


# Surprising Implications of Differences in Locations Versus Differences in Means

David Trafimow<sup>1</sup> , Nathaniel Roth<sup>1</sup>, Lina Xu<sup>1</sup>, Deborah Toomasian<sup>1</sup>, Audrey Perrello<sup>1</sup>, Tingting Tong<sup>1</sup>, Tonghui Wang<sup>1</sup>, S. T. Boris Choy<sup>2</sup>, Xiangfei Chen<sup>1</sup>, Cong Wang<sup>3</sup>, Liqun Hu<sup>1</sup>

[1] *New Mexico State University, Las Cruces, NM, USA.* [2] *University of Sydney, Sydney, Australia.* [3] *University of Nebraska, Omaha, NE, USA.*

---

Methodology, 2023, Vol. 19(2), 152–169, <https://doi.org/10.5964/meth.10969>

**Received:** 2023-01-08 • **Accepted:** 2023-05-11 • **Published (VoR):** 2023-06-30

**Handling Editor:** Katrijn van Deun, Tilburg University, Tilburg, The Netherlands

**Corresponding Author:** David Trafimow, Department of Psychology, MSC 3452, New Mexico State University, P. O. Box 30001, Las Cruces, NM 88003-8001, USA. E-mail: [dtrafimo@nmsu.edu](mailto:dtrafimo@nmsu.edu)

**Supplementary Materials:** Materials [see [Index of Supplementary Materials](#)]



## Abstract

Social science researchers depend on differences in means between experimental and control conditions to draw substantive conclusions. However, an alternative is to use differences in locations. For normal distributions, means and locations are the same, but for skew normal distributions, means and locations are different. If a difference in means and locations are similar, and in the same direction, the resulting substantive story may be similar. However, if a difference in means and locations are dissimilar, especially if they oppose directionally, the resulting substantive story may differ dramatically. We collected 51 data sets from online data repositories to check how often the differences in means versus locations are substantially different or are in different directions. Although the values depend on what one counts, the overall conclusion is that the two types of differences have a larger than trivial chance of disagreeing substantially. We suggest that when researchers report normal statistics (mean and standard deviation), they should report skew normal statistics (location, scale, and shape) too, against the nontrivial chance that the skew normal statistics imply a substantive story in opposition to that implied by the normal statistics.

## Keywords

location, scale, shape, skew normal, Cohen's  $d$



This is an open access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), [CC BY 4.0](#), which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

A recently proposed taxonomy of assumptions, termed the TASI taxonomy, provides the starting point for the present research (Trafimow, 2019b, 2022a, 2022b). The TASI taxonomy contains four categories of assumptions: theoretical, auxiliary, statistical, and inferential. Theoretical assumptions are assumptions contained in a theory, and these feature nonobservational terms such as attitude, cognition, threat, and so on. To render greater specificity, researchers must add auxiliary assumptions that provide the conduit between nonobservational terms in a theory and observational terms in an empirical hypothesis. For example, although there is no way to observe attitude, it is possible to attempt to manipulate it or measure it, and auxiliary assumptions traverse the distance between unobservable attitudes and observable manipulations or measures.

However, even with the help of auxiliary assumptions, the level of specificity may nevertheless be insufficient. For example, suppose a researcher uses an attitude manipulation to have a pro-attitude condition and a neutral-attitude condition with respect to some dependent variable of interest. The empirical hypothesis is that scores in the pro-attitude condition will exceed scores in the neutral-attitude condition. But it is not immediately clear what that means. It could mean that every score in the pro-attitude condition will exceed every score in the neutral-attitude condition, but few researchers would wish to be held to that standard. Alternatively, it could mean that the pro-attitude mean exceeds the neutral-attitude mean, the pro-attitude median exceeds the neutral-attitude median, the pro-attitude 75th percentile exceeds the neutral-attitude 75th percentile, and so on. For most social science purposes, even the empirical hypothesis is not at a sufficient level of specificity and so it is necessary to add statistical assumptions to arrive at a statistical hypothesis. For example, a researcher might assume that the distributions are normal, thereby justifying an emphasis on means and standard deviations.

But researchers can make other assumptions too, which is the main present topic. To finish off the explanation of the TASI taxonomy, there remains the issue of making inferences about populations. For example, the researcher may wish to use sample statistics to estimate corresponding population parameters. To move in this direction, it is necessary to add yet another layer of assumptions, and these are inferential assumptions.

In summary, theoretical assumptions are insufficient for empirical hypotheses because they contain nonobservational terms. It is necessary to add auxiliary assumptions to traverse the distance between nonobservational terms in theories, and observational terms in empirical hypotheses. However, even empirical hypotheses are usually not sufficiently specific, and so it is necessary to add statistical assumptions to result in a statistical hypothesis. Finally, to make inferences about populations, it is necessary to add inferential assumptions. However, the present focus is at the statistical level and on what might be considered an overdependence of researchers on differences in means (Speelman & McGann, 2013, 2016).

## Normal Parameters Versus Skew Normal Parameters

Researchers typically focus on means when analyzing the results of the experiments they conduct. The usual prediction is that the mean in one condition should be greater than the mean in another condition. Researchers interpret confirmation of the prediction to support the touted empirical hypothesis which in turn, they take to support the theory.

However, as we have seen, there are many potential statistical interpretations of a hypothesis that scores in one condition should exceed scores in another condition. There is no clear a priori necessity why means should dominate to the extent that they do (Speelman & McGann, 2013, 2016). Perhaps a reason for the dominance of differences in means is that researchers are accustomed to think in terms of the family of normal distributions, which have two parameters: mean and standard deviation. Under normality, if a researcher knows the means and standard deviations of the conditions, the researcher knows everything there is to know. Thus, it makes sense for researchers to focus on means and standard deviations of each condition. It is an economical way to convey complete information. And yet, this is true only for the family of normal distributions. If the distribution is skew normal, then means and standard deviations are not defining parameters, and reporting them no longer provides complete information (Azzalini, 1985, 2014; Azzalini & Capitanio, 1999). Multiple investigations show that skewness is the rule, not the exception (Blanca et al., 2013; Ho & Yu, 2015; Micceri, 1989).

The family of normal distributions is a subset of the family of skew normal distributions, and we present the pdf and cdf in the [Supplementary Materials](#). In contrast to normal distributions, skew normal distributions have three parameters: location, scale, and shape. If the shape parameter equals zero, the distribution is normal, the location equals the mean, and the scale equals the standard deviation. However, if the shape parameter does not equal zero, the distribution is not normal but rather skew normal. In addition, the location no longer equals the mean, and the scale no longer equals the standard deviation. When there is skew normality, as opposed to normality; then the defining parameters are location, scale, and shape; not mean and standard deviation. Thus, to present a complete picture of skew normal distributions, it is necessary to report location, scale, and shape.

At this point, it is possible to argue that the foregoing is much ado about nothing because surely differences in means between conditions are in the same direction as differences in locations between conditions. So, why make a big deal about skew normal parameters as opposed to normal parameters? However, on the contrary, it is not true that differences in means and differences in locations must be in the same direction. It is entirely possible that they are in different directions (Trafimow et al., 2019).

To see why it matters if the difference in locations is in the opposite direction of the difference in means, consider the theory of reasoned action as an example (e.g., Ajzen & Fishbein, 1980; Fishbein, 1980; Fishbein & Ajzen, 1975, 2010). According to this theory, among other assumptions we need not discuss here, attitudes cause behavioral

intentions. Imagine that a researcher randomly assigns participants to be exposed to a pro seat belt essay or a neutral essay, with a subsequent measure of intentions to wear seat belts. The theory-based prediction is that the pro essay should shift the distribution of intention scores in a positive direction, so that the location in the pro condition should exceed the location in the neutral condition. Seemingly consistent with the prediction, suppose that the mean in the pro condition exceeds the mean in the neutral condition. Such a finding would elicit practically universal agreement that the experiment “works” and supports both the empirical hypothesis and the theory.

However, let us further complicate the scenario by imagining that the difference in locations is in the opposite direction; that is, the location is larger in the neutral condition than in the pro condition. In that case, we have contradictory substantive stories. According to the difference in means, the experiment works, as explained in the previous paragraph. However, according to the difference in locations, the pro essay decreased behavioral intention scores relative to the neutral condition. The experiment still works, but in reverse of researcher hopes. That is, the difference in locations contradicts that the essay manipulation functions as is supposed to function. At the theoretical level, the difference in means supports the theory whereas the difference in locations contradicts the theory. But there is generally no way to know about the contradictory findings, if indeed they are contradictory, because researchers rarely report locations, scales, or shapes.

In cases where differences in locations are in the opposite direction of differences in means, which should we believe? The answer may depend on the goal of the research. In the foregoing attitude example, the prediction is that the pro attitudinal essay should shift the distribution of scores in the positive direction, so the difference in locations really is the better test. A problem with believing the difference in means is that one of the essays could simply have changed the shape of a distribution, without shifting the location, and thereby not fairly tested the hypothesis. In contrast, there may be applied goals, whereby changing the shape of the distribution may be sufficient, even if the location does not shift, in which case it would be possible to argue for taking the difference in means seriously. The larger point is not that the difference in locations is always superior to the difference in means, or the reverse, but that if the differences are in opposite directions, it requires intense thinking to decide which to emphasize.

The obvious soft spot in our argument, thus far, is that although it is possible for differences in locations to be opposite to differences in means, this might not happen often. Suppose, for example, that it only happens on 1% of occasions. In that case, it is reasonable to argue that the potential problem is seldom an actual problem, and so there are more pressing issues on which to concentrate research efforts. In contrast, suppose that it happens on, say, 5% of occasions, or perhaps even more than that. In that case, it is no longer plausible to argue that the possibility of effects in the opposite direction is not

a pressing issue. After all, it is possible that in 5% or more of the literature, researchers are telling the opposite substantive story of what they ought to be telling.

And thus, we arrive at the present goal. We analyzed 51 data sets haphazardly downloaded from data repositories. Our goal was to determine how often differences in locations would be in the opposite direction of differences in means. If it happens seldom, such as 1% of the time, then that would support that there is little about which to worry. However, if it happens more often, such as more than 5% of the time, that might be a reason to change data analysis and data reporting. There were no hypotheses: this was an exploratory study to determine the degree or frequency with which skew normal statistics disagree with normal statistics.

## Method

We obtained 51 social science data sets from the following data depositories: OSF.io and openicpsr.com. Data collection was haphazard, but with the stipulations that there had to be at least two conditions and at least one continuous dependent variable. If there were more than two conditions or more than one continuous dependent variables in a study, then we chose two conditions and one dependent variable randomly.

We analyzed each of the data sets using Excel to obtain means, standard deviations, and skewness for each condition in each study. In turn, we used these to estimate locations, scales, and shapes for each condition in each study.<sup>1</sup> The result is that we collapsed the 51 data sets into a single master data sheet containing normal and skew normal statistics for each condition of each study. The master data sheet contains 51 rows, with each row representing a study and containing all the relevant statistics. This master data sheet can be accessed at the [Supplementary Materials](#).

## Results

There are multiple ways to compare differences in locations with differences in means across the 51 data sets. The most obvious way, as alluded to before, is to calculate the percentage of times the difference in means is in the opposite direction of the difference in locations. There were four cases where either the difference in means or the difference in locations equaled zero, and one case where both equaled zero. If we count the instance where both equaled zero as being in the same direction, but the other three cases where only one difference equaled zero as in opposite directions, then differences in means and differences in locations were in opposite directions for 37% of the cases. Probably

---

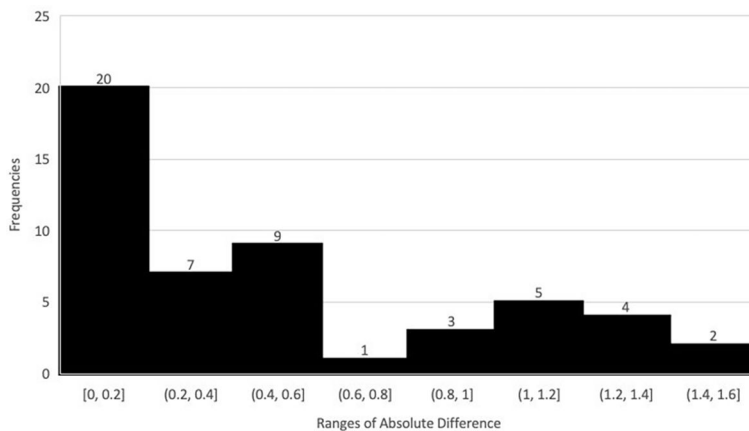
1) McCullough and Wilson (1999) and McCullough and Heiser (2008) reported problems with Excel. However, those problems are not relevant to our use of the spreadsheet.

a fairer option is to simply not count zeros, leaving 47 cases, and 34% of these resulted in differences in opposite directions. Or to be as conservative as possible, we can count all cases where there is at least one zero as exemplifying differences in means and locations as being in the same direction, rendering 31% of differences in means and locations in opposite directions. Hereafter, we will use the most conservative value of 31%, recognizing that perhaps the fairer value is 34%, but even the conservative value indicates a serious issue.

Another way to consider the difference in locations versus the difference in means is to calculate Cohen's  $d$  for each difference and investigate the extent of the absolute differences between these Cohen's  $d$  values. When basing Cohen's  $d$  on normal parameter estimates, we used the difference in means divided by the pooled standard deviation. When basing Cohen's  $d$  on skew normal parameter estimates, we used the difference in locations divided by the pooled scale. The mean absolute difference between Cohen's  $d$  based on normal versus skew normal statistics was 0.49, the median absolute difference was 0.37, and the standard deviation of the absolute differences was 0.47. In addition, [Figure 1](#) provides a histogram. Although 20 of the data sets resulted in reasonably similar values for Cohen's  $d$  regardless of whether it was computed according to normal versus skew normal statistics, the other 31 data sets resulted in varying degrees of divergence.

**Figure 1**

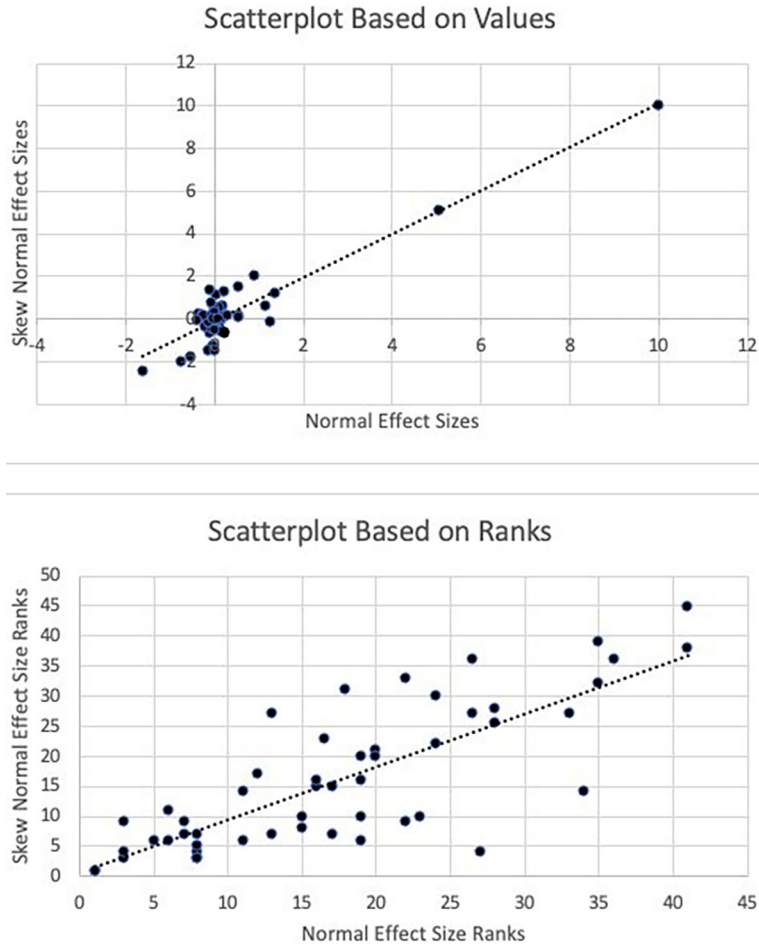
*Histogram Showing the Frequencies of Data Sets Within Each of the Ranges of Absolute Difference*



We also correlated the normal Cohen's  $d$  values and the skew normal Cohen's  $d$  values to see how well each predicts the other. Although the correlation coefficient was larger than zero, thereby supporting some predictability, it was nevertheless an unimpressive value of 0.28. However, a scatterplot (top scatterplot in [Figure 2](#)) suggests that the reason

**Figure 2**

Scatterplots of Normal Effect Size Values Versus Skew Normal Effect Size Values (Top Scatterplot) or Normal Effect Size Ranks Versus Skew Normal Effect Size Ranks (Bottom Scatterplot)



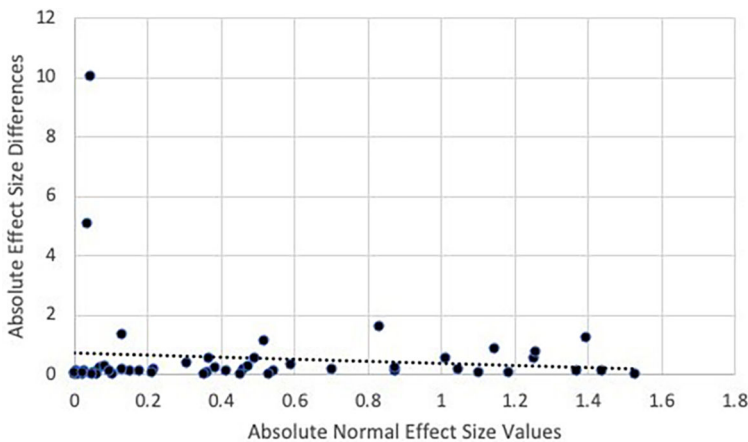
for the unimpressive correlation coefficient is too much clustering of scores. A way to address that issue is to convert all scores to ranks first. In that case, the correlation increases to 0.80 (bottom scatterplot in Figure 2). Nevertheless, if we are to take values themselves seriously, as opposed to merely using them to indicate ranks, the more pessimistic correlation coefficient should not be dismissed.

Lastly, let us consider a potential argument that goes as follows. Many calculated skew normal effect sizes diverge from normal effect sizes because too many normal

effect sizes have small magnitudes in either the positive or negative direction. If we had only included effect sizes with large magnitudes, then skew normal effect sizes would be much closer to them and more often in the same direction. One problem with this argument is that there is little prior reason to believe it. More important, according to this line of argument, we should expect that the absolute differences in the two kinds of effect sizes should decrease as the normal effect size magnitudes increase. That is, there should be a strong negative correlation coefficient. However, the correlation coefficient was only  $-0.15$ , which hardly provides convincing support for the potential argument. Nor does the scatterplot depicted in Figure 3 provide support.

**Figure 3**

*Scatterplot of Normal Effect Size Values Versus the Absolute Differences Between Normal and Skew Normal Effect Sizes*



It is also possible to investigate the extent to which the direction of differences in means versus differences in locations depends on normal effect size magnitudes. To address this, we dummy coded cases where the differences in means and differences in locations were in the same direction as +1 and different directions as -1, with zeroes deleted, leaving 47 cases. The prediction of differences in opposite directions from the normal effect sizes was only 0.21, thereby not supporting the cruciality of normal effect sizes in determining the probability of effects in opposite directions. Then, too, note that the slight  $-0.15$  value without dummy coding and the slight 0.21 value with dummy coding, are themselves in opposite directions. More generally, whether there is dummy coding or not, it is implausible that normal effect sizes are crucial for determining inconsistencies between differences in means versus differences in locations. This is not to say that the



magnitude of the effect is completely irrelevant for the difference in normal versus skew normal effects.

## Discussion

Not only can differences in means be in the opposite direction of differences in locations, but in the data sets we used, different directions occurred on 31% of occasions, which is nontrivial. Furthermore, although Figure 1 shows that sometimes effect sizes based on estimates of normal parameters are close to effect sizes based on estimates of skew normal parameters, often they are not. Nor are effect sizes based on estimates of normal parameters good predictors of estimates based on skew normal parameters, taking the values seriously. A caveat is that prediction improves markedly if the values are transformed to ranks, thereby decreasing the clustering of data points. Finally, there is little reason to believe that the divergence between normal and skew normal effect sizes, whether in extent or direction, depends crucially upon normal effect size magnitudes.

There are limitations, and we wish to be upfront about them before moving to implications, to guard against exaggeration. One limitation, and the most important one, is that the data sets were obtained haphazardly as opposed to randomly. As haphazard selection need not be the same as random selection, it is not clear how well the present findings generalize to the population of published studies or the population of studies that will be performed in the future. A related generalizability problem is the lack of assurance that data sets made publicly available adequately represent the totality of data sets. Perhaps the population percentage of cases where the difference in means and the difference in locations are in opposite directions is some amount larger or smaller than our conservative value of 31% or our perhaps fairer value of 34%. Still, it would be a stretch to assume that the population value is so much smaller that the issue becomes trivial. A reasonable interpretation might be that although the presently reported statistics provide ample reason to believe that the issue of normal versus skew normal statistics is important and demands attention, it is too early to draw strong conclusions with respect to exact probabilities of differences in direction between normal versus skew normal statistics.

Another limitation is that there are families of distributions other than normal and skew normal ones. It could be that in some cases, it would have been best not to estimate means or locations, but rather some other parameter. For example, for lognormal distributions, the parameters are the logarithmically transformed means and standard deviations. In that case, perhaps neither means nor locations are ideal, but rather means of logarithmically transformed scores should be used. And other distributions have yet different parameters that are distribution appropriate. One way to address this limitation in future research is to perform a large-scale study, across many data sets, to see how often different distributions fit the data obtained from different studies where normality

is typically assumed. A caveat is that some data sets might fit well with none of the distributions researchers typically consider, or that a data set might fit well with more than one distribution, thereby rendering categorization difficult. The performance of such a study would require rules for deciding whether data sets should be assigned as exemplifying one family of distributions as opposed to alternative families of distributions.

Finally, a limitation is that our focus was descriptive, not inferential. A reason is that there is much disagreement about that which constitutes sound inferential procedures. The special issue of *The American Statistician* contains a wide variety of viewpoints, many of which criticize null hypothesis significance testing in favor of other procedures, including Bayes factors, second generation P-values, the a priori procedure, and frequentist or Bayesian confidence intervals. Although we have strong inferential opinions, for the present paper, it is unnecessary to commit to an inferential position. One inferential direction to go in future research on the 51 data sets is to consider the log-likelihood functions under both dependent and independent assumptions to obtain maximum likelihood estimates (MLEs) of parameters.

## The Obvious Solutions

Sometimes researchers perform a significance test to see if their sample data in their different conditions depart too much from what would be expected if the population distributions were normal. If the significance tests are not statistically significant, the researcher assumes that the population distribution is normal. Or even without a significance test, researchers might calculate skewness, and find that the values are close to zero, thereby again leading to the conclusion that the population distributions are normal or near normal. Or researchers might invoke the Central Limit Theorem to argue that even if there is some non-normality, it does not matter much and it is safe to assume normality. In all these cases, the conclusion is that there is little to worry about.

But seemingly trivial differences in skewness may be extremely important. Consider again where a researcher wishes to test the theory that attitudes cause intentions to perform behaviors. The researcher randomly assigns participants to a pro attitude or neutral condition and measures intentions to use seatbelts on a scale where negative values indicate negative intentions and positive values indicate positive intentions. Suppose that the mean is +2 in the pro condition, +1 in the neutral condition, and the standard deviation is 2 in both conditions; thus, Cohen's  $d = 0.50$ . Finally, suppose the skewness is 0.10 in the experimental condition and -.10 in the control condition; these routinely would be considered trivial levels of skewness and would be unlikely to cause statistically significant differences from normality unless the sample sizes are abnormally large. Practically anyone would conclude that the findings support the theory.

Yet, running out the skew normal calculations renders a contradictory substantive story. The estimated locations are 0.77 and 2.23 in the experimental and control conditions, respectively. The estimated scale is 2.35 in both conditions and the estimated

shapes are 0.87 and -0.87 in the experimental and control conditions, respectively. Further, the skew normal effect size is -0.62, which is in the opposite direction from the normal effect size of 0.50. If we focus on normal statistics, the data support the theory; if we focus on skew normal statistics, the data contradict the theory. Even seemingly trivial skewness can render differences in means extremely misleading.

## The Other Obvious Solution

There is another obvious solution that fares better than the foregoing ones. And that is for researchers not to settle for reporting normal statistics when assuming normality, but to report skew normal statistics too. Practically all statistics programs, and even many spreadsheets such as Excel, provide the sample mean, standard deviation, and skewness. From there it is trivially easy to estimate location, scale, and shape parameters (see the Supplementary Materials). The calculations in the example required only a few minutes. If normal and skew normal statistics are reported, then reviewers, editors, and people with a stake in the research can make their own judgments about whether to emphasize normal or skew normal statistics.

## Judgment

Every research case is different, and we believe that researchers should consider each with respect to its own idiosyncrasies. That said, it is possible to have rules of thumb, though researchers should feel free to disregard these if the idiosyncrasies of the case at hand warrant it.

To explain our rules of thumb, it is convenient to divide research into the traditional categories of research designed to test theories versus applied research. For theory-testing research, where researchers randomly assign participants to experimental and control conditions, the goal is to show that the manipulation shifts the location of the distribution. If the manipulation merely changes the shape of the distribution, but does not shift it in the desired direction, then support for the theory is compromised. To assess an actual distribution shift, under skew normality, skew normal parameters are clearly better than normal parameters. Specifically, the difference in locations should be trusted over the difference in means. Or in terms of effect sizes, skew normal effect sizes should be trusted over normal effect sizes.

Matters may differ for applied research. To see this, consider again the experiment where participants are randomly assigned to a pro attitude essay or a neutral essay. If we interpret the experiment as having been performed to test the theory that attitudes cause intentions to perform behaviors, then the difference in locations should be trusted over the difference in means, as we argued in the previous paragraph. However, if the researcher does not care about theory, but merely wishes to know whether the pro attitude essay works to increase intentions to use seat belts, it is possible to argue that

the difference in means should be preferred to the difference in locations, though we emphasize that this is a judgment call.

To see why, let us ask an obvious applied question. Assuming a sufficiently large sample size to engender trust that the sample statistics are good estimates of the corresponding population parameters, what is the probability of being better off, to varying degrees, with the intervention than without it? Trafimow et al. (2022) recently provided the mathematics for answering such questions, and they also provided a free and user-friendly computer program (also see Tong et al., 2022), which can be accessed in the [Supplementary Materials](#).

To use the program, there are two preliminary steps. The first step is to use a statistical package or spreadsheet to obtain sample means, standard deviations, and skews for both conditions. The second step is to use the equations near the end of the [Supplementary Materials](#) to estimate location, scale (squared), and shape parameters for both conditions. Once the preliminary steps are completed, the next step is to activate the link, and there will be eight places to instantiate values. The estimates for location, scale-squared, and shape for both conditions can be entered into the first six of these places.

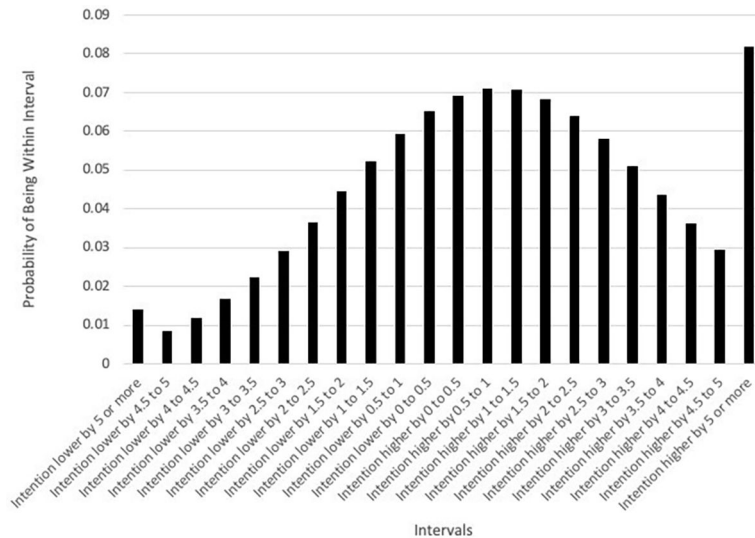
There are also two places labeled  $a$  and  $b$ . These are based on an equation:  $Z = X + aY + b$ . One strategy is to set  $a = -1$ , and then let  $b$  take on various values. Using this strategy provides probabilities of being better off or worse off by various values of  $b$ . These probabilities can be entered into a spreadsheet to facilitate constructing a G-P diagram. Another strategy is to set  $b = 0$ , and vary  $a$ . The difference is that the first strategy involves specific values whereas the second strategy involves multiples. For example, one might be interested in the probability of having higher or lower blood pressure, by 4,8,12,..., blood pressure points, if one takes a certain blood pressure medicine. In this case, it is best to use the first strategy. Alternatively, one might be interested in the probability of being 1.1,1.2,..., times better off, or worse off with the blood pressure medicine. In that case, it is best to use the second strategy. Either way, the final result will be a G-P diagram, such as that described below.

Applying the program to the example data in the attitude experiment, the result is that the probability that a randomly selected person from the pro essay condition would have a higher intention to use seatbelts than a randomly selected person from the neutral essay condition is 0.64. More generally, we can ask about the probability that a randomly selected person from the pro essay group will have lower or higher intentions to use seatbelts in various ranges. [Figure 4](#) depicts these in a single convenient gain-probability diagram.

Consider that the probability of higher intention scores is greater if one is randomly selected from the pro attitude condition than the neutral condition, and the asymmetrical nature of [Figure 4](#). From an overall applied perspective, the pro attitude essay increases the probability of intending to use seatbelts. If that is the applied goal, and theory is

Figure 4

## Gain-Probability Diagram



Note. Shows the probabilities that a randomly selected participant from the pro attitude condition will have lower or higher intentions than a randomly selected participant from the neutral condition, by varying amounts, with intervals of 0.50 scale points.

unimportant, then the pro attitude essay works. Because the difference in means gives that message, and the difference in locations seems to indicate the opposite message, for the applied goal, the difference in means may be superior to the difference in locations.

However, a caveat accompanies this conclusion. If the goal is applied, there is no reason for researchers to settle for a difference in means. In that case, it is more useful to estimate the probability of being better off in the pro condition than in the neutral condition. Or better yet, it is more useful to know the probabilities of being better off or worse off by varying degrees, such as Figure 4 illustrates. If having an applied goal is the reason for using means rather than locations, then skew normal statistics still matter because they are necessary to construct gain-probability diagrams, such as Figure 4. Therefore, our advice is to always provide skew normal statistics. Even if the researcher is uninterested in probabilities of the type illustrated in Figure 4, some readers surely will be interested, and the onus is on the researcher to provide the requisite information so that stakeholders in the research can make the desired probability calculations.

## Differences in Effect Size Magnitudes

Because social science researchers tend to focus more on effect directions than on effect sizes, we commenced the Results section with the percentage of times location differences were in the opposite direction of mean differences. However, [Figure 1](#) shows that skew normal effect sizes often differed substantially from corresponding normal effect sizes with respect to magnitude. Given the current zeitgeist, it might be tempting to argue that in those cases where both normal and skew normal effects are in the same direction, if they differ substantially in size, it does not matter much because both effect sizes support the same substantive story. However, we see at least three potential problems with this argument.

The first potential problem is that the argument tacitly assumes that all predictions are directional. Although this is usually true in the social sciences, this fact could be argued a limitation that social scientists ought to endeavor to overcome. With better assumptions across the TASI taxonomy, researchers could learn to make more precise point predictions and not settle for directional ones.

The second problem is that even under directionality, effect size matters. Imagine an experiment where the effect size is 0.1 versus 0.9. In the former case, it is easy to advance plausible alternative explanations: perhaps the randomization did not work perfectly, perhaps there was a confound, and so on. In the latter case, although these sorts of alternative explanations remain possible, they are less plausible. It would be quite a stretch, for example, to attempt to account for an effect size of 0.9 based on imperfect randomization.

Now, suppose the normal effect size is small and the skew normal effect size is large. If the normal effect size dominates, there is stronger reason to distrust the accompanying substantive story due to the plausibility of alternative explanations. In contrast, if the skew normal effect size dominates, there is stronger reason to trust that story due to the relative implausibility of alternative explanations. Or if it is the skew normal effect size that is small and the normal effect size that is large, then emphasizing the skew normal effect size might indicate distrust and emphasizing the normal effect size might indicate relative trust.

Finally, and perhaps inevitably, there is the issue of application. If the effect size is small, even if in the hoped-for direction, there is a poor case for application whereas a large effect size may militate more powerfully for an application. Interventions, policies, and so on have costs for which small effect sizes may provide insufficient justification. If either the normal or skew normal effect size is small, and the other is large, the decision to trust one or the other might be influential in the decision to incur the costs of an intervention, policy, or other application. As [Figure 1](#) shows this happens somewhat often, distinguishing whether to emphasize normal or skew normal statistics, possibly augmented by a gain-probability diagram such as that illustrated by [Figure 4](#), is a nontrivial issue.

## Conclusion

For theory-testing research, where the goal is to perform an experimental manipulation that shifts the location of one distribution relative to the other, the difference in locations is superior to the difference in means for testing that goal. However, if the goal is applied, then what matters more than means or locations is the probability of being better off, or worse off, to varying degrees, depending on condition. The difference in means may accord better than the difference in locations with such probabilities. However, even in such cases, it is still desirable to focus on skew normal statistics because they are necessary to obtain the probabilities of interest. Thus, skew normal statistics are valuable regardless of whether the goal is to test a theory or test an application. In turn, given the ease of obtaining skew normal statistics, our recommendation is as follows. Unless there is a clear reason otherwise, researchers should routinely report skew normal statistics along with the normal statistics they now report. Although this would constitute a large change in the practice of social science researchers, the change is well-justified.

Unfortunately, that a change is well-justified does not mean it will happen. [Stunt et al. \(2021\)](#) conducted focus groups with substantive researchers, journal editors, and grant funders to discuss social scientists' reluctance to deviate from null hypothesis significance testing despite its many documented problems (see [Hubbard, 2015](#); [McCloskey & Ziliak, 2010](#); [Trafimow, 2019a](#) for reviews). Although many of the substantive researchers thought such a change beneficial for science, they nevertheless indicated being unwilling to change unless journal editors or grant funders changed first. Likewise, grant funders wished to wait for journal editors to change their statistical policies. That different groups were awaiting other groups to initiate change exemplifies why change is difficult in the social sciences. In the face of such social inertia, we see two main avenues for change. If journal editors demand skew normal statistics as a condition for publication, the researchers who submit to those journals would soon follow suit. Secondly, a slower route for change would be if substantive researchers commenced reporting skew normal statistics along with normal statistics. Such 'bottom-up' change, featuring a slowly increasing groundswell of opinion, would be slower than change induced 'top-down' from journal editors. Either way, we hope and expect that the present work will stimulate change in the statistics that researchers compute and report.

---

**Funding:** The authors have no funding to report.

---

**Acknowledgments:** The authors would like to thank the editor and two reviewers for their helpful comments, which led to the improvement of this paper.

---

**Competing Interests:** The authors have declared that no competing interests exist.

---

## Supplementary Materials

For this article, the following materials are available: an Appendix defining the skew normal distributions and parameter moment estimations, an R Shiny app that calculates better score probabilities of (non)-intervention, and a master spreadsheet of 51 social science data sets (for access see [Index of Supplementary Materials](#) below):

### Index of Supplementary Materials

- Trafimow, D., Roth, N., Xu, L., Toomasian, D., Perrello, A., Tong, T., Wang, T., Choy, S. T. B., Chen, X., Wang, C., & Hu, L. (2023). *Supplementary materials to "Surprising implications of differences in locations versus differences in means"* [Skew normal distribution definitions, parameter moment estimation]. *PsychOpen GOLD*. <https://doi.org/10.23668/psycharchives.12942>
- Roth, N. (2022). *Surprising implications of differences in locations versus differences in means* [Dataset and statistics spreadsheet]. OSF. <https://osf.io/cgp5z>
- Trafimow, D., Hyman, M. R., Kostyk, A., Wang, Z., Tong, T., Wang, T., & Wang, C. (2022). *Supplementary materials to "Gain-probability diagrams in consumer research". Gain Probability: Test With and Without Intervention* [R Shiny App]. [https://probab.shinyapps.io/inde\\_prob/](https://probab.shinyapps.io/inde_prob/)

## References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice-Hall.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178. <http://www.jstor.org/stable/4615982>
- Azzalini, A. (2014). *The skew-normal and related families* (Vol. 3). Cambridge University Press.
- Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 579–602. <https://doi.org/10.1111/1467-9868.00194>
- Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78–84. <https://doi.org/10.1027/1614-2241/a000057>
- Fishbein, M. (1980). A theory of reasoned action: Some applications and implications. In H. Howe, & M. Page (Eds.), *Nebraska symposium on motivation* (Vol. 27, pp. 65–116). University of Nebraska Press.
- Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric*, *10*(2), 130–132. <https://philarchive.org/rec/FISBAI>
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology Press.
- Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. <https://doi.org/10.1177/0013164414548576>



- Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. SAGE Publications.
- McCloskey, D. N., & Ziliak, S. (2010). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.
- McCullough, B. D., & Heiser, D. A. (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics & Data Analysis*, 52(10), 4570–4578. <https://doi.org/https://doi.org/10.1016/j.csda.2008.03.004>
- McCullough, B. D., & Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics & Data Analysis*, 31(1), 27–37. [https://doi.org/https://doi.org/10.1016/S0167-9473\(99\)00004-3](https://doi.org/https://doi.org/10.1016/S0167-9473(99)00004-3)
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166. <https://doi.org/10.1037/0033-2909.105.1.156>
- Speelman, C. P., & McGann, M. (2013). How mean is the mean? *Frontiers in Psychology*, 4, Article 451. <https://doi.org/10.3389/fpsyg.2013.00451>
- Speelman, C. P., & McGann, M. (2016). Challenges to mean-based analysis in psychology: The contrast between individual people and general science. *Frontiers in Psychology*, 7, Article 1234. <https://doi.org/10.3389/fpsyg.2016.01234>
- Stunt, J., van Grootel, L., Bouter, L., Trafimow, D., Hoekstra, T., & de Boer, M. (2021). Why we habitually engage in null-hypothesis significance testing: A qualitative study. *PLoS ONE*, 16(10), Article e0258330. <https://doi.org/10.1371/journal.pone.0258330>
- Tong, T., Wang, T., Trafimow, D., & Wang, C. (2022). The probability of being better or worse off, and by how much, depending on experimental conditions with skew normal populations. In S. Sriboonchitta, V. Kreinovich, & W. Yamaka (Eds.), *Credibile asset allocation, optimal transport methods, and related topics: Studies in systems, decision and control* (Vol. 429, pp. 261–284). Springer-Verlag. [https://doi.org/10.1007/978-3-030-97273-8\\_18](https://doi.org/10.1007/978-3-030-97273-8_18)
- Trafimow, D. (2019a). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics*, 7(2), Article 26. <https://www.mdpi.com/2225-1146/7/2/26>
- Trafimow, D. (2019b). A taxonomy of model assumptions on which p is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, 22(6), 571–583. <https://doi.org/10.1080/13645579.2019.1610592>
- Trafimow, D. (2022a). Generalizing across auxiliary, statistical, and inferential assumptions. *Journal for the Theory of Social Behaviour*, 52(1), 37–48. <https://doi.org/10.1111/jtsb.12296>
- Trafimow, D. (2022b). A new way to think about internal and external validity. *Perspectives on Psychological Science*. Advance online publication. <https://doi.org/10.1177/17456916221136117>
- Trafimow, D., Hyman, M. R., Kostyk, A., Wang, Z., Tong, T., Wang, T., & Wang, C. (2022). Gain-probability diagrams in consumer research. *International Journal of Market Research*, 64(4), 470–483. <https://doi.org/10.1177/14707853221085509>
- Trafimow, D., Wang, T., & Wang, C. (2019). From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement*, 79(1), 129–150. <https://doi.org/10.1177/0013164418764801>



*Methodology* is the official journal  
of the European Association of  
Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing  
service by Leibniz Institute for  
Psychology (ZPID), Germany.