Journal of Scientific Exploration, Vol. 17, No. 4, pp. 643–650, 2003 0892-3310/03

COMMENTARY

Comments on Spottiswoode and May

MIKEL AICKIN

Center for Health Research
Kaiser Permanente Northwest Region
Portland, OR

Abstract—This comment examines the accompanying paper by S. James P. Spottiswoode and E. C. May; specifically, the statistical analysis of the primary endpoint—the comparison of the fraction of skin conductance responses preceding an audio stimulation with the fraction preceding a period of silence. It addresses the three following points: that the author's statistical analysis does not meet current professional standards; that, by under-analyzing their data, the authors have failed to note patterns that are important for their interpretation; and, that by pre-conceiving the meaning of their experiment, the authors have missed an alternative interpretation that is more highly supported by a statistical analysis of their data.

Keywords: statistical analysis—skin conductance

Introduction

Spottiswoode and May have provided an exemplary experiment to establish the existence of a prestimulus response to a loud sound. This extends earlier work that had focused on pre-response to visual stimuli. The care taken in designing, executing, and reporting this experiment is an excellent illustration of what is necessary when one wants to provide evidence in favor of a phenomenon that does not sit well with conventional science. There will, unfortunately, be no lack of professional skeptics who will want to take any potential problem with this kind of research, and turn it into a fatal flaw. For this reason, it is important that the statistical analysis be of the same high quality as the other parts of the presentation, and it is on this account that Spottiswoode and May have fallen short of achieving what is possible.

This comment is directed entirely at the statistical analysis of the primary endpoint—the comparison of the fraction of skin conductance responses preceding an audio stimulation with the fraction preceding a period of silence. I want to thank Spottiswoode and May for generously providing their raw data for this comment. The following points will be made:

1. The author's statistical analysis does not meet current professional standards.
   a. The elimination of certain data series was unjustified, and was not carried out in an even-handed manner.
   b. The analysis does not allow for dependence of responses, or alternatively, for heterogeneity of response probabilities.
2. By under-analyzing their data, the authors have failed to note patterns that are important for their interpretation.
3. By pre-conceiving the meaning of their experiment, the authors have missed an alternative interpretation that is more highly supported by a statistical analysis of their data.

The Primary Analysis

The primary analysis of Spottiswoode and May's data is stark, in contrast to the material on measurement issues. It consists of a two-by-two table, classifying all 2500 responses according to stimulus condition (noise or silence) and pre-response (present or absent). The statistical test is a conventional comparison of binomial proportions. This test depends on two assumptions; (1) the 2500 observations are all independent of each other, (2) the probability of a pre-response before a noise is the same at each observation, and the probability of a pre-response before a silence is the same for each observation. Thus, the analysis carried out might have been appropriate if the authors had conducted 2500 observations on 2500 randomly selected people. The justification would have been that (1) different people behave independently of each other, at least under these defined experimental circumstances, and (2) although the probabilities of response might vary between individuals, because they were randomly selected we can regard each individual's probabilities (under the two conditions) as being the same.

The experiment just described, which would justify Spottiswoode and May's statistical analysis, was not the one they carried out. Instead, they recruited 125 individuals (whether randomly or not we cannot tell) and made 20 observations per person ($20 \times 125 = 2500$). The fact that the observations did not come from different individuals, but rather came in packets of 20 per person, is what differentiates Spottiswoode and May's actual experiment from the experiment that would have justified their analysis. It is virtually a cliché in areas such as biostatistics that when one makes multiple observations on individuals, one should use an analysis that takes possible interdependence (within an individual) into account. The test used by Spottiswoode and May does not do this.

One can argue back and forth whether this makes any difference. There are at least three aspects of this argument that will emerge here. The first is that the analysis used by Spottiswoode and May generally estimates less variability in their endpoints than is actually present (due to inter-individual interdependence). Again in general, this would mean that the statistical significance of their results

TABLE 1
Stata Analysis of Spottiswoode and May's Data

General estimating equation for panel data		Number of obs	= 2500
Group variable:	id	Number of groups	= 125
Link:	identity	Obs/group min	= 20
Family:	binomial	Obs/group avg	= 20.00
Correlation:	exchangeable	Obs/group max	= 20
		chi^2 (1)	= 11.32
Scale parameter:	1	Prob < chi^2	= 0.0008
Pearson chi^2 (2498):	2499.82	Deviance	= 1183.61
Dispersion (Pearson):	1.00073	Dispersion	= .4738247

a	Coef.	Std. Err.	z	P < |z|	95% Conf. Interval	
s	.0321293	.0095489	3.365	0.001	.0134138	.0508447
cons	.047445	.0078976	6.007	0.000	.0319659	.0629241

would be inflated. The inflation would be due to an under-estimation of variability that is a consequence of the statistical procedure.

Methods of accounting for interdependence are well-known. Here is the output from a statistical package (Stata) that does this, using Spottiswoode and May's data (Table 1). Here "a" denotes an anticipatory response, and "s" denotes the stimulus (noise) condition. This is a model for the probability of "a", in which the effect of "s" on that probability is portrayed as a simple additive increase. The conclusion is that the probability of an anticipatory response is increased by about 3.21% (under "Coef.") by a noise condition, and that the p-value associated with this result is about 0.001 (under "p > |z|"). (The _cons term here indicates the probability of response under the silence condition, about 4.74%.) It should be clear that this analysis is simple to carry out, accounts automatically for within-person intercorrelation, and gives an eminently interpretable result. To the contrary, here is what Spottiswoode and May say:

> We computed a *Z*-Score of 3.27 and a per stimulus effect size of 0.0901 ± 0.0275 for a p-value of 5.4×10^{-4} (1-tailed). On a per participant basis we compute an effect size of 0.292 ± 0.089.

There are several strange things here. First, the term "effect size" is not defined by Spottiswoode and May, and so it is not clear how this arcane quantity relates to the very simple 0.032 effect that is clearly indicated by the above computer output. In statistics, "effect size" is an artifact created by the necessity of doing power computations for grant applications, where one does not yet know the relevant standard deviation of the estimate. In effect, an "effect size" expresses differences between groups (using some level of comparison) in terms of an unknown unit, the missing standard deviation. Presumably Spottiswoode and May used this concept, although it is clear neither how nor why.

The second oddity is the use of a one-sided p-value. It is virtually an article of faith in the statistical literature that one should use two-sided p-values (a two-sided p-value is generally twice the one-sided value). This explains why the computer output gives 0.001, whereas Spottiswoode and May give 0.0005. The third weirdness is what a "per participant effect size" might mean, and why it differs so much from the putative main result.

The reader might well ask, at this point, whether I am just quibbling, since I do not come to a different conclusion than the authors do. This is the second point about poor statistical methods, that they can produce an essentially correct result, even though they are themselves suspect, if not wrong. One should not conclude from this that there is no difference between poor and good methods. The reason we prefer the good methods is that they mislead us less often than the poor ones.

The third point is that there is no reason (in my mind) to publish incorrect and potentially misleading statistical analyses, even if in the instance they do not mislead. No stretch of the imagination is needed to foresee a would-be author in the future, who submits an article to *JSE* using the same flawed statistical approach, where a correct analysis does in fact make a difference. This future author will be outraged by a rejection, because he/she used a method *that had already been published in JSE*, and so to him/her it will look like unfair treatment. And, indeed, the future author would be right—it would be unfair. The perhaps unfortunate fact is that many researchers learn their statistical methods from their own disciplinary journals, which has given rise to some exceedingly strange schools of thought about inference. It does not seem to me a good idea for *JSE* to contribute to this process.

A second maneuver of Spottiswoode and May that is not worthy of emulation is the elimination of certain data for unsubstantiated reasons. Here is their description of what they did:

> [W]e decided in advance to reject sessions with less than six stimuli of either type, to reduce the variance of the within session effect size.

The underlying principle seems to be that it is permissible simply to delete certain data points from a designed experiment, based on some argument having to do with their variability. Nowhere in the literature of statistical analysis of designed experiments is this principle respected. It is virtually only in the privacy of scientific laboratories where this censoring occurs, usually outside the scrutiny of statisticians, or any other scientists. Moreover, given that Spottiswoode and May pooled all their data, irrespective of its clustering in 20-packet sub-experiments, their reasoning for omitting certain packets makes little sense in terms of the primary analysis they actually used.

To illustrate the dangers of this elimination strategy, from data provided by Spottiswoode and May it can be seen that there are no persons included with fewer than six noise stimuli. There is, however, one person included in the analysis who had 15 noise stimuli. This person had, therefore, five silence

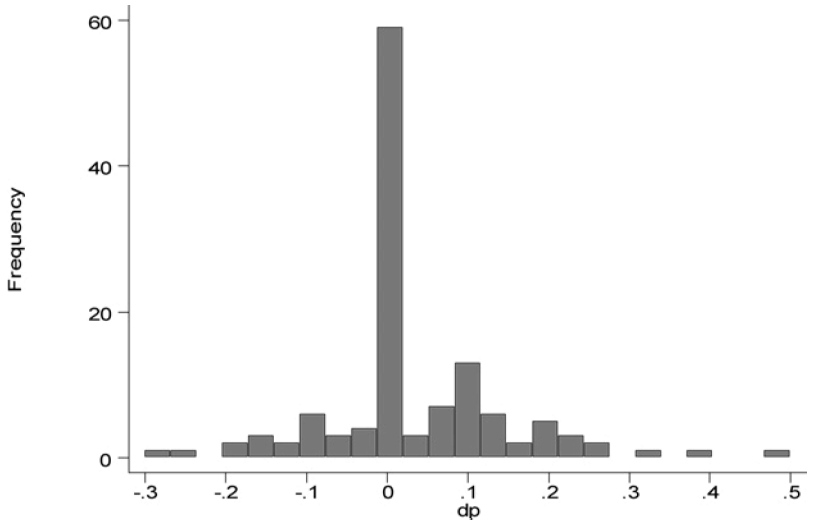

Fig. 1. Histogram of dp = proportion of responses before noise stimuli minus proportion of responses before silence stimuli.

stimuli, and so should have been excluded by the rule above, cited from the article. It is of no small moment, perhaps, that the lone individual with 15 noise stimuli had a difference of 0.48 between their probabilities of response, noise vs. silence, more than 10 times the average of such differences over the other participants. The elimination of data, especially for questionable reasons, is an activity fraught with peril.

Behind the Primary Analysis

The strategy for statistical analysis in many disciplines is to plan the experiment, figure out the analysis in advance, do the experiment, perform the analysis according to plan, and report the results. Spottiswoode and May cannot be faulted on this score, since this is precisely what they did. That this strategy can be misleading is not their fault, but it is worthy of investigation nonetheless.

It is very difficult to foresee all of the aspects of a complex data collection before the fact, particularly when the phenomena under study are not completely understood. To illustrate, the following is a histogram of the 125 participants in the data provided by Spottiswoode and May, with respect to "dp", the difference between the anticipatory response under the noise condition minus that under the silence condition (Figure 1).

The aspect of this display that cries out for explanation is the large number of participants for whom dp was essentially zero. Examination of the data shows that in fact 57 individuals (46% of all participants) had no skin conductance response in any of their 20 trials, whether followed by a noise stimulus or not.

TABLE 2
Stata Analysis of Spottiswoode and May's Data

General estimating equation for panel data		Number of obs	= 1360
Group variable:	id	Number of groups	= 68
Link:	identity	Obs/group min	= 20
Family:	binomial	Obs/group avg	= 20.00
Correlation:	exchangeable	Obs/group max	= 20
		chi^2 (1)	= 10.67
Scale parameter:	1	Prob < chi^2	= 0.0011
Pearson chi^2 (1358):	1360.89	Deviance	= 979.15
Dispersion (Pearson):	1.002126	Dispersion	= .7210243

a	Coef.	Std. Err.	z	P > |z|	95% Conf.	Interval
s	.0562237	.0172093	3.267	0.001	.0224941	.0899534
cons	.0882951	.0121144	7.288	0.000	.0645513	.1120388

These individuals comprise nearly all of those with dp essentially zero. What seems fairly clear is that either some people do not exhibit the kind of responses that Spottiswoode and May were searching for, or else the equipment failed in about half of the cases. In either circumstance, it would seem prudent to delete these cases from a secondary analysis. Even though the argument is *post hoc*, it is surprising that such a large number of people should have exhibited no responses. If Spottiswoode and May had anticipated this, then given the meticulous care in their design, they would have put each individual through a "run-in" to establish that they gave responses at least occasionally. Since they didn't, this finding falls outside the paradigm of the original design, and justifies a secondary analysis. Here is the output, from the same statistical routine shown above, but now with the complete nonresponders eliminated (Table 2).

Even with nearly half of the data discarded, the effect has risen to about 5.62%, nearly doubling that under the primary analysis, and the statistical significance is unchanged. This must be interpreted, of course, as the effect among those who give some evidence of a skin conductance response. But equally obvious, it makes little sense to perform a summary analysis that includes individuals who appear unsusceptible to the phenomenon under investigation (men are, for example, excluded from etiologic studies of ovarian cancer).

This secondary analysis would seem to support the conclusions of Spottiswoode and May even more strongly. The histogram shown above contains, however, a further message. If one looks at dp (the excess proportion of anticipatory responses under noise stimulus) over all individuals, there is marked variability. Some people do not anticipate the noise, while others do. No simple binomial model can account for these variations (the p-value is so small as to be scarcely computable, justifying the concern about heterogeneity expressed in the previous section). In other words, this is not to be explained

away as population sampling variability. The conclusion seems to be that nearly half of people (like those sampled in this study) make no anticipatory response at all, and of the remainder some show positive or negative values of dp pretty much by chance, and then there is a smaller number that either show modest, or in some cases extraordinarily large values of dp.

From this observation follows a natural question. Do those people who anticipate the noise stimulus tend to be those who give more anticipatory responses in general? The results of an analysis allowing for a trend shows a significant positive result (p uncomputably small). Those with only one total response had a modest estimated odds ratio of 1.67 (between anticipatory response and noise stimulus), while those with five responses had an estimated odds ratio of 16.6. The odds ratios rose regularly with the total number of responses, showing that this is not an artifact, but a genuine dose-response relationship.

The Misbehaving Random Numbers

There is one further mystery in the data of Spottiswoode and May, which turns out to be the most interesting. Examination of their simple primary analysis shows that there are 1319 instances of a noise stimulus (52.8%), but only 1181 of a silence stimulus. This disparity is significant with a two-sided p-value of 0.006. In other words, the random number generator that determined the stimulus type, which Spottiswoode and May defend so strongly in their article, produced more noise stimuli than can be accounted for by chance.

Careful reading of the article shows that the type of stimulus was determined after any anticipatory response. Although backwards causation cannot be ruled out in general, once one allows causes to work back in time, extremely difficult problems arise in justifying any of the conventional experimental designs. Is there any way to interpret these data while preserving forward causation?

The answer is provided by another analysis (Table 3), using the same routine as above, but now with ‘‘s'' (the noise stimulus) being explained by ‘‘a'' (the anticipatory response). Again to interpret the results, in about 51.9% of cases without an anticipatory response, there was a noise. This has a p-value of 0.069 (testing the theoretical value of 50%), worrisome but not technically statistically significant, so we can just barely believe that the random number generator worked in the absence of an anticipatory response. Among cases where there was an anticipatory response, on the other hand, there was a 13.1% rise in the occurrence of noise stimuli, statistically significant with a p-value of about 0.001. The evidence is stronger that anticipatory responses cause noise stimuli than the other way around. (The results are essentially identical if the complete nonresponders are eliminated.)

There is, of course, a third possibility—that some unidentified phenomenon was responsible for influencing both the skin conductance responses and the random number generator. While this is clearly speculative, surely part of the

TABLE 3
Stata Analysis of Spottiswoode and May's Data

General estimating equation for panel data		Number of obs	= 2500
Group variable:	id	Number of groups	= 125
Link:	identity	Obs/group min	= 20
Family:	binomial	Obs/group avg	= 20.00
Correlation:	exchangeable	Obs/group max	= 20
		chi^2 (1)	= 11.48
Scale parameter:	1	Prob < chi^2	= 0.0007
Pearson chi^2 (2498):	2499.65	Deviance	= 3447.20
Dispersion (Pearson):	1.000662	Dispersion	= 1.379984

a	Coef.	Std. Err.	z	P > |z|	95% Conf.	Interval
s	.1311564	.0387027	3.389	0.001	.0553006	.2070123
cons	.5191592	.0095617	54.295	0.000	.5004186	.5378999

scientific enterprise is to press one's analysis in an attempt to uncover novel explanations.

Summary

Reviewing this article has been difficult. In part this was due to the elegance and intricacy of its design, but it has also been due to the inadequacy of its statistical analysis. I would strongly encourage authors submitting to *JSE* either to employ the aid of a professional statistician, or at least to have their work reviewed by one, before submission. Failing this, I would encourage them to hear the complaints of statistical reviewers as being potentially helpful, rather than viewing them as an artificial barrier to publication.

The association between anticipatory responses and noise stimuli has been established by the *data* of Spottiswoode and May, if not by their specific *method* of analysis. Their data also suggest the extremely intriguing possibility that the phenomenon they have discovered is not one that is dispersed across all individuals in equal measure, but rather that it is somewhat idiosyncratic, exhibited to a modest or greater degree in a relatively small fraction of people.

Finally, a closer examination of their data suggests that rather than finding a ‘‘prestimulus response'', to use their term, the authors have instead discovered yet another way in which humans might influence physical random number generators. Without this interpretation, they have no way to account for the excess of noise stimuli in their experiment.

I again want to thank Spottiswoode and May for providing their source data for this comment.