心理学报 2010, Vol. 42, No.1, 1–11
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2010.01

Looking at Individual Subjects in Research on Judgment and Decision Making (or anything)

Jonathan Baron
(Department of Psychology, University of Pennsylvania, USA)

Abstract: Many questions in judgment and decision-making research, and, indeed, in experimental psychology generally, concern the existence of effects, and the explanation of effects shown to exist. These questions do not concern the prevalence of effects in any particular population. It is thus appropriate to look for effects in single subjects. If one person shows the effect, then it exists. This argument implies that it is sometimes appropriate to test effects across cases or rounds, without testing across subjects. It also implies that, in some experiments, effects in opposite directions may exist. I recommend looking for such effects by carrying out statistical tests on individual subjects. I describe a few methods, varying in formality, that can be used to deal with the inevitable problem of doing multiple tests of the same hypothesis: probability-probability plots; tests of the distribution of *p*-values; and correction for multiple testing with step-down resampling. I also present a few examples, some of which show effects in both directions and some of which do not.

Key words: multiple testing; one-tailed tests; experimental methods; omission bias

Introduction

In this paper I present an approach to data analysis in my field, which is the experimental psychology of judgment and decision making. The idea is to analyze the data from individual subjects, looking for effects in both possible directions, not just an overall effect in one direction. This approach is surely relevant to other fields. I will not dwell on these, but I should mention that the issues are closely related to a discussion in psycholinguistics, which goes back to Herbert Clark's (1973) landmark paper on "The language as fixed-effect fallacy." Indeed, my interest in this topic began when I read that paper and wrote a comment on it (Baron, 1975).

Within-subject experiments

In a typical experiment, each of several subjects (usually about 80) responds to several cases. The subjects are a "convenience sample," not intended to be representative of any particular population. Sometimes the cases are chosen to be compared with each other. For example, in one recent study (Baron & Ritov, 2009a), we asked for judgments of appropriate punishment for various offenses. The offenses were in pairs. The members of each pair differed in whether the offense was easy or hard to detect; for example, the offense of a waitress not reporting tips on her taxes is easier to detect if the customers pay with credit cards than if they pay with cash. In other experiments in the same paper, subjects were asked directly about the issue of interest after each case, e.g., whether the punishment should be more harsh (coded as 1), less harsh (-1), or the same (0) if the offense were more difficult to detect.

The cases (e.g., different offenses) in such studies are typically chosen for the simple purpose of getting more data out of each subject than if we used only one case per subjects. For me, they are not representative samples of any particular

Received date: 2008-10-06
Supported by a grant from the U.S.-Israel Bi-national Science Foundation (Ilana Ritov, co-PI).
Correspondence concerning this article should be addressed to Jonathan Baron. E-mail: baron@psych.upenn.edu

population of cases, although they can be representative (and some have argued that representative design is important, e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991). In this study, they were simply offenses that were easy to think of and (we hope) easy for the subjects to understand. But the use of several cases opens many possibilities for data analysis.

If we had the data from a single subject we could ask whether probability of detection had a significant effect for that subject. If we had 20 cases, 10 of each type, we could look at the penalty judgments for each of the 20 cases and do a paired *t* test comparing the cases (with 9 *df*). If, alternatively, we had 10 cases, with a direct question about the relevance of detection probability, on a rating scale in which 0 was the midpoint representing no effect, we could also do a *t* test (or some nonparametric equivalent) asking whether the mean answer was greater than 0.

The most common method for analysis of such experiments is to compute an effect for each subject, and then test across subjects, so the *df* of a *t* test would be 79 for 80 subjects. This is also what we would do if each subject had only one pair of cases. The use of many pairs still reduces the error in estimating the effect for each subject, so it should also reduce the error of such a test across subjects, increasing its power.

It is also possible to do the entire analysis across cases, so that the final test would be just like the test for one individual subject, but each number would consist of the average response of all the subject to the case in question. Some researchers do this.

The point made by Clark (1973) was that, in experiments on language, we need to pay attention to both subjects and cases, so that we can "generalize to the population," as if either the cases or subjects were representative samples of relevant populations.

Existence versus generalization

In this article, I want to extend the argument I made in Baron (1975). I argued there that many — I would say "most" except that I do not know how to define the population — hypotheses in my field, and in experimental psychology generally, are one tailed. For example, in our study of punishment (Baron & Ritov, 2009a), we wanted to ask whether subjects followed the principle of compensating for lower probabilities of detection with increased penalties, so that, from the offender's point of view, the expected punishment was independent of probability of detection. It is possible that some subjects would find some reason to *reduce* penalties with lower probabilities of detection, but this reduction would have to result from some entirely different mechanism than the one of interest to us.

I argued further that a one-tailed hypothesis of this sort asks for a demonstration of existence. In principle, if one subject shows the effect, it exists. Thus, I argued, a test across cases is sufficient. I assumed that we had no reason to think that cases should differ in the direction of the effect. Whether we did or not, then a test across subjects, with a single case or pair of cases, or more, would also be sufficient to demonstrate existence.

The same argument could be made in terms of cases rather than subjects. If we demonstrate an effect for two versions of one case, and the versions are well controlled, differing only in the property of interest, then the effect exists. Many studies draw conclusions from effects in a small number of cases. The classic findings in judgment and decision making rest on particular cases, such as the Asian disease problem (Tversky & Kahneman, 1981, showing framing effects that lead to risk aversion when the problem is stated in terms of gains and risk seeking when it is stated in terms of losses). Many papers in the field are based on a small number of cases, with no significance test across cases.

Implicitly, these arguments are recognized in recent proposals that respond to Clark's concerns about subjects and items. For example, Baayen, Davidson, and Bates (2008) recommend the use of mixed models (fixed effects and random effects)

with crossed random effects (subjects and cases) to analyze data in which both subjects and cases are sampled. Of interest is the fact that this approach can yield significant results when only a couple of cases are used, or only a couple of subjects. (The same is true of more standard approaches that treat subject effects only as random variables.) If we were to employ Clark's method, we would have to test across subjects and across cases, so that we could generalize to both. With only two subjects, it would be extremely difficult to get a significant result, yet, with a mixed model, a result can be highly significant with two subjects, only one of whom shows an effect[1].

What does it mean, anyway, to generalize to a population? I think it means this: If we think of our sample as a random sample from a population, and we find in the sample that an effect is significantly positive, in a two-tailed test, then we can conclude that the effect is positive in the population as well, with the usual caveats about the possibility of false rejections of the null hypothesis. For example, in an opinion poll about a forthcoming election, we try to sample randomly from the population of voters. We can, in principle, ask whether the support for one candidate is significantly higher than that for another.

Experimental psychologists rarely sample from populations of people, or (as I just noted) cases. We select subjects without thinking much about any population. Traditionally, psychologists have used students[2]. The phenomena that have been discovered this way are unlikely to go away. Recently, some American psychologists have made an effort to use random samples of Americans (at considerable cost). I would see some value in this if they were interested in predicting American elections. But most psychological questions are not about predicting what a particular population will do. They are, in some sense, about people in general.

But it is impossible to sample "people in general." Americans are not representative of English speakers. English speakers are a minority of people who are currently alive (even counting all Indians as English speakers). If we are optimistic, people currently alive are a tiny minority of those who will exist in the future. Thus, in an important sense, every sample is a "convenience sample," the derogatory term used by those who seem to accept the view that research necessarily involves generalizing to populations.

Some questions are inherently two-tailed. Does capital punishment for murder increase or decrease the homicide rate? We have evidence for both deterrence and "brutalization" effects, and the question is which effect is stronger. The answer surely depends on the population. For example, Shepard (2005) presents evidence that deterrence effects are larger when states in the U.S. use the death penalty a lot, but brutalization effects are larger in other states. (But see Donohue & Wolfers, 2005, who argue for no effects.)

Questions about individual differences are sometimes more difficult. Many of these depend on such issues as whether the correlation between A and B is higher than that between A and C. Such differences in correlations can determine the

[1] Here is some R code to demonstrate this for two subjects, only one of whom has a positive mean, with the MCMC p-value for the intercept being highly significant, and the raw t value over 4:

```
library(languageR)
S <-gl(2,100,200)
X <-rnorm(200)+5*(S!="1")
d1 <-data.frame(X=X,S=S)
l1 <-lmer(X 1 + (1|S),d1)
pvals.fnc(l1)
```

[2] Roddy Roediger (2004) put it this way: "Most of the people participating in psychology experiments are college students. We get a lot of grief about this, I know, and we often feel abashed and ashamed. I don't know why. All scientists use samples and techniques that are readily available and can be adapted to make rapid scientific progress. I study human memory, and to me the college student is the ideal experimental animal. Millions of years of evolution have designed a creature that is a learning and memorizing marvel. Students in my experiments have also been carefully selected through 12 or more years of education before they get to my lab. Only the ones who have shown, year after year, that they can learn and remember material in courses make it to my experiments. The world could not have arranged a more ideal subject,. . . "

outcomes of factor analysis, for example, which is central to the study of individual differences. The pattern of correlations could vary from population to population. Differences among correlations are also central to selection of items on personality scales, which can thus lose their intended meaning when they are transferred from one population to another. Cross-cultural researchers are of course aware of these problems.

The point, though, is that these questions are inherently two-tailed, unlike most of the questions that experimental psychologists ask, which are mostly about the existence of mechanisms that lead to an effect in a particular direction. As I noted, sometimes different mechanisms lead to effects in different directions, but these are best seen as two different one-tailed hypotheses. We are not usually asking which effect is bigger but, rather, whether each effect exists.

Regardless of whether questions are inherently one tailed or two tailed, how can we deal with the sampling problem? Again, I put aside those cases where the population is well defined, such as voters in an election. Much of psychology is concerned with human beings in general, or, at least, human beings as we exist now. (In the future, we may find ways to change our nature, or human nature may change by accident.) My answer is that we should strive for diverse samples of subjects, and we should look carefully at individual differences. In particular, we should look for the existence of subjects who show each possible effect that the experiment can show. We may find only a few who show an effect of one sort or another, but some other sample might show a larger proportion, so we cannot just dismiss these subjects as oddballs. Diversity within the sample increases the chance of finding such rare subjects, if they exist at all. And, if we do not find them, a diverse sample increases our confidence that they do not exist.

Effects in both directions

Let us explore further the fact that subjects may differ in the kind of effect that dominates. (Cases might differ too, but I shall put that possibility aside for the rest of this article; the argument for cases is analogous.) For example, some subjects may think that punishment should be less severe when the probability of detection is lower. For example, they might see punishment as unfair and arbitrary when most offenders do not get caught.

Another example from my own research concerns “omission bias,” the general finding that people tend to tolerate harms of omission more than harms that result from action. They are thus generally unwilling to cause harm in order to prevent greater harm (Ritov & Baron, 1990; Spranca, Minsk, & Baron, 1991). There are many reasons for this bias (Baron & Ritov, 2009b). For example, people seem to base judgments on a concept of causality that is more appropriate to the physical sciences than to decision theory or moral judgment. They think that it is worse to cause a bad outcome through a connected chain of physical events than to cause it by failing to prevent it. Yet failing to prevent can be considered causing in a legal context, where “the event would not have happened but for your failure to prevent it” is often sufficient for legal sanctions.

Yet, even in one of our earliest papers (Spranca et al., 1991) we noticed that some subjects seemed to show the opposite bias, but we carried out no statistical test of whether their existence was real, or just a consequences of random error. Intuitively, we suspected that some people, in some situations, might feel a responsibility to act, even if the action was so risky as to be expected to do more harm than good. We thought of surgeons as a possible example. Military commanders might be another, although usually their training favors an appropriate balance. If such an action bias exists, then it would have a very different basis than the omission bias. It is not simply the opposite; people would not think of physical causality as *less* serious than indirect causality. Rather, it would arise, perhaps, from an exaggerated sense of responsibility, a feeling that the expectation of someone in a particular role was

to act.

In this example and all the examples I can think of, we are not interested in simply rejecting the null hypothesis of no effect, because effects in different directions have different interpretations. Thus, they are not simply "two-tailed tests." When we look for both effects, we are, in essence doing two one-tailed tests.

Another way to make the point is that an overall test of significance can show an effect in one direction, but it does not rule out the existence of an opposite effect, even in the same experiment.

Statistical methods

The problem is that, typically, one of the two effects predominates in a given sample of subjects and cases. In this section, I shall discuss some statistical methods for demonstrating that some subgroup of subjects (or cases) shows an effect opposite to the predominant effect. Of course, exactly the same methods can be used to show that a subgroup shows the predominant effect; this demonstration is rarely needed, however, since the same conclusion follows from, say, an overall test across subjects. The idea is that we test each subject and then look for subjects showing individual significant effects, even after correcting for the fact that we are testing so many. We cannot just take the significance levels of individual tests at face value because, for example, 5% of these tests will be significant at the .05 level even if the null hypothesis is true. With 80 subjects, we are very likely to find one or two who show "significant" effects by chance.

The statistical literature most relevant is that on correction for multiple tests. This literature is vast, and I am not an expert on it, so I shall merely describe a few of the methods that I use myself. Shaffer (2002) and Dudoit, Shaffer, and Boldrick (2003; see also Ge, Dudoit, & Speed, 2003) provide excellent introductions.

An important distinction in this literature is that between two common measures: family-wise error rate (FWER) and false-discovery rate (FDR). (There are others. These are the two most used.) Both measures assume that we have some criterion for rejecting a null hypothesis. For example, for a single test, we might say that we will reject the null hypothesis if the *p*-value is .05 or less, or if t is 2 or more.

The FWER is a generalization of "type 1 error," that is, the probability of falsely rejecting the null hypothesis. For a single hypothesis, this is just the *p*-value. The "family" is the group of hypotheses being tested. Here it is the subjects. So maintaining a FWER of .05 means that there is a .05 probability (or less) of rejecting the null hypothesis — that no subject shows an effect — given that the null hypothesis is true. For example, a Bonferroni correction divides the stated *p*-values by the number of *p*-values, so that, if one corrected *p*-value is less than .05, we can reject the null hypothesis that no subject shows an effect. Of course, with 100 subjects, this requires a stated *p*-value of .0005 in the original tests. The Bonferroni method is not very powerful, but other methods, as we shall see, are substantially more powerful.

The FDR is the proportion of rejected null hypotheses that are true (no effect). Note that the FDR cannot be computed from the FWER. The FDR also depends on the proportion of true null hypotheses. For example, if 100 subjects all have no real effect (null is true), and 5 of them yield a *p*-value of .05, then the FDR is 1.00 (100%). All of the 5 rejected hypotheses are falsely rejected. But, if 50 subjects yield .05 or better, we can expect the FDR to be much lower. Most (but not all) of the 50 will probably be true rejections of the null hypothesis. The FWER is the same in these two cases, because it is calculated on the assumption that all the null hypotheses are true. Note that the FDR varies with the number of results that are "significant," so it is calculated contingently on the results. The FDR depends, of course, on the p-level. You can set a p-level to generate a given FDR; the given FDR is called the q-value.

It is interesting to ask what we would think about the FDR if it had been invented decades

before the idea of type-1 error, the reverse of what happened. Suppose I tell you that I did an experiment with 100 subjects, that 20 of them pass my test for showing significant results, and, further, that my test has a FDR of .10. You know that the expectation is that 18 of the subjects are truly significant. Of course it could be that 0 are truly significant, but this is very unlikely. (To know how unlikely, we need to know the FWER!) For some purposes, knowing that 18 subjects show an effect might be very informative. However, if we start with the basic assumption I have made throughout, namely, that we are primarily interested in the existence of effects, we would be unsatisfied. We would still want to know the FWER, because that is the measure that concerns existence versus non-existence of an effect. But the FDR can be informative, and when this is sufficiently low we may still be satisfied.

I shall not discuss FDR further. It is surely of interest if we want to estimate how many subjects in our sample are truly showing an effect, or if we want to make some decision that requires classifying subjects and we have a specific payoff function for true and false classifications. It is also a bit more formal than just looking at the graph I shall describe in the next section. But I am more interested in asking whether *any* subjects show an effect in a given direction. For further discussion, see Benjamini and Hochberg (1995), Ge, Sealfon, & Speed (2008), Storey (2002), and Storey and Tibshiran (2003).

Distributions of p-values

A very simple method for looking at significance of individual subjects is to test each subject and then look at the distribution of *p*-values. Under the null hypothesis that no subject shows a real effect, the *p*-values will be uniformly distributed between 0 and 1. This follows from the fact that, by chance, 5% of them will be .05 or lower, 10% will be .10 or lower, and so on. If we put the *p*-values from our subjects in order from lowest to highest, we should see that the 5% of the lowest values should be .05 or lower.

We can plot these *p*-values against their percentile rank[3]. An example is shown in Figure 1 (from Bonner & Newell, 2008). Each point represents the *p*-value of one subject. The horizontal axis is the percentage of *p*-values less than or equal to the given value, that is, its percentile rank. Under the null hypothesis of no real effects, the *p*-values should fall on roughly on the diagonal. It is apparent here that many *p*-values are less than .05, a lot more than 5%, and even more *p*-values are between .95 and 1. The experiment in question, in a test across subjects, showed an effect consistent with the high *p*-values, but it is apparent that many

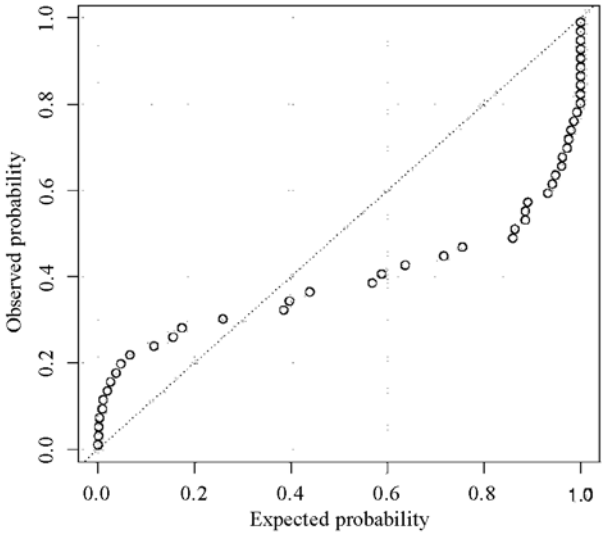

Figure 1: *P-P* plot of the observed probability of individual *p*-values against the expected prob-ability. Under the null hypothesis of no effects, a uniform distribution along the identity line is expected. Points above 0.5 on the expected probability axis indicate an effect in one direction, with the smallest *p*-values approaching 1. Points below 0.5 on the expected probability axis indi-cate an effect in the opposite direction, with the smallest *p*-values approaching 0. (From Bonner & Newell, 2008.)

[3] Here is some R code for plots of this sort, where V is the variable of interest:

```
# get one-tailed p-values and sort them
Ordinate <-sort(apply(V,1,function(x)
t.test(x,alt="greater")$p.val))
n <-length(Ordinate)
Plotpos <-seq(0.5/n, (n -0.5)/n, by = 1/n)
plot(Ordinate, Plotpos, xlab="Expected probability",
ylab="Observed probability")
abline(0,1,lty=3)
grid()
```

subjects show the opposite effect. (At issue was whether risks seemed worse when described in terms of deaths per year or deaths per day. The predominant answer was "per year.") In an example like this, the graph seems convincing by an "intra-ocular test"; the result hits you between the eyes and no further analysis is needed.

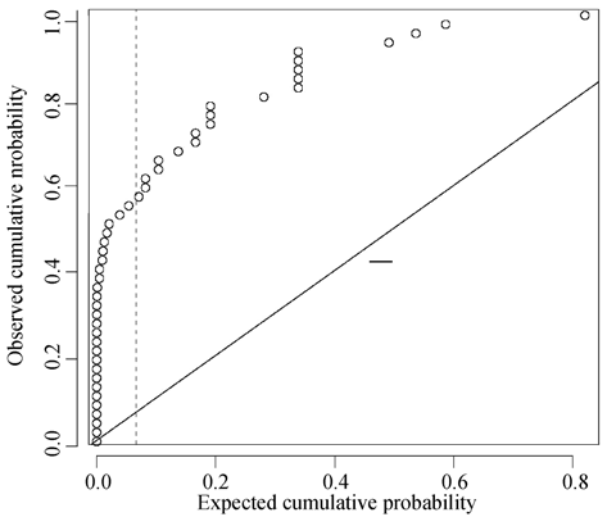

Figure 2: *P-P* plot for Experiment 2 of Baron and Ritov (2009a).

Figure 2 shows another example from the punishment study described at the beginning of this article, specifically, Experiment 5 of Baron and Ritov (2009a). In this experiment, subjects were asked about the appropriate level of punishment for crimes. Within each page, they were asked what the punishment should be if 1% of the committed crimes were detected and if 90% were detected. Many subjects made no distinction, but a large group of subjects gave harsher punishments when the probability of detection was lower, in accord with the economic theory, which holds that increased severity can compensate for the lower chance of detection, so as to maintain sufficient deterrence. In this case, no subjects showed the opposite effect. I shall return to this example.

Although the graph is sufficient in many cases, a formal test is also possible. For example, we could do a binomial test (proportion test) asking whether the proportion of *p*-values less than .05 is greater than the "null hypothesis" value of .05. Or we could use some other criterion such as .10. In the last example, 26 *p*-values were less than or equal to .05, out of 48. (The rest were undefined because all the responses were 0.) This is of course much greater, and significantly greater, than the number of such *p*-values expected by chance (48/20, or 2.4).

The trouble with this approach is that the result will depend on what cut-off we use. I know of no way of setting a non-arbitrary cutoff. Laxer cutoffs are more appropriate when the power to detect an effect for each subject is lower, e.g., when each subject does a small number of cases.

Using a somewhat different approach, Sundali and Croson (2006) demonstrated that individual gamblers in the game of roulette differed in whether they were susceptible to the "hot hand" effect (in which they thought of a given roulette wheel as having a stock of luck that would continue) or the "gambler's fallacy" (in which they believed that one outcome became more likely over rounds in which it did not occur). These effects tended to be incompatible. Sundali and Croson calculated *p*-values for each subject, for each effect. They compared the results to the expected uniform distribution using a Kolmogorov-Smirnov test. Although this method solves the problem of arbi-trariness, it lacks power, as it looks for any departure at all from the hypothesized distribution. (In this case, it worked, despite the lack of power.)

Corrections for multiple tests

The best way to determine whether any subjects shows an effect is to test each subject and correct for multiple tests, using a given FWER. The well-known Bonferroni correction is one way to do this. Suppose we set the acceptable FWER at .05, and we carry out two independent tests of the same hypothesis (e.g., 2 subjects). To maintain a FWER of .05, we must set our α level (the level for what we call "significant") at .025. Assuming that the tests are independent, then this yields a .05 chance of a Type I error (rejecting a true null hypothesis). Thus, we change our α level as a function of the number of tests. With 100 subjects, then we must

look for at least one subject with a result that is significant at .0005, in order to claim that at least one subject shows an effect significant at .05. This might be literally impossible; for example, if the data consist of two-option choices, and we want to claim that one choice is more frequent than the other, with 10 choices, the best that can be done is .0010 (one tailed), if all 10 responses go in the hypothesized direction. Even if we use a t test on continuous numbers, such a highly reliable result is difficult to get.

It turns out that it is unnecessarily difficult. The Bonferroni method is unnecessarily conservative. A variety of methods have been invented that avoid the various problems. These methods are used in fields in which the number of tests is far larger than the number of subjects in most psychology experiments: genetics and brain imaging. In genetics, a researcher might test hundreds of genes for association with some trait, using fewer subjects than genes. In brain imaging, such as the use of functional magnetic resonance imaging (fMRI), a researcher might test thousands of locations in the brain (voxels) for association with a particular stimulus or response, using only a hundred trials of each type. If you think about these cases, it is apparent that the Bonferroni correction is unfair. Particularly in the fMRI example, the voxels that are next to each other tend to behave very similarly. They are not independent. This is one problem. Analogously, if we test 100 subjects on the same 20 cases, subjects might show patterns of similarity to each other in the way they respond to the 20 cases. Groups of subjects might behave essentially identically, just as voxels in the same brain region would behave identically. Thus, instead of 100 subjects, we might have what amounts to a much smaller number of subject types. The correction should be applied to the number of types, not the number of subjects.

The example I just gave is just a way to see why the Bonferroni method might be conservative. In fact, we don't have to look for types. Methods have been developed that, in essence, take into account the correlations among subjects in other ways. I shall not (and could not) review all of these methods. (Dudoit et al., 2003, and Westfall et al., 2001, review most of them.) The best for our purposes seem to be the step-down resampling methods of Westfall and Young (1993) as implemented in the Multtest package for R (Ge, Dudoit, & Speed, 2003; R Development Core Team, 2008).

The basic idea is this. Suppose we have a matrix of numerical responses, with 100 rows as the subjects and cases as the 20 columns. Our hypothesis says that responses to cases 1–10 should be higher than those to matched cases 11–20. (If, instead of 20 cases, we have 10 cases, and our hypothesis is that the responses to these cases should be greater than zero, we make up an additional 10 cases with responses of zero to each, for each subject.) We do a t test for each subject. Suppose we want to know the "true significance" of the subject with the highest t value (or lowest p value). To determine this, we simulate the test by "permuting" the columns. That is, we switch the columns around at random, thousands of times. Each time, we test cases (columns) 1–10 against cases 1–20. Then we simply count the proportion of tests that yield a t value equal to or higher than that of the subject with the highest t value. This estimates the probability of getting that t value by chance, if no subject showed a true effect. (The permutations, by ignoring the assignment of cases to conditions, assure that there will be no effect overall.) The procedure continues for subjects other than the one with the highest t value, but here we are interested in asking whether any subject yields a truly significant result.

To see how this helps, imagine the extreme case in which all subjects had a t value of 2.83 because they all made the same responses to every case. This corresponds to a one-tailed p of .01. But of course a Bonferroni correction, with 100 subjects, would make this nonsignificant. Yet there is essentially 1 subject, for purposes of correction. One percent of the permutations, if we did enough

of them, would yield a t of 2.83 or higher.

The relevant functions in the Multtest package (Ge et al., 2003) are mt.maxT and mt.minP. The mt.maxT function is based on t tests; the minP function is based on *p*-values calculated from t tests (by default). The minP packages, which takes longer to run (but not very long), is more useful when subjects have missing data, so that the degrees of freedom are unequal. (In that case, t values do not correspond to p values.) These functions rank order the subjects in terms of their corrected p values[4].

Another experiment to try with the Multtest package is this. Generate a fairly large matrix of random data and run mt.maxT on it[5]. Then make it twice as long (twice as many subjects) by repeating the whole thing twice. This will generate two sets of random subjects, but each subject will have a perfectly correlated twin in the other set. Run the procedure again. You will find that the corrected p values of the best subjects are unchanged. That is as it should be. But, if you just generate more random subjects, the p value of the best one will increase[6].

[4] Here is an example of some R code for mt.maxT, beginning with a data matrix V in which rows are subjects, columns are cases, and the entries are scores, with a null hypothesis that they are zero:

```
library(multtest)
# create a matrix m with alternative columns of V and 0
m <-matrix(rbind(V,matrix(0,nrow(V),ncol(V))),nrow(V),)
mt <-mt.maxT(m,classlabel=rep(0:1,ncol(V)),side="lower")
```

[5] It doesn't work well with tiny data sets.

[6] Here is the relevant code:

```
library(multtest)
# generate the data
ns <-nv <-10 # 10 subjects, 10 variables
m <-matrix(0,ns,nv*2) # 20 columns
m1 <-rnorm(ns*nv)
m2 <-rnorm(ns*nv)
m[,2*1:nv] <-m1 # fill the even-numbered columns
m[,2*1:nv-1] <-m2+1 # fill the odd-numbered columns
# first analysis
mt <-mt.maxT(m,classlabel=rep(0:1,10),side="lower")
# the following shows that doubling has no effect
# because of correlations
mt.maxT(m,classlabel=rep(0:1,10),side="lower")
mt.maxT(rbind(m,m),classlabel=rep(0:1,10),side="lower")
# but this has an effect because we add noise
m3 <-matrix(rnorm(ns*nv*2),ns,nv*2)
mt.maxT(rbind(m,m3),classlabel=rep(0:1,10),side="lower")
```

Another example: omission bias

Spranca et al. (1991) speculated that some subjects showed an action bias, favoring actions that led to a harmful outcome over omissions that led to an equally harmful, or less harmful, outcome, in contrast to the opposite bias toward harmful omissions, which has repeatedly been found to dominate results when tested across subjects (Baron & Ritov, 2009b). Baron & Ritov (2004) applied the step-down resampling procedure just described to test for the existence of this reverse bias, and we found it. We explained omission bias, the dominant result, in terms of a heuristic based on direct causality. The opposite bias toward action may result from aspects of the context; for example, when subjects are asked about vaccination, they may think that vaccination is a generally good thing. Thus, the context may work against finding the normal bias toward omissions.

To illustrate further the analysis of individual subjects, I now present data from Baron & Ritov (2009b), Study 3. Seven of the items in this study assessed omission bias, in a situation where the consequences of acts and omissions were the same. For example, one item read:

> Joe is angry at a neighbor.
>
> A. When the neighbor's car starts rolling down a hill, Joe sees that a brick in front of it will stop it from rolling. Joe pushes the brick away, and the car suffers expensive damage.
>
> B. When the neighbor's car starts rolling down a hill, Joe sees that it is possible to stop the car by pushing a brick in front of it. Joe does nothing, and the car suffers expensive damage.
>
> Is option A morally wrong? [yes, no, equal or hard to say]
>
> Is option B morally wrong? [yes, no, equal or hard to say]
>
> Which option is better or less wrong? [A, B, hard to say]
>
> . . . [Other questions followed.]

None of the seven cases involved vaccination.

I assessed omission bias by counting "equal" as 0, "yes" as 1, and "no" as -1 and then subtracting

the second question from the first, and adding the third, thus combining the two morality questions with the comparison question. The mean bias was 0.35 on a scale ranging from -3 to 3; this was highly significant across subjects and even significant across the 7 cases (t_6 =2.67, p =0.0368, two tailed). The p-values for t tests on 75 individual subjects are shown in Figure 3. (Some subjects had all 0 values and were omitted.) It is apparent that no subjects showed the opposite effect, as found in the vaccination cases used by Baron and Ritov (2004). This results is consistent with the claim that vaccination is a special case because of the pro-vaccination norm. The mt.maxT method (described earlier) yielded only two significant subjects. Evidently, with only seven cases, it is difficult to find many that survive correction. However, 15 of the 75 subjects had results significant at .05, two tailed, significantly more than expected by chance.

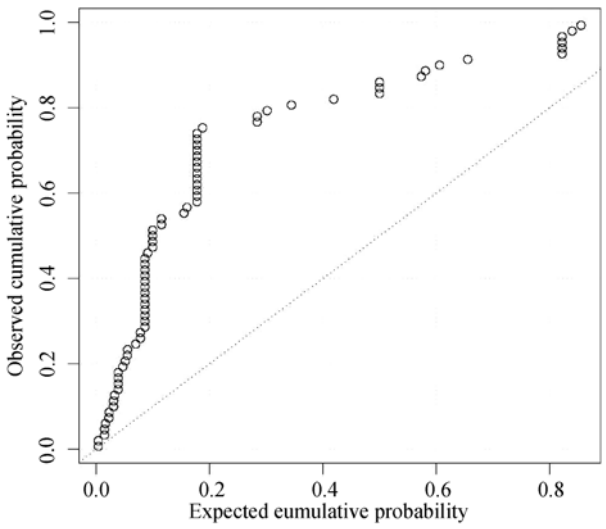

Figure 3: *P-P* plot for Study 3 of Baron and Ritov (2009b).

Conclusion

I have discussed the value of looking at individual subjects in research on judgment and decision making. I have presented some examples in which individual subjects show effects in both directions, and other examples in which only one direction of effect is found. This seems like an important distinction.

It is also consistent with the view that, in much of experimental psychology, we are interested mainly in the existence of effects, and, given that they exist, their explanation. We are not so much interested in their prevalence, if only because we cannot begin to sample the population of real interest, all human beings, present and future.

Such analyses might be useful in other areas. For example, a great deal of research shows that people are biased toward beliefs that are already strong, both in the selection of evidence and the assimilation of that evidence. (Baron, 2008, ch. 8, provides a review.) Yet it is possible that some people do not show these effects, and others might even be too self-critical.

Perhaps a more important case is the study of treatment effects, in education, in psychotherapy, and elsewhere. Famously, it is often been claimed that psychotherapy is beneficial on the average, but this is only because it hurts some people while helping more people. (The ratio of 1 to 2 is commonly mentioned.) These claims seem to be based on comparison of measures taken before and after treatment. Yet these measures contain error. It is quite possible that nobody is hurt. If we could evaluate treatment effects for individuals by looking at data collected over longer periods of time, we might be able to answer this question. Perhaps some sort of regression discontinuity design could make this possible.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, in press.

Baron, J. (1975). Stimuli and subjects in one-tailed tests. *Bulletin of the Psychonomic Society, 6*, 608–610.

Baron, J. (2008). *Thinking and deciding* 4th edition. New York: Cambridge University Press. (Chinese translation of 4th edition to published by China Light Industry Press.)

Baron, J. & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74–85.

Baron, J. & Ritov, I. (2009a). The role of probability of detection in judgments of punishment. *Journal of Legal Analysis, 1*, 553–590.

Baron, J., & Ritov, I. (2009b). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral Judgment and decision making*, Vol. 50 in B. H. Ross (series editor), *The Psychology of*

Learning and Motivation, pp. 133–167. San Diego, CA: Academic Press.

Benjamini, Y, & Hochberg, Y. (1995). Controlling the false discover rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 298–300.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Donohue, J. J. & Wolfers, J. (2005). Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review, 58*, 791–846.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science, 18*, 71–103.

Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data hypothesis. *Test 12*, 1–44 (with discussions on 44–77).

Ge, Y., Sealfon, S. C., & Speed, T. P. (2008). Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica 18*, 881–904.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roediger, R. (2004). What should they be called? *APS Observer, 12* (4), April. http://www.psychologicalscience. org/observer/getArticle.cfm?id=1549

Shaffer, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psycholog-ical Methods, 7*, 356–359.

Shepard, J. (2005). Deterrence versus brutalization: Capital punishment's differing impact among the states. *Michigan Law Review, 104*, 203–255.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B, 64*, 479–498.

Storey, J. D., & Tibshiran, R. (2003). Statistical significance for genomewide studies. *PNAS, 100*, 9440–9445

Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making, 1*, 1–12.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons.

Westfall, P. H., Zaykin, D. V., & Young, S. S. (2001). Multiple tests for genetic effects in associa-tion studies. In S. Looney (Ed.), *Methods in Molecular Bioloty, vol. 184: Biostatistical Methods*, pp. 143–168. Totowa, NJ: Humana.

决策与判断研究中的个体分析

Jonathan Baron
(Department of Psychology, University of Pennsylvania, USA)

摘　要　决策与判断研究中(甚至是实验心理学研究中)的许多问题关注某效应是否真实存在, 及其背后的解释是什么。这些问题不关注该效应在某一特殊群体中是否显著。因此, 可以通过分析单个被试来检验效应的显著性。如果有一个被试表现出了该效应, 那么, 这个效应就是存在的。根据这一观点, 有时也可通过跨案例或者轮次(across cases or rounds)分析来验证效应的显著性, 而不需要进行跨被试分析(across subjects)。这一观点也暗示在一些实验中可能存在反方向的效应。本文建议通过进行基于被试个体的统计分析来检验这样的效应, 并介绍了一些不同形式的方法：PP 概率图(probability probability plots); P 值分布检验(tests of the distribution of *p*-values); 分层取样多重检验的矫正(correction for multiple testing with step-down resampling)。这些方法都可以用于处理在对同样假设进行多重检验时无法避免的问题。另外, 本文也列举了一些例子, 其中有一部分例子存在反方向的效应, 另一部分例子不存在。

关键词　多重检验; 单尾检验; 实验方法; 冗余偏差
分类号　B841; B842.5