GLOBALP/GETTY IMAGES

Common Statistical Mistakes in Entomology:

Ignoring Interactions

Editor's Note: This article is the third in a series of commentaries that address common statistical mistakes in entomology.

DALE W. SPURGEON

The previous article of this series (Spurgeon 2019b) addressed the consequences of using an analysis of variance (ANOVA) model that does not faithfully represent the experimental design. However, even when the ANOVA model accurately represents the respective treatment and design structures of an experiment, it is fairly common that F-tests of the model effects are not appropriately interpreted. In particular, analysts often ignore interaction effects, and instead take the simpler approach of addressing only the main effects. Milliken and Johnson (1984) explicitly addressed this problem and suggested that it occurs because some analysts do not understand how to address interactions, or they hold an unwarranted belief that the interactions are not important. This article addresses why the interaction effects in multifactor experiments must not be ignored, and how their misinterpretation can lead to erroneous conclusions.

Both previous commentaries in this series (Spurgeon 2019a, b) showed how interactions between fixed (treatment) effects and random (blocking or repetition) effects are often needed in the ANOVA model. I also pointed out that there usually are no valid *F*-tests for these fixed-effect*random-effect interactions; their purpose is to partition nuisance variation from experimental error and to serve as error terms for testing the fixed effects. Although some software will report *F*-tests for these variance components (e.g., PROC GLM of SAS [SAS Institute 2012]), those tests and their *p*-values are irrelevant or uninterpretable. The interactions of interest, in the context of hypothesis testing, are solely those between the fixed treatment effects.

An *F*-test of an interaction between two fixed effects assesses whether the responses to each of these effects are more or less independent, or whether the responses are dependent (or *conditional*) on the levels of the other effect. If the *p*-value of the *F*-test is *negligible*, then the response to a given treatment is said to be consistent over levels of the other effect. If the *F*-test suggests the interaction is *non-negligible* (but not necessarily *significant* at $\alpha = 0.05$), then the response to a given treatment varies among the levels of the other, interacting treatment.

Consider the mortality responses to two

hypothetical toxicants evaluated at different temperatures (Fig. 1). If the responses to the two toxicants (Tox1 and Tox2) are roughly parallel at different temperatures (Temp), then there is no interaction and inferences or conclusions regarding the toxicants and temperatures should be based on their respective main effect tests. In this case (Fig. 1a), Tox1 is more effective than Tox2, and efficacy of both materials increases similarly with temperature. As in this example, the presence or absence of an interaction is often easy to visualize from the data.

Alternatively, suppose the relative effects of the two toxicants vary among temperatures, and the Tox*Temp interaction is non-negligible (Fig. 1c–d, and likely Fig. 1b). In that case, interpretation of the main effect of toxicant is *conditional* on the level of temperature, and the effect of temperature is conditional on toxicant. The corresponding *F*-tests and *p*-values for either main effect are, at best, irrelevant, and may be completely misleading. Note that the interaction is described as *non-negligible*, but not necessarily statistically significant at $\alpha = 0.05$ (the nominal error rate typically used for hypothesis testing). An interaction can be both non-negligible and non-significant (e.g., $p > 0.05$). This is especially true when only one or a few of many cells (treatment combinations) are responsible for the non-negligible interaction (Stroup 2013). There are no formal guidelines regarding the *p*-value at which one should declare a non-negligible interaction, because the *p*-value will be influenced by the total number of cells, the number of cells exhibiting the interaction response, and the magnitude of the differences among cells of different treatment combinations. Fortunately, these types of interactions (Fig. 1b) are usually evident from visual inspection of the data and often have a biological interpretation that makes sense.

Once the analysis provides evidence of a non-negligible interaction, the nature of the interaction should be explored. Two common but unsatisfactory approaches of exploring an interaction are to examine all pairwise comparisons among the treatment combinations, or to examine subsets of the data in individual *t*-tests or one-way ANOVAs. Pairwise comparisons among the treatment combinations are statistically inefficient if there are many treatment combinations, because of the statistical power lost in making the necessary adjustments for multiplicity. In addition, many of the comparisons may be uninteresting or irrelevant. This approach, applied to the examples in Fig. 1, will provide nine meaningful comparisons: three comparisons among temperatures for each toxicant (six total comparisons) and a comparison between toxicants at each temperature (three total comparisons). However, this approach will also make four meaningless comparisons between the two toxicants at different temperatures (e.g., Tox1 at the low temperature versus Tox2 at the high temperature). Inclusion of the meaningless comparisons inflates the multiplicity-adjusted *p*-values. The use of *t*-tests or one-way ANOVAs on subsets of the data is unsatisfactory because the analytical models no longer accurately represent the experimental design, the consequences of which were discussed in the previous commentary (Spurgeon 2019b). The disadvantages of both of these approaches become more severe as the interaction becomes more complex (i.e., contains more main effects).

A simple and appropriate way to examine an interaction is through the use of *simple effect tests*, in which the effects of one factor or treatment are compared within levels of the other factor. This can be accomplished in planned contrasts if the analyst can correctly identify the appropriate coefficients in a group of CONTRAST statements in PROC MIXED or PROC GLIMMIX (SAS Institute 2012). However, both of these procedures provide a much simpler means of obtaining these tests, which is the SLICE option of the LSMEANS statement.

Using the examples in Fig. 1, the statement

```
lsmeans tox*temp / slice=tox;
```

would test the effect of temperature within each toxicant. The effect of temperature within toxicant, and toxicant within each temperature, can be tested at the same time using the statement SLICE (TOX TEMP). Because the TOX main effect has only two levels (Tox1, Tox2), the simple effects test of toxicant "sliced" by temperature (SLICE=TEMP; the two toxicants are compared at each temperature) will indicate

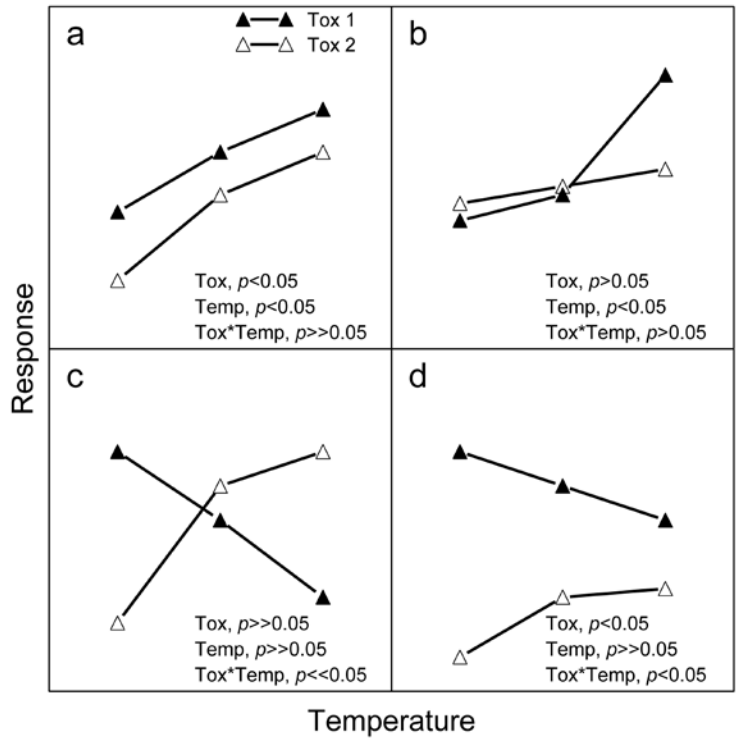

Fig. 1. Graphical representations of hypothetical responses of insects to two toxicants (TOX) at three temperatures (TEMP), illustrating a scenario without a toxicant × temperature interaction (a) and three scenarios in which toxicant and temperature interact (b–d). Note that in (b), the interaction appears non-negligible but is not statistically significant at $\alpha = 0.05$.

whether the responses to toxicant 1 and toxicant 2 are the same at each temperature, and no further testing or multiplicity adjustment is needed. Simple effect tests of temperature "sliced" by toxicant (SLICE=TOX; the influence of temperature is tested within each toxicant) will indicate whether temperature influenced the response to toxicant 1, toxicant 2, both, or neither. Because temperature has three levels, if the simple effects test of toxicant 1 or toxicant 2 is significant, one may wish to compare responses to temperature within one or both toxicants. In PROC GLIMMIX, this would be accomplished by the statement

```
lsmeans tox*temp / slicediff=tox
  adjust=simulate;
```

which will make all pairwise comparisons among levels of temperature within toxicant 1 and within toxicant 2 (PROC MIXED does not currently implement the SLICEDIFF option). This approach eliminates the irrelevant comparisons between toxicants 1 and 2 when they are each at different temperatures (e.g., toxicant 1 at temperature 1 vs. toxicant 2 at temperature 3), which are output when one makes all pairwise comparisons among levels of the Tox*Temp interaction. Additional options in GLIMMIX are available to control which comparisons are made, including restricting comparisons of each treatment level to a control (ADJUST=DUNNETT). Because this approach makes multiple comparisons, the SLICEDIFF statement should be accompanied by one of the various experiment-wise multiplicity adjustments in an ADJUST= statement (e.g., BON, TUKEY, SIMULATE), as illustrated in the lsmeans statement above.

The approach of examining simple effects is expandable to more than two interacting treatments (e.g., SLICE=*EFFECT1 EFFECT2 EFFECT3*), but even this approach can become cumbersome with interactions containing more than three effects. However, one of the major advantages of examining simple effects is the likelihood of finding that one or more levels of a given effect are contributing little or nothing to the interaction. In that case, the ability to focus on the treatment levels that are important can greatly simplify interpretation of the results.

Virtually any experiment in which treatment levels of one or more of the main effects are quantitative (e.g., date, dose, temperature) and involve a range of levels that produce a corresponding range of responses will likely involve an interaction. These interactions are generally easy to see in the graphical or tabular results of manuscripts and papers, but they are often not addressed. When these interactions occur, some reviewers or editors still insist on reporting or interpreting the *F*-tests of the main effects. The only justification for reporting *F*-tests of the corresponding main effects is to document the degrees of freedom so the reader can look for other potential problems with the analysis. Otherwise, tests of the main effects are irrelevant and uninterpretable, regardless of their respective *p*-values.

Finally, when an analysis indicates a likely interaction and only one or a few treatment combinations are responsible, the analyst should consider whether the interaction is biologically reasonable. This is especially true in cases where sample sizes are small or the entire experiment has not been replicated to demonstrate or assess repeatability of the results, which unfortunately is common. In those cases, one should be concerned for the likelihood of a Type-I error (in which a difference is declared based on $p < 0.05$ but is actually caused by random chance). There is a fairly widespread misconception that adjustments for multiplicity eliminate, or nearly eliminate, Type-I errors. This is simply not so. Type-I errors can and do occur, and the only way to completely avoid them is to set the Type-I error rate (α) so low that a difference cannot be detected. Occurrence of a Type-I error should not doom a sound piece of work to rejection. Instead, the researcher should provide a frank and honest rationale for suspecting a Type-I error, and an explanation of how such an error might influence the interpretation, impact, or application of the research.

In summary, interactions between or among ANOVA model main effects are often tested but dismissed as unimportant or too complicated for interpretation or presentation. When demonstrated and repeatable interactions occur, they contain all of the interpretable information from an analysis, and the main effect tests are irrelevant. Especially when an interaction is unexpected, its recognition and exploration represent an opportunity to gain a more astute understanding of the studied system than is possible in its absence.

ESPECIALLY WHEN AN INTERACTION IS UNEXPECTED, ITS RECOGNITION AND EXPLORATION REPRESENT AN OPPORTUNITY TO GAIN A MORE ASTUTE UNDERSTANDING OF THE STUDIED SYSTEM THAN IS POSSIBLE IN ITS ABSENCE.

Acknowledgments

Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References Cited

Milliken, G.A., and D.E. Johnson. 1984. Analysis of messy data, volume 1: designed experiments. Van Nostrand Reinhold, NY.

SAS Institute. 2012. SAS release, ed. 9.4. SAS Institute, Cary, NC.

Spurgeon, D.W. 2019a. Common statistical mistakes in entomology: pseudoreplication. American Entomologist 64: 16–18.

Spurgeon, D.W. 2019b. Common statistical mistakes in entomology: models inconsistent with the experimental design. American Entomologist 64: 87–89.

Stroup, W.W. 2013. Generalized linear mixed models: modern concepts, methods and applications. CRC Press, Boca Raton, FL.

Dale W. Spurgeon, *USDA, ARS, Arid-Land Agricultural Research Center, 21881 N Cardon Lane, Maricopa, AZ (retired). E-mail: dwspurg@gmail.com*

DOI: 10.1093/ae/tmz042