Psychological Bulletin
1993, Vol. 113, No. 3, 566-579

Transformations for Within-Subject Designs: A Monte Carlo Investigation

Lauren K. Bush, Ursula Hess, and George Wolford

We explored the use of transformations to improve power in within-subject designs in which multiple observations are collected for each S in each condition, such as reaction time and psychophysiological experiments. Often, the multiple measures within a treatment are simply averaged to yield a single number, but other transformations have been proposed. Monte Carlo simulations were used to investigate the influence of those transformations on the probabilities of Type I and Type II errors. With normally distributed data, *Z* and range correction transformations led to substantial increases in power over simple averages. With highly skewed distributions, the optimal transformation depended on several variables, but *Z* and range correction performed well across conditions. Correction for outliers was useful in increasing power, and trimming was more effective than eliminating all points beyond a criterion.

Psychologists are fond of transforming their data, occasionally with good reason. The most common reason for transforming data is either that the transformed measure more closely reflects the psychological process of interest (Levey, 1980) or that the data do not meet the assumptions for the desired analysis (Winer, 1971). In the latter case, two assumptions that are often violated are the homogeneity of variance across cells and the degree to which the data are approximated by a normal distribution. These two problems are related, and a transformation that addresses one often improves the other. The appropriate transformation depends on the shape of the underlying distribution. Various transformations have been proposed for specific cases, including log, reciprocal, square root, and arcsine transformations. Kruskal (1968) and Smith (1976) have provided excellent overviews of these transformations and when to use them.

A second, and related, reason for transforming data is to lessen the impact of outliers. Two common techniques are trimming and eliminating all points beyond a certain criterion. Wilcox (1992) has provided a nice overview of trimming as a technique for dealing with outliers.

A third reason for transforming data is to eliminate differences in reactivity or variability across subjects. These transformations apply only to designs in which each subject receives all levels of the independent variable and in which each subject yields multiple observations at every level. The three types of transformations are interrelated in their effects. For instance, eliminating outliers tends to normalize the distributions and equate variances. In the present investigation, we concentrated on the third type of transformation (equating reactivity), but in several situations we used these transformations in concert with the other types of transformations.

Two examples illustrate the paradigms in which the reactivity transformations apply (both are somewhat simplified from their usual implementation). In the first example, subjects view two different videotapes, one depicting a humorous situation and one depicting a neutral situation. During the showing of each videotape, heart rate is recorded every few seconds. In the second example, subjects participate in the cuing paradigm developed by Posner (1978). On each trial, subjects try to detect a target as rapidly as possible. On some trials, the probable location of the target is cued in advance; on other trials, it is not cued. Latencies are recorded on numerous trials of each of the two conditions, cued versus uncued.

In both paradigms, each subject contributes, several observations to each of the two conditions. Prior to analyzing the data, most researchers collapse those several numbers into a single number for each condition. The most common procedure is to use the mean of all of the numbers within a condition as the score for that condition. In the present article, we examined several alternatives to the simple mean to determine whether any are superior.

Most of the transformations performed to eliminate differences in reactivity scale the effect of an independent variable for a given subject in terms of that subject's variability. The logic behind such transformations is that subjects differ substantially in variance or reactivity and that a change of *n* units from Condition 1 to Condition 2 might imply more or less of an effect for different subjects. Such transformations would be successful if the transformed data more accurately reflected the influence of the independent variable. Our measure of success was the statistical power of the various transformations as determined in Monte Carlo simulations.

Lauren K. Bush (now at Department of Psychology and Counseling, Dartmouth College); Ursula Hess (now at Department of Psychology, University of Quebec, Montreal, Quebec, Canada); and George Wolford, Department of Psychology, Dartmouth College.

This work was supported in part by National Institute of Mental Health Grant 5-T32-MH18931 to Lauren K. Bush and by National Science Foundation Grant BNS-8820653 to George Wolford. We would like to thank Arvid Kappas for his help during all phases of the project; James Townsend for his comments on foundational measurement and meaningfulness; and Gershon Ben-Shakhar, Richard Gonzales, David Lykken, Victor McGee, Gregory McHugo, and Glen Shafer for their extremely helpful comments.

Correspondence concerning this article should be addressed to George Wolford, Department of Psychology, Dartmouth College, Hanover, New Hampshire 03755.

Before describing the transformations we examined, we want to emphasize that there are situations in which it would be inappropriate to use them (Townsend & Ashby, 1984). For example, if the experimenter has reason to believe that the dependent measure achieves a ratio-level scale of measurement and that the ratios are of theoretical interest, then the only transformation that would be permissible is a multiplicative transformation, because that is the only one that preserves meaningful ratios (Townsend, 1992). The transformations that we discuss would be inappropriate because they would destroy the meaningfulness of the ratios. If, however, the experimenter was interested in whether two treatments were significantly different and not in the exact ratios, then our transformations would be appropriate. A readable treatment of measurement theory has been provided by Roberts (1979). We return to the issue of meaningfulness in the Discussion section.

There are other reasons to be wary of transformations besides loss of meaningfulness. It is essential to remember that when data are transformed, some information is removed, and a different question is addressed than was addressed with the raw data.[1] A transformation might make it more difficult to communicate the results if it were contrary to normal convention. It might also make it difficult to compare the data with previous data (Cacioppo, Tassinary, & Fridlund, 1990). In the present analysis, we judged the success of our transformations solely on the basis of statistical power. If a transformation improved statistical power but was contrary to normal convention, it would be easy to report both the transformed and the raw data.

The two paradigms illustrated earlier—psychophysiology and information processing—diverge in their concern over appropriate transformations. We discuss each paradigm in turn.

In psychophysiology, the recommended procedure varies somewhat with the dependent measure. The transformations that have been used with skin conductance include range-corrected scores, ratio-corrected scores, within-subject standardization, and log and square root transformations. Range-correction transformations (Lykken, 1972; Lykken, Rose, Luther, & Maley, 1966) express an individual's response to a particular stimulus as a proportion of his or her range of responsivity and presumably reduce variance due to individual differences in basal level. Ratio-correction transformations (Paintal, 1951) express a particular skin conductance measure as the ratio of that response to the subject's maximal response. The within-subject Z transformation expresses each individual's responses relative to his or her mean and standard deviation (e.g., Ben-Shakhar, 1985, 1987). Log and square root transformations have been proposed specifically for the purpose of normalizing the sample distribution (Venables & Christie, 1980).

The decision to represent data in a particular way depends not only on the sensitivity of that representation to treatment effects, but also on the specific process that one is trying to study (for detailed discussions of the processes that underlie skin conductance responses, see Edelberg, 1972; Fowles, 1986). When the range correction and within-subject Z transformation methods are used, information about individual differences is lost, and for this reason some researchers do not favor their use (Stemmler, 1987a). Researchers who favor the use of such transformations often justify the loss of basal-level information by arguing that basal levels usually reflect individual difference variance unrelated to the psychological process of interest and that transformation enhances their analysis's sensitivity to "real" treatment effects (e.g., within-subject Z transformation, Ben-Shakhar, 1985; range correction, Lykken & Venables, 1971).

Many researchers favor transformations that enhance sensitivity to "real" effects (Levey, 1980), but not all agree that these transformations enhance power (cf. Ben-Shakhar, 1985, 1987; Stemmler, 1987a, 1987b). Recently, Stemmler (1987a) suggested that within-subject standardization could actually reduce an analysis's sensitivity to "real" treatment effects by disproportionately reducing the contributions of the more reactive subjects. He used an artificially constructed data set to illustrate his concern. Our goal in the present investigation was to examine systematically the effect of these transformations on statistical power.

Researchers who use reaction time paradigms (e.g., cognitive psychologists) have worried less than psychophysiologists about techniques for combining the multiple measures within a condition into a single number. The vast majority of such researchers take the mean of all the reaction times within a cell for a subject and carry out subsequent analyses on those means. On rare occasions, the reaction times are subjected to a log or square root transformation, or medians are used rather than means. Researchers who use reaction time paradigms do worry about outliers, however, and often perform some procedure to eliminate deviant reaction times. The most common procedure is to eliminate all data points more than k standard deviations from the mean of a subject's condition. We examined the effect of outliers and the procedures for dealing with them in conjunction with the various transformations.

Previous researchers (e.g., Ben-Shakar, 1985) have examined the power of various transformations by taking an empirically obtained data set, carrying out several different transformations, and determining which transformation yielded the highest F value for that data set. This method has advantages and disadvantages. The advantage is that the data are guaranteed to be realistic (the underlying distributions are correct, the values are appropriate, etc.). The disadvantage is that because the true outcome is necessarily unknown, the transformation that maximizes the F value may not be the best one; it may be yielding a Type I error. In addition, it is difficult to draw conclusions from a small number of data sets.

We examined the power of the various within-subject transformations by using Monte Carlo simulations. We generated hundreds of thousands of data sets with and without real effects, employed each of the recommended transformations, and recorded the probabilities of Type I and Type II errors for each transformation. We varied the shape of the underlying distribution, the way in which effects were added, the number of subjects, the number of observations per subject, the effect size, and the presence or absence of outliers.

[1] It should be pointed out that any procedure for reducing the multiple observations to a single number per condition, including a simple average, is a transformation and removes some of the information.

Overview of the Approach

A description of the general flow of the Monte Carlo program will allow a better appreciation of the procedures and results. For purposes of description, we used heart rate as the dependent measure. For a given set of parameters, we carried out either 1,000 or 10,000 experiments. We used a two-stage sampling procedure. In the first stage, we chose n subjects and the sampling parameters for those subjects. We assumed that each subject had his or her own heart rate distribution and that those individual distributions differed in both mean and variance. For each subject, then, we chose that subject's mean heart rate from a normal distribution and his or her standard deviation of heart rate from a chi-square distribution.[2] In the second sampling stage, we chose m observations for each of two conditions from a normal distribution using *that* subject's mean and standard deviation. Next, we added an effect to each of the observations in the second condition. This process was repeated for all n subjects. We reduced the m observations for each subject in each condition to a single number for that condition using each of the transformations. The resulting pairs of numbers were analyzed with t tests, and the number of significant results was recorded. We measured Type I errors by setting the effect size to zero.

We generated data using both normal and positively skewed distributions, because both have been reported in the psychophysiological and reaction time literature (Venables & Christie, 1980). Specific parameters for the distributions were chosen to approximate the distributions that have been reported for skin conductance and heart rate (Graham, 1980; Siddle & Turpin, 1980; Venables & Christie, 1980) and found in data from our laboratory. We compared five different transformations in the initial simulations: means of the raw scores, means of the logs of the raw scores, Z scores, range-corrected scores, and ratio scores for repeated measures data. All of these transformations have been recommended for one situation or another. In later simulations, we added medians and the Z transformation of log scores to the set of transformations, and we examined the influence of outliers.

We had some concern about the procedures for choosing effects. Most Monte Carlo simulations set the effect size at some fraction of the standard deviation. In our case, there were a couple of different possibilities for the standard deviation. One possibility was the mean of the chi-square distribution from which each subject's standard deviation was chosen, and the other was the particular subject's standard deviation. In the former case, every subject would have the same absolute effect size; in the latter case, every subject would have the same relative effect size. It also seemed possible that effects are not constant but are random variables from a distribution. Because all of these possibilities seemed reasonable, we carried out the initial simulations adding effects in three different ways.

Simulation 1: Normally Distributed Data

Method

Design

Simulation 1 consisted of 1,000 experiments per cell. We varied the effect size (0, ¼, ½, or ¾ of the standard deviation), the number of subjects (10 vs. 20), the number of observations per subject (10 vs. 20), and the procedure for generating effects. The individual responses were drawn from normal distributions.

Computations

The simulations were carried out on two different computers: a Macintosh II and a Convex super minicomputer. The simulations were programmed in BASIC on the Macintosh and in FORTRAN on the Convex. Both languages and machines make use of high-precision numbers (10-byte representation on the Macintosh and 8-byte representation on the Convex). With identical parameters, the two machines yielded highly similar results; the advantage of the Convex was a manifold increase in speed.

Random numbers were generated using the built-in random-number generators with random seeds at the beginning of each run. We then converted the random numbers to normal deviates using an algorithm developed by Box and Muller (1958).

Raw Data Generation

For the first set of simulations, we used data from normal distributions. The parameters for the normal distributions were chosen to reflect the type of distribution obtained for heart rate data (Siddle & Turpin, 1980). Likewise, heart rate data from our laboratory have been approximately normal in all studies (although a slight positive skew was occasionally evident). The following parameters were derived from several studies in our laboratory using college-age students. Each subject's mean heart rate was chosen from a normal distribution with a mean of 69 beats per minute (bpm) and a standard deviation of 12 bpm; each subject's standard deviation of heart rate was chosen from a chi-square distribution with 4 degrees of freedom. (A chi-square distribution with 4 degrees of freedom has a mean of 4 and a standard deviation of 2.83.)

We chose the mean and standard deviation for each of the subjects and then, using those parameters, chose the given number of observations from normal distributions. In the first simulation, the independent variables were number of subjects (10 vs. 20), number of observations per condition (10 vs. 20), and the size and nature of the effect size. For each subject in each cell, there were two sets of numbers: one intended to serve as the control or neutral condition and the other as the experimental condition.

We added an effect to the scores in the experimental condition in one of three ways. In the absolute case, we added either 0, ¼, ½, or ¾ of the mean of the chi-square distribution. In the relative case, we added 0, ¼, ½, or ¾ of that subject's standard deviation (which had been chosen from the chi-square distribution). In the distributional case, we added a number from a normal distribution that had a mean equal to 0, ¼, ½, or ¾ of the mean of the chi-square distribution. In the absolute case, the same number was added to every score in the experimental condition for every subject, but the number represented a different fraction of each subject's standard deviation. In the relative case, different numbers were added to different subjects' scores in the experimental condition, but those numbers represented the same fraction of each subject's standard deviation. The distributional case was similar to the absolute case, except that the effects were not a constant but were drawn from a distribution with the same mean as the constant. The

[2] Differences in mean heart rate did not affect any of our analyses, because all of the statistics that we carried out were within-subject analyses. We varied mean heart rate so that the numbers we generated appeared reasonable. This was useful in debugging the simulations.

expected value of the effects for a particular size effect (e.g., ¼ of the standard deviation) was the same in all three cases.

Transformations

Once the data for each of the subjects were generated, we applied each of the transformations to those data. Each transformation resulted in a single number for each condition. Those 10 or 20 pairs of numbers were then analyzed with a *t* test for correlated scores, and the number of significant outcomes using an apha of .05 was recorded.

Means of raw scores. This transformation was accomplished by computing the mean of all the observations in each cell for each subject. Each subject was left with two scores: the mean of the observations in the neutral condition and the mean of the observations in the experimental condition.

Means of log scores. This transformation was accomplished by computing the mean of the logs of the observations in each condition for each subject. Log transformations have been recommended to reduce the impact of deviant scores, particularly if the underlying distribution is likely to be skewed.

Z scores. This transformation was accomplished by first finding the mean and standard deviation of all of the scores for a given subject across the two conditions. The raw scores were then transformed to *Z* scores by subtracting the mean and dividing by the subject's standard deviation. The mean of the *Z* scores within each condition was then computed to represent the single number for each condition.[3]

Range-correction scores. This transformation involved finding the minimum score in either condition and then subtracting that score from each of the subject's responses; this difference was then divided by the range of scores across the two conditions (maximum response minus minimum response).[4]

Ratio scores. This transformation involved finding the maximum score for each subject across the two conditions and then dividing all the scores by that maximum. Finally, the mean of those ratios was computed for each condition.

Simplified calculation. Using a *t* test for correlated measures, the last three transformations are equivalent to computing the mean difference between conditions for a particular subject and then dividing the mean difference by that subject's standard deviation, range, or maximum.

Analysis

The result of each of the transformations was a pair of scores for each subject. We analyzed those pairs (10 or 20, depending on the number of subjects) using a *t* test for correlated observations (Winer, 1971). The result was considered significant if the absolute value of the obtained *t* value exceeded the critical value (i.e., a two-tailed decision rule) with a *p* of .05.

Results and Discussion

The results of Simulation 1 are presented in Tables 1–3. The numbers in the tables refer to the percentages of significant *t* values from the 1,000 experiments in each cell. Table 1 contains the results for absolute effect size, Table 2 the results for relative effect size, and Table 3 the results for effects chosen from a distribution. The pattern of results is clear. In all the conditions, the *Z* transformation and the range-corrected transformation were superior in statistical power to the other transformations, often by 20% or more. Although the *Z* and range-corrected transformations led to similar results, the *Z* transformation was slightly superior when there was a difference. The log transformation was ineffective, as expected with normally distributed data.

None of the transformations led to spurious Type 1 errors. Notice, however, that the Type 1 errors were always less than .05 for the means of raw scores. The standard error of that cell was about 0.7% when the binomial distribution was used. In several cases, then, the probability of a Type 1 error was significantly below .05 for the means of raw scores, and certainly in composite there were too few Type 1 errors with the simple means. This implies that the *t* test was overly conservative for the untransformed data in the paradigm that we used. The probabilities of Type 1 errors occurred in approximately correct frequencies for the most effective transformations, that is, *Z* and range correction.

We believe that within-subject transformations improve power by scaling effects relative to a subject's own variability or reactivity (see Simulation 2). The slight superiority of the *Z* transformation over the range-corrected transformation was probably due to the fact that the standard deviation is a more stable estimate of dispersion than is the range. Estimates of the minimum and the maximum are influenced strongly by outliers. Range-corrected scores, however, may have a more natural interpretation in some cases because the transformed cell means represent a particular fraction of the available range. The ratio transformation improved power over the simple average but did not fare as well as the *Z* and range transformations.

Reversals of Effect Direction

Several investigators have expressed concern that transformations of the type we examined might distort the pattern of effects. For example, Cacioppo et al. (1990) illustrated how a different pattern may be obtained with raw means than with *Z* means. Such reversals occur when a small number of subjects have both high variance and effects in the opposite direction from the other subjects. Because of this possibility, Cacioppo et al. recommended caution in using transformations and making comparisons across studies in which different transformations were used.

All the transformations that we considered are monotonic within a single subject. That is, if the mean of Condition 2 is larger than the mean of Condition 1 for a given subject, all the transformations preserve that ordering. However, the transformations are not necessarily monotonic with respect to group means. They change the relative weighting of a subject in the group and can change the ordering of the group means. Is this

[3] One reviewer suggested that it might be more appropriate to compute the standard deviation within each condition and then pool those estimates rather than compute the standard deviation across all scores. We tried both methods in Simulation 1, but they led to similar results. Whenever there was a difference, it was in favor of computing the standard deviation across conditions, which is the method that we report.

[4] In developing the range correction transformation, Lykken (1972) measured the range for a subject in a separate session and then used that range to transform the scores from the experimental sessions. For purposes of symmetry and convenience, we did not use a separate set of scores to estimate the minimum and maximum.

Table 1
Percentage of Significant Results With Normal Distribution and Constant Effect

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
	10 subjects per sample							
Mean raw	34.2	78.5	93.6	3.0	53.2	93.3	99.0	3.8
Z score	56.2	92.4	99.3	4.9	75.0	98.9	100	5.0
Ratio	36.7	81.0	95.1	3.4	56.8	94.6	99.4	4.6
Range	54.8	91.2	98.4	5.0	72.2	97.5	100	5.0
Mean log	33.1	77.3	92.8	2.8	52.7	92.5	98.7	4.0
	20 subjects per sample							
Mean raw	57.4	95.2	99.5	4.0	79.0	99.6	100	4.2
Z score	89.6	100	100	4.9	98.2	100	100	5.8
Ratio	60.6	96.9	99.6	3.5	82.3	99.9	100	4.3
Range	88.9	100	100	5.0	97.9	100	100	5.4
Mean log	55.3	94.0	98.8	3.1	76.7	99.2	100	4.1

really a problem? We addressed this question in two ways. First, in all our simulations, we recorded the direction of every significant result. In all the simulations reported herein (approximately 500,000 experiments), every significant finding was in the correct direction (i.e., the mean of the experimental group was greater than the mean of the control group.)[5]

Second, we further examined the direction of mean differences for the data in the first column of Table 1 (i.e., 10 subjects, 10 observations, ¼ of the standard deviation, and absolute effect). The data represent 1,000 experiments; the results of 342 were significant using raw means, and the results of 562 were significant using the *Z* transformation. We examined the direction of the effect for those experiments without regard to significance. The expanded results for the raw and *Z* means appear in Table 4. In the vast majority of cases (927 of 1,000), the means had the correct ordering (i.e., the experimental mean was greater than the control mean) for both procedures. In 9 other cases, both transformations yielded an incorrect ordering. There were, however, 64 cases in which the raw means yielded an ordering different from that yielded by the *Z* means. In 63 of those cases, the *Z* means were in the correct direction and the raw means were in the incorrect direction. In short, the transformations that we considered occasionally reversed the pattern of effects seen with raw means, but the reversal was almost always advantageous in this situation where we knew the correct answer. Furthermore, such reversals never occurred with significant effects.

Numbers of Subjects and Observations

We varied the number of subjects and the number of observations in Simulation 1. Increasing either number increased power as expected, with a slight advantage for number of subjects. Although we found the same ordering of transformation effectiveness when numbers of subjects and observations were varied, a colleague expressed concern that the usefulness of transformations might diminish if we used large numbers of subjects or observations. In many reaction time experiments, there are hundreds of observations per condition per subject. To test this possibility, we carried out one simulation with 10 subjects and 100 observations and a second simulation with 100 subjects and 10 observations. With these larger numbers, we used an effect size of 5/100 of the standard deviation to prevent ceiling effects. In both cases, we found the same ordering of transformations as shown in Tables 1-3. For example, with 10 subjects and 100 observations, 16.8% of the results were significant with raw means, and 31.0% of the results were significant with *Z* means. With 100 subjects and 10 observations, 13.5% of the results were significant with raw means, and 44.9% of the results were significant with *Z* means. The effectiveness of using the *Z* transformation, therefore, did not appear to diminish with large numbers of subjects or observations provided that ceiling effects were avoided.

Procedure for Choosing Effects

We used three procedures for adding effects to the experimental condition: absolute, relative, and distributional. The rank ordering of the effectiveness of the different transformations was identical in all three cases. The numerical results from the absolute case and the distributional case were nearly identical, although that similarity would probably diminish if we increased the variance of the effect distribution for the distributional case. The relative case yielded lower power for small effect sizes but higher power for larger effect sizes compared with the other cases. The relative and distributional cases were probably more realistic than the absolute case. Because the procedure for adding effects never influenced the ordering of the effectiveness of the transformations and because most paramet-

[5] This was not true for the null case, in which no effect was added to either group. In this situation, there was no "correct" direction, and the significant results split approximately equally in the two directions.

Table 2
Percentage of Significant Results With Normal Distribution and Relative Effect

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
Mean raw	22.4	71.6	94.3	3.7	42.1	90.8	98.5	4.3
Z score	35.0	88.4	99.7	5.4	59.2	99.2	100	5.5
Ratio	23.4	74.3	95.7	3.9	43.7	93.5	99.7	4.7
Range	34.6	87.7	99.7	5.4	59.1	99.1	100	4.8
Mean log	22.1	69.3	92.0	3.8	40.0	89.7	98.4	4.5
	20 subjects per sample							
Mean raw	48.7	96.2	99.9	4.6	79.8	100	100	4.2
Z score	64.0	99.6	100	4.5	92.7	100	100	5.8
Ratio	51.2	97.0	99.9	5.1	81.8	100	100	4.3
Range	63.7	99.2	100	4.4	91.6	100	100	5.4
Mean log	47.3	95.9	99.9	5.3	77.7	99.8	100	4.1

ric tests assume the absolute case, we used it in all subsequent simulations.

Simulation 2: Equal-Variance Case

Method

We hypothesized that the success of the *Z* and range-corrected transformations resulted from scaling a subject's effect size in terms of that subject's dispersion. As a check on this hypothesis, we repeated one quadrant of Table 1 (i.e., 10 subjects and 10 observations), assigning every subject the same standard deviation. All of the details were the same as in Simulation 1, except that instead of choosing each subject's standard deviation from a chi-square distribution, we assigned the mean of that chi-square distribution to every subject. The individual subject distributions thus differed in mean but not in variability. If the *Z* and range-corrected transformations succeed by scaling effects in terms of an individual's variability, then they should have little effect in Simulation 2 because all subjects had the same expected variance. The results of Simulation 2 appear in Table 5. As is shown, none of the transformations was superior to taking a simple average when all subjects had the same variability. This supports the conclusion that the *Z* and range-corrected transformations achieved their additional power in Simulation 1 by scaling effects in terms of each subject's variability.

Do real subjects differ in variability? Certainly. We based the parameters in Simulation 1 on the data of subjects from our previous research. Those individual subjects differed substantially in the standard deviation of their heart rate distributions, and our impression is that the differences are also pronounced for reaction time and skin conductance. For example, in Hess, Kappas, McHugo, Lanzetta, and Kleck's (1992) study, each subject's heart rate was measured 20 times, and the standard deviation of heart rate within a subject ranged from a low of 1.76 bpm to a high of 6.66 bpm. Typical effect sizes were on the order of 3 bpm. In a reaction time study involving speech recognition,

Table 3
Percentage of Significant Results With Normal Distribution and Distributional Effect

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
	10 subjects per sample							
Mean raw	35.3	75.1	92.2	3.9	55.0	91.1	98.5	3.0
Z score	55.0	89.4	97.8	6.5	75.5	98.4	99.8	4.9
Ratio	35.9	76.2	92.8	4.1	56.2	91.7	98.9	3.2
Range	54.8	89.8	97.8	6.3	75.8	97.8	99.8	4.6
Mean log	34.9	75.5	91.8	4.1	54.9	91.0	98.3	3.1
	20 subjects per sample							
Mean raw	56.2	94.0	99.3	5.2	80.2	98.9	100	5.2
Z score	87.1	99.8	100	5.5	97.6	100	100	5.4
Ratio	57.7	94.7	99.3	5.3	82.2	99.3	100	5.3
Range	86.9	99.8	100	5.8	97.3	100	100	5.1
Mean log	55.9	93.8	99.3	5.1	80.4	98.8	100	5.1

Table 4
Reversals in the Pattern of Effects With Transformations

	Z transformation	
Raw means	Experimental > control	Control > experimental
Experimental > control	927	1
Control > experimental	63	9

Fowler, Treiman, and Gross (in press) had subjects contribute hundreds of reaction times. The standard deviation of the reaction times within a subject ranged from a low of 109 ms to a high of 729 ms. Typical effect sizes were on the order of 125 ms.

Simulation 3: Skewed Distribution

Method

In Simulation 3, we examined the efficacy of the various transformations with a skewed distribution. Dependent variables such as reaction time and skin conductance often yield skewed distributions. For example, several investigators have suggested that reaction times are well fit by a gamma distribution. The gamma distribution is equivalent to the sum of *r* exponentials, each having the intensity parameter lambda (λ). Researchers have often modeled reaction times by assuming that some number of synapses intervene between the stimulus and the response and that the transmission time of any one synapse is distributed as an exponential. The chi-square distribution is a special case of the gamma distribution in which $r = \nu/2$ (where ν represents the degrees of freedom for the chi-square distribution) and $\lambda = 0.5$. We used the chi-square distribution in our simulations because it is slightly easier to work with than the generalized gamma distribution. A chi-square distribution can be generated as the sum of ν squared normal distributions.

Unlike the normal distribution, the mean, variance, and skew of a chi-square distribution are interrelated (all three are functions of ν). For that reason, it is not possible to have those parameters vary independently. Their interdependence is shown empirically in most reaction time experiments by a positive correlation between the mean and variance.

In Simulation 3, then, we began each experiment by randomly choosing a ν for each of the *n* subjects from a uniform distribution ranging from 3 to 14. We chose *m* observations for each condition from a chi-square distribution with that subject's value of ν. As in Simulation 1, we varied the number of subjects (10 vs. 20) and the number of observations per condition (10 vs. 20). We added effect sizes of 0, ¼, ½, or ¾ of the average standard deviation to each of the observations in the experimental condition. With ν ranging from 3 to 14, the expected value of standard deviation across subjects was 4.03.

Results

The results of Simulation 3 are presented in Table 6. For the smallest effect size, the overall power was similar to that shown for normally distributed data (see Table 1), but with larger effect sizes more significant differences were found with the skewed distributions. The log transformation led to a benefit over the raw scores and was superior in every case to the other transformations as applied to the raw scores. The advantage of the log transformation is consistent with the recommendation found in most statistics texts for dealing with skewed distributions. The *Z*, range, and ratio transformations were modestly superior to the mean of the raw scores. On the basis of the initial results of Simulation 3, we tried several additional transformations. Namely, we applied the *Z*, range, and ratio transformations to the logs of the individual scores. Those transformations generally yielded a small but consistent improvement over the use of logs alone. We found in Simulation 5 that these results with skewed distributions were affected substantially by outliers.

Simulation 4: Outliers, Corrections, and Medians With Normal Distribution

Method

Both reaction time and psychophysiological measures occasionally yield scores that appear to be outliers. These deviant data points result from numerous sources: faulty equipment, shifting electrodes, inattentive subjects, and so forth. Investigators often use some technique to eliminate these outliers. We considered three common techniques: medians, trimming, and the elimination of all points beyond some criterion. Medians are sensitive only to the ordering of the data and are therefore less affected than means by extreme outliers. Medians, however, are insufficient estimators because they do not use all of the data, namely, the magnitude of the numbers. Miller (1988) has also criticized medians as being biased, with the magnitude of the bias being a function of the number of subjects.

We considered two other techniques for dealing with outliers; both consist of eliminating some of the data points. Winer (1971) described the use of trimming for dealing with outliers (see also Wilcox, 1992). In our version of trimming, we eliminated the highest and lowest score for each subject in each condition. The final, and perhaps most popular, technique is to eliminate all data points beyond some criterion. In an informal poll of several colleagues, the most common criterion was ±3 standard deviations from a subject's condition mean. In the same informal poll, we asked colleagues what percentage of their data were eliminated with this procedure. The modal response was "around 2%."

In Simulation 4, then, we generated the data as in the upper right quadrant of Table 1 (i.e., 10 subjects, 20 observations, and normally distributed data). There were two new variables in this simulation, and their orthogonal combination yielded six cells. The first variable was the presence or absence of outliers, and the second variable was the procedure for dealing with outliers: doing nothing, trimming, or eliminating scores exceeding 3 standard deviations. In the conditions with outliers, we chose 4% of the data points at random and added ±5 standard deviations to those data points. We chose the frequency and the

Table 5
Percentage of Significant Results With Normal Distribution, Constant Effect, and Equal Variance Among Subjects

	Effect size			
Transformation	1/4	1/2	3/4	Null
Mean raw	36.9	87.9	99.6	5.1
Z score	36.8	88.0	99.7	4.7
Ratio	36.1	87.2	99.6	4.1
Range	36.6	87.7	99.7	5.0
Mean log	35.8	86.2	99.5	4.3

Note. Percentages are based on 10 observations per subject and 10 subjects per sample.

Table 6
Percentage of Significant Results With Skewed Distribution

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
			10 subjects per sample					
Mean raw	35.9	87.1	98.8	4.6	61.7	98.8	100	5.0
Z score	40.4	89.9	99.6	5.2	65.9	99.4	100	5.2
Ratio	40.3	89.3	99.7	5.2	66.0	99.1	100	5.1
Range	39.6	88.7	99.5	5.3	65.7	99.2	100	5.1
Mean log	44.9	93.2	99.9	5.4	72.7	99.7	100	5.1
Z score log	45.3	93.3	99.9	5.7	73.9	99.7	100	5.3
			20 subjects per sample					
Mean raw	64.1	99.3	100	5.7	92.0	100	100	4.3
Z score	70.8	99.7	100	5.1	94.0	100	100	5.1
Ratio	70.2	99.7	100	5.8	93.7	100	100	4.9
Range	70.9	99.7	100	5.7	93.5	100	100	4.9
Mean log	76.2	99.8	100	6.0	96.2	100	100	4.7
Z score log	77.8	99.7	100	5.4	96.0	100	100	5.2

magnitude of the outliers to yield approximately 2% outliers as revealed by the 3-standard-deviation (3-*SD*) elimination procedure. Because outliers affect the mean and standard deviation of the sample, the 3-*SD* procedure does not identify all outliers. We also added medians to the list of transformations.

Results

The results of Simulation 4 appear in Table 7. We expected outliers to reduce the power of the various tests because they add to the error variance. In this simulation with normally distributed data, outliers did reduce the power of all the tests, in some cases by as much as 10%–15%. Note that with outliers, the probability of a Type 1 error is less than .05 in most cases. The ordering of the effectiveness of the different transformations was the same as in the previous simulations with normally distributed data. The *Z* transformation remained optimal, yielding the highest number of significant results.

Medians

Medians were affected minimally by outliers, as expected, but were quite low in power compared with the other transformations (with or without the presence of outliers). Given the lack of power and the bias problems cited earlier, it is hard to imagine a justification for using medians with normally distributed data.[6]

Corrections for Outliers

When no outliers were added to the data, neither technique for correcting outliers had much effect on power compared with not using a correction. Surprisingly, the trimming technique was slightly superior to 3-*SD* elimination, even though trimming always eliminated 10% of the data points (2 of 20) and the 3-*SD* correction eliminated very few data points. (The percentages of outliers eliminated with the 3-*SD* technique are listed at the bottom of Table 7.)

When 4% outliers were added to the data, the trimming procedure was clearly superior to doing nothing and to the 3-*SD* technique. The 3-*SD* technique yielded little change from doing nothing.[7] Note that even though 4% outliers were added to the data, the 3-*SD* technique generally eliminated fewer than 2%. This is because the outliers influenced the mean and standard deviation of a condition and therefore affected the criterion for elimination.

Simulation 5: Outliers, Corrections, and Medians With Skewed Distribution

Method

Simulation 5 was similar to Simulation 4 except that we generated the raw data using a skewed distribution, as in Simulation 3. The procedure we used to add outliers also differed from that used in Simulation 4. If highly deviant points were added equidistantly above and below the mean, the skew would change substantially. As a practical example, consider reaction times. In a simple cuing experiment, the mean reaction times might be around 300 ms, with a standard deviation of 75–100 ms. Data points more than 5 standard deviations above the mean can occur in such experiments, but data points more than 5 standard deviations below the mean are physically impossible. In Simulation 5,

[6] In all the present simulations, we assumed that the level of measurement of the dependent variable was at least interval. If the data were ordinal, medians would be the appropriate transformation, and the parametric tests and other transformations would be inappropriate.

[7] In Simulation 4, the outliers were ±5 standard deviations from the population mean of each condition. If more extreme outliers are used, then the 3-*SD* technique is superior to doing nothing at all but is still worse than trimming.

Table 7
Percentage of Significant Results With Normal Distribution and Effect of Outliers

	Effect size							
	No outliers				Outliers			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
			No correction					
Mean raw	54.4	93.3	98.7	3.6	40.8	82.8	96.1	3.2
Z score	75.9	99.2	100	4.9	62.4	96.2	99.6	4.6
Ratio	55.8	94.1	99.2	3.7	42.7	85.6	97.0	3.2
Range	72.0	98.5	99.9	4.6	57.4	92.3	98.8	4.0
Mean log	54.6	93.3	98.7	3.4	40.8	83.1	95.8	3.4
Medians	42.5	86.7	97.6	3.5	41.2	83.0	96.3	4.1
			Trimming highest and lowest score					
Mean raw	56.2	90.9	99.0	4.1	46.2	84.5	97.6	3.1
Z score	78.0	98.8	100	6.5	68.2	96.5	100	5.5
Ratio	57.3	91.6	99.1	4.5	47.9	86.2	98.0	3.5
Range	75.9	98.5	100	6.2	66.4	95.2	99.6	5.3
Mean log	56.3	91.0	99.0	4.1	46.4	84.4	97.1	3.4
Medians	45.7	83.6	97.7	4.2	42.1	82.2	96.5	3.6
			Eliminating all scores beyond ±3 standard deviations					
Mean raw	54.3	90.7	98.5	3.0	41.8	83.0	96.0	3.5
Z score	75.2	98.5	99.9	4.9	64.0	95.7	99.8	4.7
Ratio	56.1	91.5	98.6	3.0	43.8	85.1	96.7	4.0
Range	71.8	97.4	99.8	4.4	58.4	93.1	99.4	5.3
Mean log	54.2	90.3	98.5	3.0	41.6	82.5	96.1	3.4
Medians	41.8	84.0	96.7	4.1	37.7	81.1	96.0	3.7
% rejected	0.08	0.08	0.08	0.07	1.7	1.7	1.7	1.7

therefore, we added 2% outliers that were 3 standard deviations below the mean and 1% outliers that were 6 standard deviations above the mean. This procedure preserved the expected mean of the distributions, did not affect the skew too much, and led to just more than 2% of the data points' being eliminated using the 3-*SD* procedure.[8]

Results

The results of Simulation 5 appear in Table 8. The results differed from those obtained with normally distributed data in several ways. Outliers drastically reduced the power of the different procedures, especially the use of the raw means. With outliers, the transformations were especially important, in some cases leading to three- and fourfold improvement. The rank ordering of the effectiveness of the different transformations varied with condition. For instance, the log transformation yielded better results than the *Z* transformation in the absence of outliers or in the presence of outliers when no correction for outliers was used, but it yielded worse results than *Z* when there were outliers and outlier correction was used. Outliers and the two corrections for outliers reduced the skew in the distributions, rendering the log a less appropriate transformation. The range transformation was similar to *Z*, that is, slightly better in some cases and slightly worse in others. The combination of the log and *Z* transformations performed quite well, often yielding the highest power.

Medians

In the absence of outliers, medians were consistently low in power, as with the normally distributed data in Simulation 4. The presence of outliers did not affect the power of the medians, but because it reduced the power of the other transformations, the medians fared relatively well. In several cases with outliers, medians yielded more significant results than any other transformation. Corrections for outliers had a limited effect on medians but large effects on other transformations. Any advantage of medians over other transformations was reduced or eliminated by the use of corrections for outliers.

Corrections for Outliers

Corrections for outliers were more important with skewed data than with normal data, and the effect of the correction

[8] We tried various procedures for adding outliers in both the normal and the skewed case. With normally distributed data, the effectiveness of the different analysis procedures was not affected by the procedure for adding outliers. More frequent or more distant outliers reduced overall power more, but the *Z* transformation was always best, trimming always helped, and so on. With skewed data, the outcomes were more sensitive to the technique for adding noise. In particular, the relative effectiveness of medians and logs varied markedly with different procedures. The procedure that we report met the overall properties that we wanted to satisfy and seemed reasonable to us based on the data from our laboratory.

Table 8
Percentage of Significant Results With Normal Distribution and Effect of Outliers

	Effect size							
	No outliers				Outliers			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
	No correction							
Mean raw	58.3	98.8	100	5.9	13.5	35.1	65.2	4.9
Z score	62.2	99.1	100	5.0	21.5	54.8	82.2	4.8
Ratio	61.5	99.2	100	5.1	21.4	52.9	79.4	4.7
Range	60.8	99.2	100	5.5	21.6	54.7	79.8	4.9
Mean log	69.1	99.6	100	5.3	34.9	84.8	99.5	4.5
Medians	47.2	94.8	99.7	5.5	43.5	93.2	99.8	4.5
Log *Z*	69.5	99.4	100	5.3	40.5	89.2	99.7	5.0
	Trimming highest and lowest score							
Mean raw	60.8	98.5	100	5.0	13.7	29.4	43.8	1.7
Z score	65.4	99.0	100	5.3	44.4	84.1	96.5	4.6
Ratio	65.0	99.0	100	5.0	41.4	78.2	93.6	4.4
Range	64.6	98.8	100	5.2	45.5	86.2	96.9	4.6
Mean log	69.3	99.6	100	5.4	41.5	83.1	97.3	3.2
Medians	50.9	96.3	100	5.7	48.3	92.5	99.7	3.7
Log *Z*	70.2	99.7	100	5.4	52.7	94.9	99.5	4.8
	Eliminating all scores beyond ±3 standard deviations							
Mean raw	56.6	98.4	100	5.0	8.7	23.1	37.7	1.5
Z score	60.0	98.9	100	5.0	37.9	77.5	93.3	5.4
Ratio	60.6	98.9	100	5.2	32.9	66.8	86.9	5.0
Range	60.8	98.9	100	5.2	38.8	78.6	94.0	5.3
Mean log	65.9	99.3	100	5.1	33.1	69.8	92.3	4.4
Medians	44.5	93.5	99.8	5.8	41.5	92.8	99.8	5.9
Log *Z*	67.9	99.3	100	5.0	47.7	91.9	99.0	5.3
% rejected	0.59	0.59	0.57	0.59	2.22	2.22	2.23	2.20

interacted with the type of transformation. The corrections for outliers decreased the power obtained with raw means, had little effect on medians or logs, and considerably increased the power of the other transformations. As with normally distributed data, trimming was superior to 3-*SD* elimination. We strongly recommend that researchers interested in using the *Z* or range transformation consider trimming. Trimming has little effect in the absence of outliers but a beneficial effect in the presence of outliers. In some case, trimming increased the power of the *Z* and range transformations by a factor of 2.

Simulation 6: Baseline Measures

Method

The designs used in the first five simulations were somewhat simplified from their normal implementation. In psychophysiology, in particular, researchers are interested in changes in the baselines of the various dependent measures. As a result, many experimenters take baseline measures prior to each condition. The measures for each condition are corrected for any change in baseline. Several procedures have been used for those corrections. One procedure is to calculate difference scores consisting of the value for a condition minus the value for that condition's baseline. A second procedure is to test for significant differences in the two baselines and apply a correction only if there is a significant difference.

In Simulation 6, we looked at the effect of the various transformations in a design with four cells: two conditions (experimental and control), each preceded by a baseline. In this simulation, we modeled an experiment in which there was no difference in the expected values of the two baselines. Except for the use of four cells rather than two, we used the same parameters as in Simulation 1. After choosing each subject's mean from a normal distribution and standard deviation from a chi-square distribution, we chose *m* (10 vs. 20) observations for each of the four cells. Effects were added only to the one cell, the experimental condition. Effects were added as in the absolute case in Simulation 1. The results were based on 10,000 rather than 1,000 experiments per condition, and the simulation was executed on the Convex.

We used two sets of transformations. In both sets, the parameters used in the calculation of the range, ratio, and *Z* transformations were calculated by using all four cells. In other words, the mean and standard deviation for the *Z* transformation were calculated using the scores from all four cells; likewise for the minimum, maximum, and range used in the other transformations. In one set, we applied the transformations only to the neutral and experimental conditions, ignoring the baseline conditions. In the other set, we applied the transformations to the difference scores for the neutral and experimental conditions from their respective baselines. We have omitted the log transformation from the report of this simulation. As in the other simulations in which normal distributions were used, logs were slightly worse than the simple mean of the raw scores.

Table 9
Percentage of Significant Results With Normal Distribution and Baseline and Difference Scores

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
	10 subjects per sample							
Mean raw	29.1	73.3	93.5	5.0	48.3	92.3	99.0	4.8
Z score	41.4	85.8	98.3	5.1	64.5	97.8	100	5.0
Ratio	31.3	74.0	94.7	5.4	51.4	93.4	100	4.6
Range	40.1	84.4	98.1	5.0	63.4	97.8	100	5.1
Mean difference	19.7	50.4	78.7	5.1	30.9	74.9	94.5	3.8
Z score difference	29.5	67.1	90.1	6.2	45.7	87.3	98.1	5.0
Ratio difference	19.2	52.3	79.2	4.9	32.5	76.4	96.3	4.6
Range difference	28.1	66.5	89.0	5.0	44.9	87.1	98.7	5.2
	20 subjects per sample							
Mean raw	51.5	95.3	99.9	5.0	77.4	100	100	4.7
Z score	75.3	99.5	100	4.9	95.8	100	100	5.8
Ratio	53.9	96.1	100	5.5	80.4	100	100	5.3
Range	73.9	99.6	100	5.2	94.6	100	100	5.1
Mean difference	30.1	78.7	97.7	5.0	53.5	95.3	100	4.3
Z score difference	54.4	95.2	100	4.9	80.2	100	100	5.1
Ratio difference	32.9	80.6	97.3	5.4	56.0	97.7	100	4.6
Range difference	52.5	94.1	100	5.2	78.5	99.9	100	5.4

Results

The results of Simulation 6 are shown in Table 9. As in Simulation 1, the use of the Z and range transformations led to substantial increases in power, with the Z transformation slightly superior. The use of difference scores lowered power, but this was expected because difference scores have a higher expected standard error. The expected variance of a set of difference scores is equal to the sum of the variances of the two sets of scores used in computing the differences. Assuming that the baseline and condition scores have the same variance, the difference scores will have twice the variance of either the baseline or condition scores. The standard error of the difference scores will be the square root of the variance (i.e., it will be 1.414 times as large as the standard error of either condition). Notice that the power for the difference scores was reduced by approximately that ratio for the smallest effect sizes.

Simulation 7: Baseline With Drift

Method

Judging only from the data in Table 9, it appears disadvantageous to use difference scores. Most investigators who recommend against using difference scores either try to allow baselines to stabilize at prestimulus levels (the closed-loop baseline procedure; see McHugo & Lanzetta, 1983) or begin by carrying out a significance test comparing the two baseline measures. Only if this test fails to reach significance do they carry out the test comparing the two conditions. We did not follow that latter procedure in Simulation 6. If we had, 5% of the baseline tests would have been significant. This follows from the fact that the baselines were generated from the same distributions as the conditions and that a test comparing the baselines is equivalent to the test comparing the conditions when the effect size is 0. In calculating the number of significant results obtained without using difference scores, therefore, we would have to subtract those cases in which the baselines were significantly different. That would be equivalent, on average, to reducing the number of significant results by 5% for each of the cases that did not involve difference scores. For cases with large effect sizes (e.g., 20 subjects, 20 observations, and of the standard deviation ¾ effect), difference scores actually would yield higher power than the baseline test technique because of the 5% artificial loss with the latter.

To illustrate the problem of not using difference scores, we simulated a case in which the baseline drifted in the direction opposite the effect. Simulation 7 was identical to Simulation 6 except that the baseline was reduced by the appropriate effect size, and that effect size was added back to the experimental condition.

The results are presented in Table 10. Because of the baseline drift, comparisons of the conditions that do not use difference scores have essentially zero power. The baseline drift causes only a small loss of power with difference scores. This is obviously a contrived situation, but it illustrates the danger of not taking difference scores. We did not apply the baseline drift to the null condition. If we had, the failure to take difference scores would have resulted in a substantial number of Type I errors.

The question of how to interpret change scores from different baselines remains a controversial topic. Subjects who exhibit different baselines may be experiencing different psychological states, making it impossible to interpret the effect of a treatment. We agree that the situation is problematic and that change scores do not solve the problem. We feel that the use of change scores is superior to not using change scores for the reasons just described, but the problem of the changing baseline is probably best solved through experimental design rather than data analysis. The closed-loop baseline procedure described by McHugo and Lanzetta (1983) is one design procedure that addresses the problem of unequal baselines.

Discussion

We examined the effect of various transformations in a particular paradigm, namely, one in which several observations are

Table 10
Percentage of Significant Results With Normal Distribution and Baseline Drift

	Effect size							
	10 observations per subject				20 observations per subject			
Transformation	1/4	1/2	3/4	Null	1/4	1/2	3/4	Null
	10 subjects per sample							
Mean raw	4.6	5.8	4.6	5.0	5.6	4.7	5.3	4.4
Z score	5.0	5.6	5.7	5.1	5.4	5.6	5.8	5.9
Ratio	4.0	5.1	4.0	4.9	4.7	5.5	5.1	4.6
Range	5.5	4.9	5.4	5.0	5.3	5.3	4.4	5.9
Mean difference	16.3	46.5	73.1	5.1	28.2	70.3	92.1	4.3
Z score difference	23.1	58.2	83.5	4.9	38.0	82.3	97.0	5.0
Ratio difference	17.3	48.0	75.4	4.7	29.4	73.9	93.3	4.8
Range difference	22.6	57.9	83.2	5.0	37.9	82.0	97.1	5.2
	20 subjects per sample							
Mean raw	5.4	4.7	5.2	5.0	5.3	5.1	4.2	4.7
Z score	5.3	5.0	5.5	4.6	4.8	5.4	5.3	5.2
Ratio	4.3	4.4	5.8	5.7	4.5	4.7	5.7	5.5
Range	5.9	4.3	5.6	5.1	5.0	5.5	5.5	4.9
Mean difference	28.5	74.6	95.7	5.1	49.2	94.8	100	4.7
Z score difference	43.9	90.7	99.8	5.9	69.3	99.6	100	4.9
Ratio difference	30.6	76.8	96.8	5.1	52.7	96.8	100	5.1
Range difference	42.6	90.4	99.1	5.3	68.5	99.1	100	5.4

collected for each subject in each condition. Such designs are common in psychophysiological and reaction time experiments. Although several different transformations have been suggested for these paradigms, most investigators compute a simple mean for each subject in each condition and then analyze those means, often after performing some correction for outliers, such as 3-*SD* elimination. Our simulations show that the most common analysis procedures are far from the most powerful. In almost every situation that we examined, there were transformations that yielded higher power than simple means. That increase in power was often substantial, in some cases a three- to fourfold increase. In no case did those transformations adversely affect the probability of a Type I error. Corrections for outliers were often quite influential, but, contrary to popular practice, trimming was more effective than 3-*SD* elimination.

In situations in which baselines might change over time, we examined the effect of the transformations using difference scores between baseline and treatments. The transformations had the same effect with difference scores that they did with raw scores. Logically, if baselines are likely to change, difference scores offer a more accurate picture of the effect of the independent variable.

Number of Subjects and Observations

The rank ordering of the effectiveness of the various transformations did not differ as a function of the number of subjects or the number of observations per subject, provided the effect size did not yield ceiling effects. For instance, the *Z* transformation was optimal with normally distributed data whether there were 10 subjects or 100 subjects or 10 observations per subject or 100 observations per subject.

In several of the simulations, we varied both the number of subjects and the number of observations between 10 and 20. In every case, power was increased more by increasing the number of subjects than by increasing the number of observations, but the difference between the two was not large. In most experiments, it is easier and cheaper to increase the number of observations than the number of subjects. Of course, in the present simulations we assumed that subjects were independent of one another and that observations were independent of one another. The latter assumption may be harder to justify in practice.

Types of Effect

We added three different types of effects to the experimental conditions: absolute, relative, and distributional. The rank ordering of the effectiveness of the various transformations did not differ as a function of effect type. Most parametric tests assume that effects are absolute (i.e., a constant is added to every score in a particular condition). We believe that some combination of relative and distributional effects is the most realistic, but the distinction does not appear important for our present purpose.

Why Transformations Help

Parametric tests, such as the *t* test, assume that the scores are identically distributed random variables. One aspect of that assumption is that each subject is assumed to have the same variance. As long as each subject contributes only one observation, or one observation per condition, there is no way of examining the equality of variance across subjects. In our simulations, we explicitly manipulated the variance across subjects. When variability differed across subjects, the transformations had a positive and, often, large effect. When all subjects were assigned the same variance, none of the transformations had any effect (see Simulation 2 in Table 5). The Z, ratio-, and range-correction transformations scale each subject's effect in terms of that subject's dispersion, thereby equating variance across subjects. The log transformation tends to equate variance across subjects and across conditions, but only if the data have a particular underlying distribution. Thus, if the data are normally distributed, logs do not increase the homogeneity of variance, but if the data follow a gamma distribution, logs do increase the homogeneity of variance.

On the basis of all the data that we have examined (from our laboratory and elsewhere), subjects do differ in variability in most tasks, often substantially. We believe that these differences are the rule rather than the exception.

Reasons for Not Transforming One's Data

Although in almost every case the appropriate transformation yields a substantial increase in power, there are several reasons for not transforming the data.

Strong Scales and Strong Inference

Sometimes the experimenter has reason to believe that the data represent an absolute, ratio, or interval scale and is interested in those absolute magnitudes, ratios, or intervals. There are some experiments in which the absolute magnitudes of the dependent measure are of interest for theoretical or logical reasons. For instance, a researcher might be interested in estimating nerve conduction rates in humans using a reaction time procedure. In that case, the investigator would wish to determine the absolute magnitude of the estimated time. Any transformation would destroy those absolute magnitudes. In a similar fashion, Townsend (1992) argued that there are situations in which reaction times can be demonstrated to lie on a ratio scale and the ratios are of theoretical interest. Again, any transformation other than simple multiplication would destroy the meaningful ratios.

In our opinion, though, many experimenters are primarily interested in whether an independent variable produces a significant effect and not in the absolute magnitudes, ratios, or intervals that were obtained. The values that are obtained are a function of the measuring instrument that is chosen and there often is not a compelling reason to choose one instrument over another. For example, in a maze experiment, the experimenter might measure motivation by measuring the time to reach the goal. Those reaction times could be converted to speeds by taking reciprocals. This is a nonlinear transformation and would certainly change the magnitudes, ratios, and so on. However, the experimenter could have measured speed in the first place with a radar gun. Because there might be no a priori basis for choosing between reaction time and speed, it is hard to know which one is correct.

Weak Scales

In some experiments, the data may not meet the assumptions for an interval scale. Perhaps only the ordinal information is valid. In that case, any monotonic transformation would preserve the meaningfulness of the data, and the transformations discussed herein are monotonic, at least within a subject. Unfortunately, because it is inappropriate to use parametric tests with such data, none of the results of the present simulations would apply. Furthermore, there are only a few distribution-free tests available for within-subject data, and our transformations would not affect those tests. For example, a sign test is affected only by the ordering of each subject's two scores. All the transformations that we examined in the present simulations would preserve the ordering of the two scores and leave the sign test unchanged.

Convention

Another reason for not choosing some transformation is the desire or need to compare the results with results in the literature. Differences in magnitude across similar studies might signal important procedural differences that had been overlooked. There is a danger, though, in choosing an analysis procedure solely on the basis of convention. It is possible that some null results in the literature arose from a lack of power rather than a lack of effect and that an appropriate transformation might have uncovered those effects. An obvious solution to this problem is for the researcher to analyze his or her data using both the conventional procedure and the most powerful procedure based on the results of our simulations. Both analyses could be reported, and differences in outcomes could be highlighted. Differences between analyses could provide useful information about individual differences.

Distortion of Pattern

As we described earlier, many investigators shy away from transformations, such as Z, because they worry that such transformations might distort the pattern of results. It is true that one can obtain a different ordering of conditions with raw means versus the Z transformation. However, as shown in Table 4, when there is a difference in ordering, the Z transformation is more likely than raw means to yield the correct ordering. Finally, reversals of orderings never occurred for significant effects.

Recommendations

On occasion, theory or logic requires the preservation of absolute magnitudes, and transformations should be avoided. On other occasions, convention dictates a particular form of data analysis (but see earlier section on convention). In the absence of either of those circumstances, we recommend the following.

Normally Distributed Data

For normally distributed data, the Z transformation yielded the highest power, followed closely by the range-correction transformation. Winer's (1971) trimming procedure enhanced power in the presence of outliers and did not reduce power in the absence of outliers. Therefore, we recommend the use of trimming in the type of design that we employed. The 3-*SD* elimination procedure was inferior to trimming in every way and should be avoided. Logs and medians reduced power with normally distributed data and should be avoided.

Skewed Data

For skewed data, the optimal transformation varies according to circumstance in complex ways. For instance, medians yielded the highest power with outliers but the lowest power without outliers. Furthermore, it is impossible to determine conclusively with real data whether extreme points are outliers or not. Therefore, we recommend the use of trimming coupled with the Z or range-correction transformation. This combination did not yield the highest power in many cases with skewed data, but it worked well in every case, was far superior to raw means, and is easy to employ and understand. We do not recommend the use of the combination of log and Z, even though it generally yielded the highest power, because the advantage over the Z transformation alone (provided trimming was used) was too small to justify the additional complexity. A final advantage of our recommendation to use the Z transformation with skewed data is that it is the same as our recommendation for normally distributed data. This consideration is potentially important. For our simulations with skewed data, we chose distributions that were highly skewed. As the degree of skew diminishes, the results converge on the results from the normally distributed data.

References

Ben-Shakhar, G. (1985). Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology, 22,* 292–299.

Ben-Shakhar, G. (1987). The correction of psychophysiological measures for individual differences in responsivity should be based on typical response parameters: A reply to Stemmler. *Psychophysiology, 24,* 247–249.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29,* 610–611.

Cacioppo, J. T., Tassinary, L. G., & Fridlund, A. J. (1990). The skeletomotor system. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social and inferential elements* (pp. 325–384). Cambridge, England: Cambridge University Press.

Edelberg, R. (1972). Electrical activity of the skin: Its measurement and uses in psychophysiology. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 367–418). New York: Holt, Rinehart & Winston.

Fowler, C. A., Treiman, R., & Gross, J. (in press). The structure of English syllables and polysyllables. *Journal of Memory and Language.*

Fowles, D. C. (1986). The eccrine system and electrodermal activity. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 51–96). New York: Guilford Press.

Graham, F. K. (1980). Representing cardiac activity in relation to time. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 139–246). New York: Wiley.

Hess, U., Kappas, A., McHugo, G. J., Lanzetta, J. T., & Kleck, R. E. (1992). The facilitative effect of facial expression on the self-regulation of emotion. *International Journal of Psychophysiology, 12,* 251–265.

Kruskal, J. B. (1968). Statistical analysis: Transformations of data. In *International Encyclopedia of the Social Sciences* (pp. 182–193). New York: Macmillan.

Levey, A. B. (1980). Measurement units in psychophysiology. In I. Martin & P. Venables (Eds.), *Techniques in psychophysiology* (pp. 597–628). New York: Wiley.

Lykken, D. T. (1972). Range correction applied to heart rate and GSR data. *Psychophysiology, 9,* 373–379.

Lykken, D. T., Rose, B., Luther, B., & Maley, M. (1966). Correcting psychophysiological measures for individual differences in range. *Psychological Bulletin, 66,* 481–484.

Lykken, D. T., & Venables, P. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology, 8,* 656–672.

McHugo, G. J., & Lanzetta, J. T. (1983). Methodological decisions in social psychophysiology. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 630–655). New York: Guilford Press.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 539–543.

Paintal, A. S. (1951). A comparison of the galvanic skin responses of normals and psychotics. *Journal of Experimental Psychology, 41,* 425–428.

Posner, M. I. (1978). *Chronometric explorations of mind.* Hillsdale, NJ: Erlbaum.

Roberts, F. (1979). *Measurement theory with applications to decision making, utility, and social sciences.* Reading, MA: Addison-Wesley.

Siddle, D., & Turpin, G. (1980). Measurement, quantification, and analysis of cardiac activity. In I. Martin & P. Venables (Eds.), *Techniques in psychophysiology* (pp. 139–246). New York: Wiley.

Smith, J. E. K. (1976). Data transformations in analysis of variance. *Journal of Verbal Learning and Verbal Behavior, 15,* 339–346.

Stemmler, G. (1987a). Implicit measurement models in methods for scoring physiological reactivity. *Journal of Psychophysiology, 1,* 113–125.

Stemmler, G. (1987b). Standardization within subjects: A critique of Ben-Shakhar's conclusions. *Psychophysiology, 24,* 243–246.

Townsend, J. T. (1992). On the proper scales for reaction time. In H. Geissler, S. Link, & J. Townsend (Eds.), *Cognition, information processing, and psychophysics: Basic issues* (pp. 105–120). Hillsdale, NJ: Erlbaum.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Venables, P., & Christie, M. (1980). Electrodermal activity. In I. Martin & P. Venables (Eds.), *Techniques in psychophysiology* (pp. 3–68). New York: Wiley.

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power and what can you do to correct the problem? *Current Directions in Psychological Science, 1,* 101–105.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Received December 18, 1990
Revision received September 17, 1992
Accepted September 18, 1992 ■