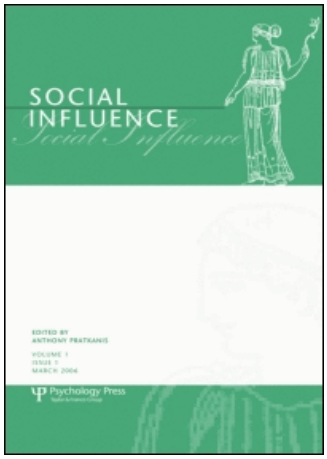

Social Influence
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100705

Reconsidering the impact of behavior prediction questions on illegal drug use: The importance of using proper analytic methods

Online Publication Date: 01 September 2007
To cite this Article: Schneider, Daniel, Tahk, Alexander and Krosnick, Jon A. (2007) 'Reconsidering the impact of behavior prediction questions on illegal drug use: The importance of using proper analytic methods ', Social Influence, 2:3, 178 - 196
To link to this article: DOI: 10.1080/13506280701396517
URL: http://dx.doi.org/10.1080/13506280701396517

SOCIAL INFLUENCE, 2007, 2 (3), 178–196

Psychology Press
Taylor & Francis Group

Reconsidering the impact of behavior prediction questions on illegal drug use: The importance of using proper analytic methods

Daniel Schneider, Alexander Tahk and Jon A. Krosnick
Stanford University, Stanford, CA, USA

Social scientists often habitually employ ANOVA methods when analyzing data from experiments when other analytic approaches are required instead. This paper illustrates how traditional analytic approaches can lead to incorrect research conclusions by reanalyzing data from a recent study by Williams, Block, and Fitzsimons (2006a). Because the non-negative dependent variable (illegal drug use) was super skewed and had a large majority of zero values, the use of improper statistical tests and the presence of just a few extreme, outlying observations produced the illusion that asking people to predict their likelihood of drug use increased that behavior significantly, when in fact it did not. The effect of behavior prediction questions on frequency of exercise also turns out to be non-significant when analyzed properly. As this example illustrates, experimental researchers should choose and implement appropriate analytic approaches carefully.

Although social scientists are often trained to use a wide variety of different statistical tools, in practice we often fall into habits and simply use the most common and familiar techniques, over and over again. The tradition of testing the effects of experimental manipulations using analysis of variance runs particularly deep: it is overwhelmingly common to see reports of *t*-tests and *F*-tests in print. Indeed, many of us are on "automatic pilot" when it comes to this aspect of our work, as if operating according to this script: "Step 1: Collect experimental data. Step 2: Conduct analysis of variance." A simple count of the articles in an arbitrarily selected recent issue of the

Address correspondence to: Daniel Schneider, McClatchy Hall, Stanford University, Stanford, CA 94305, USA. E-mail: daniel.schneider@stanford.edu

Jon Krosnick is University Fellow at Resources for the Future. We thank Patti Williams, Lauren Block, and Gavan Fitzsimons for providing us with their data sets for the analyses reported here.

http://www.psypress.com/socinf DOI: 10.1080/13506280701396517

Journal of Personality and Social Psychology (2006, *91*[2]), one of the most prominent journals reporting experimental studies, revealed that all 12 articles in that issue employed analysis of variance to assess effects.

Yet analysis of variance is biased and misleading when applied to particular types of data. Interestingly, this bias is one that is likely to perpetuate itself: it increases the likelihood that researchers will conclude that a manipulation had a statistically significant effect, when in fact it did not. Because significant effects are more likely to be published than non-significant effects (Atkinson, Furlong, & Wampold, 1982; Greenwald, 1975; Rosenthal, 1979; Shadish, Doherty, & Montgomery, 1989), any bias toward finding significance is likely to be appealing, even if unconsciously, to researchers seeking to make contributions to the published literature.

The problem has to do with assumption violation. All statistical approaches entail making assumptions, and we have come to act as if our results are always robust in the face of violating these underlying assumptions. But in one particular type of situation, they are not. And it is in this situation that we find ourselves with a recently published paper by Williams, Block, and Fitzsimons (2006a).

In short, Williams et al. (2006a) conducted *t*-tests typical of many experimental studies to assess the impact of manipulations and concluded that these manipulations' effects were statistically significant. Because of the importance of the phenomenon under study, and implications of their findings for the expenditure of millions of dollars per year by the federal government on surveys of children and adolescents, the work quickly gained visibility and set off a small firestorm of controversy and concern among survey researchers. But because the assumptions underlying the *t*-tests were violated in the data, the published findings were incorrect in a practically very important way.

We therefore take this opportunity both to correct the record on this study and to offer a more general presentation of the problem and solutions to it. This problem may be lurking in the data of many, many published studies, so the lessons learned in this instance may apply to the work of many experimental social scientists.

We begin below by describing the Williams et al. (2006a) study and its published findings. Next, we describe how the *t*-statistics reported in the paper were accidentally computed incorrectly. Correcting this accidental error makes one of the two central empirical findings disappear. Then we turn to the article's other central empirical finding and show that when properly analyzed, this effect disappears as well. In explaining why this is so, we describe a variety of data analysis techniques that social scientists should routinely consider using when analyzing experimental data. In the Appendix, we provide simple instructions on how to implement each test with widely available statistical software.

THE STUDY BY WILLIAMS ET AL. (2006a)

Williams et al.'s (2006a) study explored the effects on future behavior of asking behavior prediction questions. Continuing a long line of prior experiments, Williams et al. (2006a) gauged the impact of reporting the likelihood of exercising and using illegal drugs during the next 60 days on subsequent exercising and drug use among college students. Many past studies have indicated that asking people to predict their behavior or to indicate the likelihood of engaging in a behavior led them to perform that behavior more often (for a review, see Sprott, Spangenberg, Block, Fitzsimons, Morwitz, & Williams, 2006). Two streams of research have addressed this issue, one under the label of "self-prophecy effects" (e.g., Spangenberg, Sprott, Grohmann, & Smith, 2003), and the other labeled "mere measurement effects" (e.g., Morwitz, Johnson, & Schmittlein, 1993). Recently, researchers have come to think of these as the same phenomenon (e.g., Sprott et al., 2006).

The self-prophecy effect hypothesis asserts that participants are driven by either positive or negative social norms when predicting their future behavior and that the prediction exerts a force that brings later behavior in line with it. For behaviors that social norms suggest are admirable, participants are presumed to over-predict their likelihood of behavior performance (due to wishful thinking, or intentional or unintentional social desirability response bias), and then to behave in accordance with those predictions. For behaviors that social norms suggest are embarrassing, participants are expected to under-predict their likelihood of behavior performance and then to behave in accordance with those predictions. A common explanation of the mechanism of the self-prophecy effect is effort to avoid cognitive dissonance between the prediction and the actual behavior.

The mere measurement hypothesis involves a different line of reasoning, built on different assumptions. From this perspective, reporting the likelihood of performing a behavior in the future activates a person's attitude toward the behavior in working memory. Thus the temporary accessibility of the attitude is enhanced, and that increased accessibility is thought to remain high for a considerable amount of time afterwards and to cause the attitude to have stronger impact on later behavior (Morwitz & Fitzsimons, 2004). It is not a general social norm but rather the individual's attitude toward the behavior that determines whether the person manifests increased or decreased behavior frequency. People with favorable attitudes toward the behavior presumably become more likely to perform it, while people with negative attitudes become less likely to perform it. Thus, this approach can be applied to behaviors about which there are no powerful or universally shared social norms.

Williams et al. (2006a) were the first to explore whether a behavior prediction question could alter performance of a negatively sanctioned behavior, in this case illegal drug use. Their study also investigated the effect of behavior prediction on a positively valued behavior: exercising. In an initial survey, half of the participants indicated their likelihood of exercising, and the other half indicated their likelihood of using illegal drugs during the next 2 months. Two months later, all participants reported how many times they had exercised and used illegal drugs.

The authors described their expectations as follows (2006a, p. 120):

> According to self-prophecy, the direction of these effects should be consistent with respondents' understanding of social norms towards the behavior: for health behaviors viewed as socially positive, responding to an intent question should increase the behavior, while the opposite should be true for harmful health behaviors. According to attitude accessibility explanations, the direction of the behavior should be consistent with respondents' own attitudes toward the behaviors, regardless of external social norms. Thus, for behaviors towards which respondents have positive attitudes, responses to an intent question should increase the behavior, while the opposite should be true for behaviors for which respondents hold negative attitudes.

According to Williams et al.'s (2006a) published paper, indicating the likelihood that participants would exercise during the next 2 months apparently significantly increased the frequency with which these individuals did exercise (reported $t=1.64$; reported $p=.05$; $N=167$). This replicated findings of Spangenberg (1997) and can be explained with both the self-prophecy explanation (assuming that exercising is socially desirable behavior) and the attitude accessibility explanation (assuming most of the participants had positive attitudes toward exercising).

Indicating the likelihood that participants would use illegal drugs was also said to have significantly increased drug use frequency (reported $t=2.0$; reported $p<.05$; $N=167$). This is consistent with the self-prophecy explanation if we assume that drug use is socially desirable among college students and is also consistent with the attitude accessibility explanation if the participants mostly had positive attitudes toward illegal drug use.

CORRECTING ACCIDENTAL CALCULATION MISTAKES

The reported statistics in the published paper were not accurate, due to accidental miscalculations of the *t*-values using Excel. In fact, the *t*-values

for the tests the authors intended to report should have been $t=1.26$ ($N=165$) for the effect of predicting exercising and $t=1.71$ ($N=165$) for the effect of predicting illegal drug use (Williams, Block, & Fitzsimons, 2006b).

These are ordinary *t*-statistics assuming equal variances of the dependent variable in the experimental conditions. But in fact, the variance of exercising was significantly greater among people asked to predict their exercise behavior than among people who did not make that prediction (variance=6.64 vs 77.14, $p<.001$, $N=167$), and the variance of drug use was significantly greater among people asked to predict drug use than among people who did not make that prediction (variance=195.71 vs 583.16, $p<.001$, $N=167$). Therefore, *t*-statistics assuming unequal variances should have been computed. In this instance, the *t*-values did not change notably when using this more suitable computational approach: $t=1.25$ for exercise and $t=1.74$ for drug use.

The *p*-values reported in the paper were also incorrect. Although the authors set out to report one-tailed *p*s, they accidentally reported half-tailed *p*s by dividing one-tailed *p* values in half. However, in light of the theoretical perspectives offered in the paper's introduction and quoted above, even one-tailed *p*s seem inappropriate.

In the absence of data on the distributions of attitudes toward exercise or drug use, or on the social norms regarding drug use among these college students, it was impossible to predict with confidence the direction of the effects of the prediction questions, so two-tailed *p*s are most appropriate. When two-tailed *p*s are computed correctly, the difference between the treatment and control groups in their engagement in exercise is not significant ($p=.21$; $N=167$). Thus, this study failed to replicate prior studies indicating that behavior prediction questions could increase the frequency of performance of a socially admirable behavior.

When a two-tailed *p* is computed for drug use, the effect of the behavior prediction question appears to be marginally significant ($p=.086$; $N=167$). However, even this corrected result is misleading, as we will explain next.

PROBLEMS CAUSED BY AN EXTREMELY SKEWED, NON-NORMAL DISTRIBUTION WITH A LARGE MAJORITY OF ZEROS

The problem

The use of *t*-tests is based on a number of assumptions, as are related estimation methods like ordinary least square regressions (OLS; Greene,

2003; Kirk, 1995). Two of these assumptions are especially relevant to the Williams et al. data:

1. Errors of prediction are assumed to be normally distributed.
2. The variance of the prediction errors is assumed to be uncorrelated with the observed values of the dependent variable.

Table 1 displays the distributions of drug use frequency in the full sample and separately for the two experimental conditions. Three aspects of these distributions raise concerns about violation of assumptions:

1. A large number of participants indicated no drug use at all (73.05%).
2. The remainder of the distribution is extremely skewed, with a very long tail and only a few observations at the upper end (skewness=5.51; N=167).
3. Consequently, the distribution is severely non-normal, which almost always means that the first assumption above is violated.

In addition, the unequal variance of drug use in the treatment and control groups violates the second assumption, indicating heteroscedasticity in the disturbances.

Delucci and Bostrom (2004, p. 1159) argued that "to ignore the distribution of the observed data or to blindly use methods based on untenable assumptions about the characteristics of the data is to court

TABLE 1
Distribution of illegal drug use among participants in Williams et al. (2006a)

Uses of illegal drugs	*All participants*		*Treatment*		*Control*	
	Frequency	*Percent*	*Frequency*	*Percent*	*Frequency*	*Percent*
0	122	73.05	62	72.94	60	73.17
1	13	7.78	7	8.24	6	7.32
2	7	4.19	3	3.53	4	4.88
3	3	1.80			3	3.66
4	7	4.19	4	4.71	3	3.66
5	2	1.20	1	1.18	1	1.22
6	1	0.60			1	1.22
7	1	0.60			1	1.22
10	3	1.80	1	1.18	2	2.44
12	2	1.20	2	2.35		
15	1	0.60			1	1.22
20	1	0.60	1	1.18		
25	2	1.20	2	2.35		
50	2	1.20	2	2.35		
N	167	100	85	100	82	100

statistical trouble that may lead to invalid estimates of effects and *p*-values." Furthermore, they said, when dealing with highly non-normal distributions with many zeros, "… we encourage researchers to be wary of using standard parametric methods such as the *t*-test and the ANOVA and suggest using robust or nonparametric/distribution-free methods (p. 1167)." Likewise, McClelland (2000, p. 409) said that "it is a serious abuse to report classical statistical tests … when the assumptions are substantially violated."

In general, *t*-tests are thought to be quite robust as long as (1) the sample size is sufficiently large, (2) the skewness in the two groups is similar, and (3) normality is not too severely violated (Kirk, 1995). But it is difficult to specify criteria for determining what sample size is needed, whether skewness is sufficiently similar, and whether normality is too severely violated (Barber & Thompson, 2000). Furthermore, these three criteria are not independent. For example, the greater the deviation from normality, the larger the sample size must be (Barrett & Goldsmith, 1976).

Many experts have concluded that non-normality in the dependent variable distribution and insufficient sample size make *t*-tests unsuitable for analyzing the Williams et al. (2006a) data. For example, Pocock (1982) concluded that *t*-tests require at least 100 observations in each group to be trustworthy, but he investigated a distribution with substantial fewer than 50% zero values, so an even larger N is no doubt required in the Williams et al. study. Lumley, Diehr, Emerson, and Chen (2002) found that in the presence of extreme skewness, *t*-tests will only be accurate if the total sample size is more than 500. And Sullivan and D'Agostino (1992) concluded that *t*-tests should not be used at all if the proportion of zero values in the dependent variable distribution is larger than 50%. Thus, there is good reason to believe that alternative testing methods should be applied to the Williams et al. (2006a) data on drug use.

This same conclusion is supported by another perspective on these data. The number of incidents of drug use can be thought of as an accumulation of a series of discrete events, which is called a "count" variable (Cohen, Cohen, West, & Aiken, 2003). Results of OLS-based estimation predicting count data are known to be "inefficient, inconsistent, and biased" (Long, 1995, p. 217) and lead to inflated *t*-values for the coefficients (Gardner, Mulvey, & Shaw, 1995). This, too, is reason to consider alternative analytic approaches.

To this end, we reanalyzed the drug use data using a variety of different statistical tests, more suited to the distribution. Most statistical tests we report can be easily implemented using current statistical software. In the Appendix we list the commands used to generate each test statistic using Stata 9.2.

Statistical tests

The first three tests we explored involved OLS regressions with a transformed dependent variable, using a natural-log transformation,[1] a square-root transformation (which is specifically suggested for count variables; see Cohen et al., 2003; Kirk, 1995), and a square root of the square root (fourth root) transformation (Clarke & Green, 1988; Field, Clarke, & Warwick, 1982). Although these transformations are often used to reduce non-normality and heteroscedasticity with skewed distributions, they only produce small improvements with distributions that have many zeros, because those observations are simply moved around in the distribution, leaving the clumping intact and violating the normality assumption of the OLS approach (Gardner et al., 1995).

In the Williams et al. (2006a) data, these three transformations only partially reduced heteroscedasticity, and the variances of the transformed variables were still significantly different in the "treatment" and "control" groups (log-transformation: $p=.002$; square root transformation: $p<.001$; fourth root transformation: $p=.08$). We therefore also conducted a Zhou Z-score test that is designed to replace t-tests with log-transformed variables when the variance of the log-transformed variables differs between the treatment and control groups (Zhou, Gao, & Hui, 1997; Zhou, Melfi, & Hui, 1997).

Next, we implemented an analytic method designed for predicting counts: negative binomial regression, which is often used for predicting rare events with many zeros in the distribution (Long, 1995).[2] This approach is usually an improvement over simple transformations of the dependent variable (Cohen et al., 2003).[3] However, count estimation methods assume that the events counted up are independent of each other (Cohen et al., 2003; Long, 1995). This may not hold for drug use, because use of drugs once may

[1] Because of the presence of zeros in the distribution, the dependent variable must be transformed by adding a constant, as in $\ln(y+c)$, where c is a constant that moves the distribution to the right to eliminate the zero values. Selection of c is arbitrary. In the results reported here, we set $c=1$, as is routinely done (Clarke & Green, 1988; Field et al., 1982; Kirk, 1995). In addition, we also tried setting $c=.0001$ and $c=.05$ and found that the associated p-values were even less statistically significant than the results we report.

[2] Count data are sometimes analyzed via regression based on a Poisson distribution as well, but we restricted our investigation to the more general negative binomial distribution, following recommendations by Gardner et al. (1995) and Long (1995).

[3] The huge proportion of zero drug uses means that an estimator for a zero-inflated distribution might seem appropriate (Long, 1995). Such an estimator assumes that the observed distribution results from two different processes, one that determines whether the dependent variable is zero or non-zero, and a second that determines, among the non-zeros, what the observed value will be. We did not implement this approach, because we saw no basis for assuming that different processes governed use versus non-use of drugs and amount of drug use.

influence subsequent use of drugs by the same person. Therefore, the negative binomial results we report should be treated with caution.

The above tests are all steps in the right direction, but there are reasons to be skeptical about them in this particular context.

The sixth and seventh tests we implemented are non-parametric and have a different sort of potential drawback. The Mann-Whitney-Wilcoxon test and the Kolmogorov-Smirnov test assess whether the distributions of drug use in the two groups are identical to one another, although these tests do not directly and exclusively test whether the arithmetic means of the distributions are different. It is possible that two distributions might be different from one another but have the same mean, leading to rejection of the null hypothesis in a misleading way regarding the mean in particular. But if these tests do not yield significant *p*-values, that is reassuring evidence that the means do not differ. Unfortunately, these tests may be misleading in the presence of a large number of zeros, which can "compromise the power of the Mann-Whitney-Wilcoxon test" (Delucchi & Bostrom, 2004, p. 1163), thus enhancing the likelihood of failing to reject the null hypothesis. Nonetheless, Delucchi and Bostrom (2004) recommended use of these tests for skewed distributions, so we implemented them.

To eliminate *a priori* distributional assumptions from the testing process entirely, we implemented a permutation test and a non-parametric bootstrap test (see Barber & Thompson, 2000; Delucchi & Bostrom, 2004; Efron & Tibshirani, 1993; Ludbrook & Dudley, 1998). The non-parametric bootstrap derives the mean and its variance in each group from a series of bootstrap samples. Each bootstrap sample consists of a random draw of observations from the sample with replacement up to the number of observations in the original sample (i.e., some observations may be sampled numerous times in the same final distribution, while others may not be sampled at all). We constructed 2000 bootstrap samples for the control group and 2000 for the treatment group. Then, the mean of drug use and its standard error were calculated based on all bootstrap samples of the treatment group, and the mean and standard error were calculated separately using all bootstrap samples of the control group. Lastly, the mean and its standard error for each group were used to calculate a regular *Z*-test.

To conduct a permutation test, we started by first randomly assigning each of our observations to either a hypothetical "treatment" group or a hypothetical "control" group and calculating the difference between the groups in the dependent variable's mean for this specific random assignment. In a true permutation test, we would repeat this process for all possible combinations with the given data. Then we could determine the probability that the actual observed difference between the real treatment and control groups in the drug use data occurred by chance alone. This probability is the proportion of permutations that yield a difference equal to or larger than the

difference observed in the real drug use data. For example, if that proportion is smaller than .05, the probability that the observed result occurred by chance alone would be 5% or less. However, permutation tests with more than just a few observations require the calculation of a huge number of possible combinations, making them computationally intensive. We therefore generated a random sample of 2000 of all possible permutations to generate the *p*-values for the permutation test.

Like the Mann-Whitney-Wilcoxon test, the permutation test also focuses on the difference between the distributions of drug use in the two groups, so results can only be used to make confident inferences about a difference in the mean if the test statistic is non-significant (Barber & Thompson, 2000). The non-parametric bootstrap relies on the assumption that the sample distribution is an adequate reflection of the population distribution, and this assumption only holds if a very large sample is used. Considering the small sample size in the Williams et al. (2006a) study, this assumption might be violated, so caution is again advisable.

Finally, we implemented three different versions of Lachenbruch's (1976, 2001, 2002) two-part test. This test was designed for situations with a large number of zeros in the dependent variable distribution. First, a simple χ^2-test was computed for a 2×2 cross-tabulation of a variable indicating whether the participant was in the treatment group or the control group and another variable indicating whether the participant did or did not use drugs. Then a second test was computed to compare the distributions of drug use in the treatment and control groups only among participants who used drugs at least once. This second test can be a *t*-test, and the resulting *t*-value can be transformed into a χ^2 by squaring it. The resulting χ^2 was then added to the χ^2 from the first test, and the result is itself a χ^2 with two degrees of freedom. For the second test, we used a *t*-test, a Mann-Whitney-Wilcoxon-test, and a *t*-test with a log-transformation of the non-zero values.[4]

Results

The first two rows of column 1 in Table 2 show the mean and standard deviation of drug use in the treatment and control groups. The next row shows the two-tailed *p* for an ordinary *t*-test (with unequal variances) that we reported earlier. Two-tailed *p*-values produced by the various remaining estimation methods described above are shown in the remaining rows of the first column of Table 2.

The *Z*-score and the negative binomial regression indicate that the difference between the means was statistically significant ($p<.05$). The OLS

[4] The log-transformed distribution was generated by using ln(y) instead of ln(y+1), because the distribution did not include any zero values, so addition of a constant was not necessary.

TABLE 2
Means, *SD*, and tests of the significance of the effect of behavior prediction questions on drug use

	All participants	*Excluding one outlier*	*Excluding two outliers*
Mean of the control group (*SD*)	1.07 (2.58)	1.07 (2.58)	1.07 (2.58)
Mean of the treatment group (*SD*)	2.80 (8.78)	2.24 (7.13)	1.66 (4.83)
p-value: Ordinary least squares (*t*-test)	.084	.162	.329
p-value: OLS on log-transformed DV	.409	.590	.828
p-value: OLS on square-root-transformed DV	.268	.438	.704
p-value: OLS on fourth-root-transformed DV	.582	.762	.972
p-value: Zhou *Z*-Score test	.028	.110	.369
p-value: Negative binomial regression	.026	.092	.290
p-value: Mann-Whitney-Wilcoxon	.849	.982	.881
p-value: Kolmogorov-Smirnov	.979	.998	1.000
p-value: Permutation test	.087	.185	.342
p-value: Bootstrap test	.082	.154	.319
p-value: Lachenbruch 2-part model with a *t*-test	.140	.236	.374
p-value: Lachenbruch 2-part model with a Mann-Whitney-Wilcoxon test	.673	.802	.897
p-value: Lachenbruch 2-part model with a *t*-test and log transformation	.448	.620	.799
N	167	166	165

t-test, the permutation test, and the bootstrap test indicate that the difference was marginally significant ($p<.10$). The remaining eight tests indicate that the effect of the behavior prediction question was non-significant. This ambiguity in results might be a cause of frustration for an analyst who wants a single answer to this question. Fortunately, a single answer is generated when we take a final step in this adventure by noting a particularly troubling fact about the distribution of drug use, involving outliers.

THE DISTORTING IMPACT OF TWO OUTLIERS

The problem with outliers

The investigation of outliers is an important part in any data analysis (e.g., Cohen et al., 2003; McClelland, 2000). Outliers have particularly pronounced undesirable impact on statistical estimation procedures designed to minimize sums of squares (like *t*-tests and OLS regressions). Indeed, these statistics are "very sensitive to outliers" (McClelland, 2000, p. 393). McClelland (2000) warned against overconfidence in the robustness of statistical analyses without first assessing the possible impact of extreme observations (i.e., outliers).

Quite often investigations into the impact of outliers have been pursued when a researcher produced results that violated his or her expectations, and such an approach has been rightfully criticized. The removal of outliers can be done simply to generate significant results where none is to be found. However, McClelland (2000, p. 409) pointed out that "… the concern of abuse should not be one-sided. It is a serious abuse to report classical statistical tests when the data contain serious outliers." Similarly, Cohen et al. (2003) argued that significant findings that are generated only by the presence of one or two outliers will most likely not hold up upon replication and should therefore not be reported.

In this light, we looked closely at the upper tail of the distribution in Table 1. Two observations are extremely far away from the others, at 50 uses in 60 days, whereas the next most frequent uses are 25 and 20. The two people at 50 are both in the treatment group, which might be viewed as consistent with the conclusion that the behavior prediction question increased drug use. But surprisingly, these two participants indicated during the initial survey that they were extremely unlikely to use drugs during the next 60 days, placing themselves at the lowest point on the likelihood scale. Neither the self-prophecy nor the mere measurement hypothesis predicts that making such extremely low likelihood predictions should have increased drug use. Thus, the scores of 50 for these two individuals seem highly implausible and unlikely to be traceable to the effect of the manipulation.

Following McClelland's (2000) suggestion, we calculated the studentized deleted residual for the top two outliers and found it to be quite high (8.81) and substantially above the suggested threshold of 4. In situations with distant outliers and such implausibility, the conventional approach to take is to re-estimate test statistics removing the outliers (Cohen et al., 2003).

New results

As shown in the second column of Table 2, removing just one of the outliers caused all of the marginally significant statistics to become non-significant and caused the conventionally significant statistics to become non-significant and marginally significant, respectively. Thus only one test, the negative binomial regression, indicated a marginally significant effect of the behavior prediction question. But when we removed the second outlier as well, all test statistics became non-significant (as shown in the third column of Table 2).[5]

[5] Removing these two outliers also caused the non-significant correlation that Williams et al. (2006a) reported between the estimated likelihood of using drugs and the actual frequency of drug use within the treatment group (reported: $r=.034$; $p=.761$; $N=85$) to become stronger and marginally significant ($r=.204$; $p=.065$; $N=83$), showing again how sensitive the presented conclusions are to outliers.

Conclusions

Our investigation identified two problems with the statistical analyses of the drug use data reported by Williams et al. (2006a). First, the very large number of zeros and the general shape of the distribution of drug use violated the assumptions underlying *t*-tests, and this cannot easily be solved simply by transformation. Second, two participants were extreme outliers in the drug use distribution, and implausibly so; removing them caused all tests of the treatment effect to become non-significant. And when accidental calculation errors were corrected, the impact of the behavior prediction question on exercise also became non-significant.

We are therefore inclined to conclude that the data collected in that study do not provide a solid basis for inferring that the behavior prediction questions reliably increased exercise or drug use. The most sensible conclusion to reach is not necessarily that these questions cannot and did not alter behavior patterns, but simply that these data do not afford a justification for concluding that they did. Perhaps with a larger sample these effects would have been more clearly significant. But in the absence of data from additional participants, the conclusions reached in the Williams et al. (2006a) paper must be withdrawn.

Of course, this conclusion sets the stage for future studies of the impact of behavior prediction questions on behavior. In such investigations, conventional parametric statistics such as *t*-tests will sometimes be perfectly fine, when the assumptions underlying these statistics are not violated. But as this literature moves into the domain of socially undesirable behaviors that are performed rarely, it will become increasingly important for researchers to implement the array of tests we reported here. None of these tests is without its potential drawbacks and Achilles' heel, but when viewed as a set, the package may sometimes generate a consistent set of results strongly supporting a single conclusion. In cases when the results are more heterogeneous, investigators will have to investigate carefully which analytic methods are most suitable to the data under study, on both theoretical and empirical grounds.

The lessons learned here do not apply only to studies of the impact of behavior prediction questions. More generally, this experience reminds us that when analyzing the results of experiments (or indeed, doing any statistical social science), we must not just jump into computing test statistics, celebrating significant *p*-values, and considering leaving the profession when *p*s are not significant. Before such analyses are done, we need to pause to inspect the distributions of the variables involved, test the adequacy of the assumptions made by the statistical procedures we implement, check for the presence and distorting impact of outliers, and conduct replication studies to be sure of the robustness of the effects we find.

Of course, this advice is not even slightly new—every introductory scientific methods and statistics course preaches this gospel. But we must all resist the temptation to ignore it in the hopes of generating significant effects as quickly and easily as possible, so we can be sure that our contributions to the published literature are indeed justified by our data.

Manuscript received 12 March 2007
Manuscript accepted 12 April 2007

REFERENCES

Atkinson, R. C., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189–194.

Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: An application of the non-parametric bootstrap. *Statistics in Medicine*, *19*(23), 3219–3236.

Barrett, J. P., & Goldsmith, L. (1976). When is n sufficiently large? *American Statistician*, *30*(2), 67–70.

Clarke, K. R., & Green, R. H. (1988). Statistical design and analysis for a 'biological effects' study. *Marine Ecology – Progress Series*, *46*, 213–226.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry*, *161*(7), 1159–1168.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Field, J. G., Clarke, K. R., & Warwick, R. M. (1982). A practical strategy for analysing multispecies distribution patterns. *Marine Ecology – Progress Series*, *8*, 37–52.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404.

Gayen, A. K. (1950). Significance of difference between the means of two non-normal samples. *Biometrika*, *37(3/4)*, 399–408.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Lachenbruch, P. A. (1976). Analysis of data with clumping at zero. *Biometrische Zeitschrift*, *18*(5), 351–356.

Lachenbruch, P. A. (2001). Comparisons of two-part models with competitors. *Statistics in Medicine*, *20*, 1215–1234.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, *11*(4), 297–302.

Long, J. S. (1995). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, *54*(3), 217–224.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *American Statistician*, *52*(2), 127–132.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Publich Health*, *23*, 151–169.

McClelland, G. H. (2000). Nasty data. Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 393–411). Cambridge, UK: Cambridge University Press.

Morwitz, V. G., & Fitzsimons, G. J. (2004). The mere-measurement effect: Why does measuring intentions change actual behavior? *Journal of Consumer Psychology*, *14*(1&2), 64–74.

Morwitz, V. G., Johnson, E., & Schmittlein, D. (1993). Does measuring intent change behaviour. *Journal of Consumer Research*, *20*(June), 46–51.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, *62*(2), 223–241.

Pocock, S. J. (1982). When not to rely on the central limit theorem – An example from absenteeism data. *Communications in Statistics: Theory and Methods*, *11*(19), 2169–2179.

Ratcliffe, J. F. (1968). The effect on the t distribution of non-normality in the sampled population. *Applied Statistics*, *17*(1), 42–48.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology*, *69*(2), 240–243.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family martial psychotherapy literature. *Clinical Psychology Review*, *9*, 589–603.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*(2), 211–221.

Spangenberg, E. R. (1997). Increasing health club attendance through self-prophecy. *Marketing Lettters*, *8*(1), 23–32.

Spangenberg, E. R., Sprott, D. E., Grohmann, B., & Smith, R. J. (2003). Mass-communicated prediction requests: Practical application and a cognitive dissonance explanation for self-prophecy. *Journal of Marketing*, *67*(2), 47–62.

Sprott, D. E., Spangenberg, E. R., Block, L. G., Fitzsimons, G. J., Morwitz, V. G., & Williams, P. (2006). The question–behavior effect: What we know and where we go from here. *Social Influence*, *1*(2), 129–137.

Sullivan, L. M., & D'Agostino, R. B. (1992). Robustness of the t test applied to data distorted from normality by floor effects. *Journal of Dental Research*, *71*(12), 1938–1943.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Williams, P., Block, L. G., & Fitzsimons, G. J. (2006a). Simply asking questions about health behaviors increases both healthy and unhealthy behaviors. *Social Influence*, *1*(2), 117–127.

Williams, P., Block, L. G., & Fitzsimons, G. J. (2006b). Corrigendum. *Social Influence*, *1*(3), 248.

Zhou, X., Gao, S., & Hui, S. L. (1997). Methods for comparing the means of two independent log-normal samples. *Biometrics*, *53*(3), 1129–1135.

Zhou, X., Melfi, C. A., & Hui, S. L. (1997). Methods for comparison of cost data. *Annals of Internal Medicine*, *127*(8 part 2), 752–756.

APPENDIX: STATISTICAL TESTS IN STATA 9.2

All statistical tests reported in this paper were performed using the statistical software package Stata, Version 9.2, but other statistical software can also be used to conduct the same analyses. In this appendix, we document the commands used in Stata.

OLS regressions with heteroscedasticity-consistent standard errors

To reduce the impact of heteroscedasticity on the standard errors calculated in OLS regressions, heteroscedasticity-consistent estimators can be used (White, 1980). We used the simple command-option "robust", but others are available in Stata, such as HC3 and HC2 (see Long & Ervin, 2000, for further discussion of different techniques to obtain heteroscedasticity-consistent standard errors). Therefore, the following command was used for the first four tests:

```
regress drugs treatment, robust
```

where drugs is the dependent variable and should be replaced by the researcher with the appropriate name of the dependent variable; treatment is the independent variable (the experimental treatment coded either 0 or 1 for control and treatment group) and also needs to be replaced by the researcher with the name of the treatment variable. After the comma, the option robust is added to obtain the corrected standard errors. The commands for the transformed variables are identical, but the dependent variable is transformed by manually implementing a natural log, square-root, or fourth-root transformation of the dependent variable before implementing the regression.

An OLS regression is equivalent to a two-sample *t*-test. The *t*-test also allows the user to adjust the results for unequal variance in the dependent variable between groups:

```
ttest drugs, by(treatment) unequal
```

Z-score test

The *Z*-score test as proposed by Zhou et al. (Zhou, Gao, & Hui, 1997; Zhou, Melfi, & Hui, 1997) is not available in Stata, but its calculus is simple and can easily be implemented. The equation used is:

$$Z = \frac{\hat{\mu}_2 - \hat{\mu}_1 + (1/2)\left(S_2^2 - S_1^2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + (1/2)\left(\frac{S_1^4}{n_1 - 1} + \frac{S_2^4}{n_2 - 1}\right)}}$$

μ_1 = mean of group 1
μ_2 = mean of group 2
S_1 = standard deviation of group 1
S_2 = standard deviation of group 2
n_1 = number of observations in group 1
n_2 = number of observations in group 2

The resulting value of Z is compared to the standard normal distribution. The implementation in Stata requires the computation of the mean and standard deviation, the calculation of the Z-score, and finally a comparison with the normal distribution (lndrugs is the log-transformed dependent variable):[6]

```
tabstat lndrugs, by(treatment) save statistics(mean variance n)
scalar mu_1=el(r(Stat1),1,1)
scalar var_1=el(r(Stat1),2,1)
scalar n_1=el(r(Stat1),3,1)
scalar mu_2=el(r(Stat2),1,1)
scalar var_2=el(r(Stat2),2,1)
scalar n_2=el(r(Stat2),3,1)
scalar Z=(mu_2-mu_1+(.5*(var_2-var_1)))/(sqrt((var_1/n_1)+(var_2/
   n_2)+(.5*(((var_1^2)/(n_1-1))+((var_2^2)/(n_2-1))))))
scalar p=(1-norm(abs(Z)))*2
display "Z-score: " [Z] _newline "p = " [p]
```

Negative binomial regression

Running a negative binomial regression is virtually identical to running an OLS regression with a different command name:

```
nbreg drugs treatment, robust
```

[6] The log-transformation has to be conducted before the test is used. The log-transformation in Stata would be done by using the command generate lndrugs=log(drugs) where drugs is the original variable and lndrugs is the transformed variable.

Non-parametric tests

The Mann-Whitney-Wilcoxon and the Kolmogorov-Smirnov test are implemented similarly to the *t*-test mentioned earlier.

Mann-Whitney-Wilcoxon:

```
ranksum drugs, by(treatment)
```

Kolmogorov-Smirnov:

```
ksmirnov drugs, by(treatment)
```

Permutation and bootstrap tests

To generate permutation and bootstrap tests, we used the tabstat-command to compute means and other statistics for the permutation test. The command used was:

```
permute drugs diff=(el(r(Stat1),1,1)-el(r(Stat2),1,1)), reps(2000): tabstat drugs, by(treatment) save
```

The command for the bootstrap test is similar, but slightly more complicated because the variances and the means are bootstrapped independently for each group, and then a regular *z*-test is performed:

```
bootstrap r(mean) if treatment==0, reps(2000):sum drugs
matrix mu_1=e(b)
matrix sterrsq_1=e(V)
bootstrap r(mean) if treatment==1, nodots reps(2000):sum drugs
matrix mu_2=e(b)


matrix sterrsq_2=e(V)
scalar Z=((mu_1[1,1]- mu_2[1,1])/sqrt(sterrsq_1[1,1]+ sterrsq_2[1,1]))


scalar pp=(1-norm(abs(z)))*2
display "z-value: " [Z] _newline "p = " [p]
```

Two-part tests

To implement the two-part tests, we first generated a variable coded 0 for participants who did not use drugs and 1 for those who did (called

useddrugs in the following commands). Then we computed a χ^2 for the relation between group membership (0=control group, 1=treatment group) and useddrugs:

```
tab2 useddrugs treatment,chi
scalar firstpart=r(chi2)
```

Then a second χ^2 was computed using only participants who reported non-zero drug use by including an if-statement in the second command. Here we show the command for the simple *t*-test, for the Mann-Whitney-Wilcoxon and the *t*-test on the log-transformed variable, as discussed before:

```
ttest drugs if drugs > 0, by(treatment)
```

or

```
ranksum drugs if drugs > 0, by(treatment)
```

or

```
ttest lndrugs if drugs > 0, by(treatment)
```

The resulting t-value is squared afterwards:

```
scalar secondpart=r(t)^2
```

Then these two χ^2s are added together:

```
scalar chisq= firstpart+secondpart
```

Then the two χ^2-values are compared to a χ^2-distribution with 2 degrees of freedom to obtain a *p*-value:

```
scalar p=chi2tail(2, chisq)
display "Chi^2: " [chisq] _newline "p = " [pp]
```