), and many others have opined similarly on this point.

Another example Casella presents (p. 196) as a split-split-unit design involved growth chambers each containing six different kinds of plants, with two of these chambers under each of four different ozone levels. At the end of the experiment, some variable (unspecified) is measured on both the root and the shoot of each plant. These two parts of the plant are treated as "subunits," thus defining, putatively, a second level of "splitting." They do not have the requisite physical independence, however, for them to be regarded as independent subunits. The experiment has only a standard split-unit design structure and a response structure involving two different response variables measured on each subunit (or on each individual plant in each subunit, if each species is represented by more than a single plant).

In a third experiment that Casella (p. 201) claims has a split-split-unit design, each of three laboratories (= blocks) tests three different washing solutions for their ability to retard bacterial growth in milk cans. Effects are assessed by conducting two assays on the contents of each can on each of 4 d. The experiment, in fact, has a simple randomized complete block

design structure with a response structure involving a nested sampling design. The experimental unit (milk can) is not "split" in any way.

Implicit in all these misuses of the split-unit label is the general misconception that all entities nested within the experimental unit should and can, without confusion, be labeled as "subunits" regardless of their physical interdependencies and even if no separate treatments are applied to them.

Split-unit experiments are classically defined as experiments where the experimental unit is defined at two or more scales and where different levels of one or more treatment factors are assigned to whole (experimental) units and different levels of one or more other treatment factors are assigned to discrete sub- (experimental) units of the whole units (e.g., Yates, 1935, 1937; Kempthorne, 1952; Cochran and Cox, 1953; Cox, 1958; Mead, 1988; Mead et al., 2003; Hinkelmann and Kempthorne, 1994, 2008). It is understood that the subunits within a whole unit must be just as physically independent of each other as are the whole units of each other, as discussed above. The researcher must have confidence that what transpires on one subunit will not be able to affect the responses of other subunits.

In an example concerning different species of plants (subunit factor) grouped in growth chambers with different light intensities (whole-unit factor), Federer (1975) pointed out that analyzing this experiment as one with a split-unit design would be an "erroneous procedure" if the putative subunits (different species) within a chamber influenced each other.

Interestingly, Federer's good advice in that example was contrary to that he gave in a later one. Federer (1977) presented as a putative split-unit design an experiment where different diets (whole-unit factor) were assigned to different pens, each containing a litter of piglets. Different "minor nutritional elements" (subunit factor) were then fed to different piglets. The high potential for interactions among piglets (putative subunits) within a pen disallows this being classified as a split-unit design and the piglets being treated as physically independent subunits.

Again, these improper usages of split-plot or split-unit terminology did not originate with *Statistical Design*. They have long been abundant in statistics textbooks (e.g., Steel and Torrie, 1960; Gill, 1978; Kirk, 1982; Milliken and Johnson, 1984, 2009; Steel et al., 1997; Federer and King, 2007; Montgomery, 2009). The misuses, usually accompanied by misunderstandings of *experimental unit*, also have resulted in many cases of *pseudofactorialism* in the disciplinary literatures. That error is defined as an "invalid statistical analysis that results from the misidentification of two or more response variables as representing different levels of an experimental variable" (Hurlbert and White 1993; Hurlbert, 2013).

The same confusion is often found in books on experimental design in chapters or sections titled "Nested designs" or "Hierarchical designs." Such terms are superfluous and also problematic in a number of other ways. They are general terms lacking specific definitions, yet they carry the flavor of technical terms that do have specific technical definitions. Nested or hierarchical structure can be found *in any of the three aspects* of a design—treatment structure, design structure, and response structure—but has very different implications for analysis and interpretation in each case. The use of these terms as formal labels typically is associated with a failure to recognize that response structure, conceptually and operationally, is an aspect distinct from both treatment structure and design structure, as Finney (1955) and Urquhart (1981) emphasized.

Repeated Measures

Casella (2008, p. 216) states that "in a repeated measures design, we typically take multiple measurements on a subject over time. If any treatment is applied to the subjects, they immediately become the whole plots, and the treatment 'Time' is the split plot treatment." The first sentence is fine. The second reflects the confusion discussed above. "Time" is not properly regarded as either a treatment or a treatment factor in the examples presented.

The widespread confusion about the meanings of *split unit* and *repeated measures* perhaps derives from a tendency on the part of more theoretically oriented statisticians (and those who follow them) to focus on mathematics and unitary models and to prefer lean formal terminologies. It is true that split-unit design structures, crossover design structures, repeated-measures response structures, and multivariate response structures can (but not necessarily should) be analyzed with models of similar if not identical structure. That "elegance" may entrance the theoretician. The applied statistician or researcher, however, is likely to be more interested in the fact that these structures are designed to answer different kinds of questions and the fact that the resulting similar ANOVAs and their interpretations are based on different kinds of assumptions. The different aspects of experimental design can be discussed clearly only with a fairly rich terminology that is independent of, although related to, mathematical terminologies. At least for experimenters, if not for mathematicians, texts using experiment-focused conceptual and terminological frameworks will always be clearer and more useful than texts based primarily on model-focused frameworks.

CONCLUSIONS

Because *Statistical Design* has been recently published and is getting widespread notice and favorable reviews, it seemed that a critique of the terminological problems in it would be timely, useful, and focused enough to have a positive impact on both researchers and statisticians. *Statistical Design* has not been the originator of any of the improper usages I criticize. These are the inheritance from a discipline that has been badly language challenged for a long time. Revising the quotation that opens this essay, I suggest, "It would be convenient to *agree on* a standard terminology!"

In 2013 we still have a long way to go. Progress will be accelerated if experiment-focused and model-focused statisticians adopt a common set of terms and definitions with which to discuss "experimental components." We need the best, down-to-earth applied statisticians to serve as mediators—and as pre-publication reviewers for the next generation of statistics textbooks.

ACKNOWLEDGMENTS

For their suggestions on drafts of this paper, I thank David R. Cox, Klaus Hinkelmann, Kathy Mier, N. Scott Urquart, and 12 anonymous reviewers.

REFERENCES

Casella, G. 2008. Statistical design. Springer, New York.

Cochran, W.G., and G.M. Cox. 1953. Experimental designs. John Wiley & Sons, New York.

Cox, D.R. 1958. Planning of experiments. John Wiley & Sons, New York.

Cox, D.R. 1961. Design of experiments: The control of error. J. R. Stat. Soc. Ser. A 124:44–48. doi:10.2307/2343152

de Klerk, N.H. 1986. Repeated warnings re repeated measures. Aust. N.Z. J. Med. 16:637–638. doi:10.1111/j.1445-5994.1986.tb00001.x

Federer, W.T. 1955. Experimental design: Theory and application. Macmillan, New York.

Federer, W.T. 1973. Statistics and society: Data collection and interpretation. Marcel Dekker, New York.

Federer, W.T. 1975. The misunderstood split plot. In: R.P. Gupta, editor, Applied statistics. North Holland Publ., Amsterdam. p. 9–39.

Federer, W.T. 1977. Sampling, blocking, and modeling considerations for split plot and split block designs. Biom. J. 19:181–200. doi:10.1002/bimj.4710190304

Federer, W.T. 1984. Principles of statistical design with special reference to experiment and treatment design. In: H.A. David and H.T. David, editors, Statistics: An appraisal. Iowa State Univ. Press, Ames. p. 77–104.

Federer, W.T. 1986. Whither statistics? In: C.E. McCulloch et al., editors, Statistical design: Theory and practice. Proceedings of a conference in honor of Walter T. Federer. Cornell Univ., Ithaca, NY. p. 211–231.

Federer, W.T. 1993. Statistical design and analysis for intercropping experiments. Vol. 1. Two crops. Springer-Verlag, Berlin.

Federer, W.T., and F. King. 2007. Variations on split plot and split block experiment designs. John Wiley & Sons, Hoboken, NJ.

Finney, D.J. 1955. Experimental design and its statistical basis. Univ. of Chicago Press, Chicago.

Finney, D.J. 1990. Repeated measurements: What is measured and what repeats? Stat. Med. 9:639–644. doi:10.1002/sim.4780090610

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh, UK.

Gill, J.L. 1978. Design and analysis of experiments in the animal and medical sciences. Vol. 2. Iowa State Univ. Press, Ames.

Hinkelmann, K., and O. Kempthorne. 1994. Design and analysis of experiments. Vol. 1. Introduction to experimental design. 2nd ed. Wiley-Interscience, New York.

Hinkelmann, K., and O. Kempthorne. 2008. Design and analysis of experiments. Vol. 1. Introduction to experimental design. 3rd ed. Wiley-Interscience, New York.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211. doi:10.2307/1942661

Hurlbert, S.H. 1990. Pastor binocularis: Now we have no excuse. Ecology 71:1222–1223. doi:10.2307/1937395

Hurlbert, S.H. 2004. On misinterpretations of pseudoreplication and related matters. Oikos 104:591–597. doi:10.1111/j.0030-1299.2004.12752.x

Hurlbert, S.H. 2009. The ancient black art and transdisciplinary extent of pseudoreplication. J. Comp. Psychol. 123:434–443. doi:10.1037/a0016221

Hurlbert, S.H. 2013. Pseudofactorialism, response structures and collective responsibility. Austral Ecol. (in press). doi:10.1111/aec.12010

Hurlbert, S.H., and C.M. Lombardi. 2004. Research methodology: Experimental design, sampling design, statistical analysis. In: M.M. Bekoff, editor, Encyclopedia of animal behavior. Greenwood Press, London. p. 755–762.

Hurlbert, S.H., and C.M. Lombardi. 2009. Final collapse of the Neyman–Pearson decision theoretic framework and rise of the neoFisherian. Ann. Zool. Fenn. 46:311–349. doi:10.5735/086.046.0501

Hurlbert, S.H., and M.D. White. 1993. Experiments with freshwater invertebrate zooplanktivores: Quality of statistical analyses. Bull. Mar. Sci. 53:128–153.

Kempthorne, O. 1952. Design and analysis of experiments. John Wiley & Sons, New York.

Kempthorne, O. 1982. Edgington, Eugene S., *Randomization Tests*. Biometrics 38:864–867. doi:10.2307/2530071

Kirk, R.E. 1982. Experimental design: Procedures for the behavioral sciences. 2nd ed. Brooks/Cole Publ., Pacific Grove, CA.

Kozlov, M., and S.H. Hurlbert. 2006. Pseudoreplication, chatter, and the international nature of science: A response to D.V. Tatarnikov. (In Russian.) Zh. Obshch. Bio. 67:145–152.

McCulloch, C.E., S.J. Schwager, G. Casella, and S.R. Searle, editors. 1986. Statistical design: Theory and practice. Proceedings of a conference in honor of Walter T. Federer. Cornell Univ., Ithaca, NY.

Mead, R. 1988. The design of experiments. Cambridge Univ. Press, Cambridge, UK.

Mead, R., and R.N. Curnow. 1983. Statistical methods in agriculture and experimental biology. Chapman and Hall, London.

Mead, R., R.N. Curnow, and A.M. Hasted. 2003. Statistical methods in agriculture and experimental biology. 3rd ed. Chapman & Hall, New York.

Milliken, G.A., and D.E. Johnson. 1984. Analysis of messy data. Vol. I. Designed experiments. Van Nostrand Reinhold, New York.

Milliken, G.A., and D.E. Johnson. 2009. Analysis of messy data. Vol. 1. Designed experiments. 2nd ed. CRC Press, New York.

Montgomery, D.C. 2009. Design and analysis of experiments. 7th ed. John Wiley & Sons, Hoboken, NJ.

Puntanen, S. 2008. *Statistical Design* by George Casella. Int. Stat. Rev. 76:443. doi:10.1111/j.1751-5823.2008.00062_8.x

Raktoe, B.L., A. Hedayat, and W.T. Federer. 1981. Factorial designs. John Wiley & Sons, New York.

Rowell, J.G., and D.E. Walters. 1976. Analyzing data with repeated observations on each experimental unit. J. Agric. Sci. 87:423–432. doi:10.1017/S0021859600027763

Steel, R.G.D., and J.H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill, New York.

Steel, R.G.D., J.H. Torrie, and D.A. Dickey. 1997. Principles and procedures of statistics. 3rd ed. McGraw-Hill, New York.

Urquhart, N.S. 1981. The anatomy of a study. HortScience 16:100–116.

Vahl, C.I. 2008. *Statistical Design* by Casella, G. Biometrics 64:1304–1305. doi:10.1111/j.1541-0420.2008.01138_6.x

Valiela, I. 2001. Doing science: Design, analysis, and communication of scientific research. Oxford Univ. Press, Oxford, UK.

Verkuilen, J. 2010. A review of *Statistical Design*. J. Educ. Behav. Stat. 35:248. doi:10.3102/1076998609341363

Wiley, R.H. 2003. Is there an ideal behavioural experiment? Anim. Behav. 66:585–588. doi:10.1006/anbe.2003.2231

Yates, F. 1935. Complex experiments. Suppl. J. R. Stat. Soc. 2:181–247.

Yates, F. 1937. The design and analysis of factorial experiments. Tech. Commun. 35. Imperial Bur. of Soil Sci., Harpenden, UK.

Yates, F. 1982. [Comentary on regression models for repeated measurements and M. Aitkin's use of the term 'split-plot']. Biometrics 38:850–853.

Supplementary Material

for

Hurlbert, S.H. 2013. Affirmation of the classical terminology for experimental design via a critique of Casella's *Statistical Design*, Agronomy Journal 105:412-418

APPENDIX 1: Book reviews for *Statistical Design*

I have seen three reviews of Casella's (2008 *Statistical Design*. These are brief and largely complimentary.

Puntanen (2008) says the book "exceeds exceptionally well..[in describing] the principles that drive good design…[and would be] an excellent course book."

Vahl (2008) says it "does a nice job demonstrating many definitions and concepts though real world examples …[and he] intends to use this book as supplementary reading material for [his] own design course." However, he also found the book "uneven in its depth" in places, finds the treatment of multiple comparison methods to be insufficient, and is "conflicted" about portions of the chapter on split plot experiments.

Verkuilen (2010) found the book "very clear and readable" and says it would be "very useful for advanced students who are in a statistics or experimentally oriented behavioral science program…[and] would also be a very useful reference work for anyone likely to consult with experimentalists." He is dissatisfied with the lack of attention to outlier detection, generalized linear models and a "thin" treatment of repeated measures designs.

I attempt no evaluation of the cogency of these reviews but cite them only to document the largely positive reception the book has had so far.

APPENDIX 2: *The American Statistician* opines on 'experimental unit' and 'blocking'

The pervasive misunderstanding on these basic matters within the statistics profession was well illustrated when an earlier version of this paper was submitted to and rejected by *The American Statistician (TAS).* The major technical objection to the manuscript, concurred in by two referees and two editors, was that with respect to the fish experiment discussed, Casella (2008, p. 4) was correct and that "the author [S.H.] errs in some of the criticisms, especially concerning the fundamental notion of experimental unit and its relationship to blocking" (*TAS* associate editor, pers. comm. to S. Hurlbert.).

TAS Reviewer #1 opined that in Casella's fish example, "the fish would be the experimental unit if the fish within a tank were given different food types. In this case the tank would be a blocking factor rather than a treatment factor." TAS Reviewer #2 opined, "… it is indeed the fish that are the experimental units. There is no requirement for experimental units to be independent." Casella only *implied* that 'tank' would be validly treated as a blocking factor; the *TAS* statisticians were *stating,* erroneously, that it could validly be so treated. This puts them at odds with R.A. Fisher, F. Yates, D.R. Cox, W.T. Federer (in 1975 and 1993, at least) and R. Mead, among many others.

Have we espied at TAS the tip of an iceberg bigger than that which sank the *Titanic*?