Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation, 9(6)* Online at http://pareonline.net/getvn.asp?v=9&n=6

The Power of Outliers (and Why Researchers Should *Always* Check for Them)

Jason W. Osborne and Amy Overbay

North Carolina State University

There has been much debate in the literature regarding what to do with extreme or influential data points. The goal of this paper is to summarize the various potential causes of extreme scores in a data set (e.g., data recording or entry errors, motivated mis-reporting, sampling errors, and legitimate sampling), how to detect them, and whether they should be removed or not. Another goal of this paper was to explore how significantly a small proportion of outliers can affect even simple analyses. The examples show a strong beneficial effect of removal of extreme scores. Accuracy tended to increase significantly and substantially, and errors of inference tended to drop significantly and substantially once extreme scores were removed.

The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or nonparametric tests (e.g., Zimmerman, 1994, 1995, 1998). Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort. This inference is supported empirically by Osborne, Christiansen, and Gunter (2001), who found that authors reported testing assumptions of the statistical procedure(s) used in their studies--including checking for the presence of outliers--only 8% of the time. Given what we know of the importance of assumptions to accuracy of estimates and error rates, this in itself is alarming. There is no reason to believe that the situation is different in other social science disciplines.

What are Outliers and Fringeliers and why do we care about them?

Although definitions vary, an outlier is generally considered to be a data point that is far outside the norm for a variable or population (e.g., Jarrell, 1994; Rasm□ussen, 1988; Stevens, 1984). Hawkins described an outlier as an observation that "deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980, p.1). Outliers have also been defined as values that are "dubious in the eyes of the researcher" (Dixon, 1950, p. 488) and contaminants (Wainer, 1976).

Wainer (1976) also introduced the concept of the "fringelier," referring to "unusual events which occur more often than seldom" (p. 286). These points lie near three standard deviations from the mean and hence may have a disproportionately strong influence on parameter estimates, yet are not as obvious or easily identified as ordinary outliers due to their relative proximity to the distribution center. As fringeliers are a special case of outlier, for much of the rest of the paper we will use the generic term "outlier" to refer to any single data point of dubious origin or disproportionate influence.

Outliers can have deleterious effects on statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests. Second, if non-randomly distributed they can decrease normality (and in multivariate analyses, violate assumptions of sphericity and multivariate normality), altering the odds of making both Type I and Type II errors. Third, they can seriously bias or influence estimates that may be of substantive interest (for more information on these issues, see Rasmussen, 1988; Schwager & Margolin, 1982; Zimmerman, 1994).

Screening data for univariate, bivariate, and multivariate outliers is simple in these days of ubiquitous computing. The consequences of not doing so can be substantial.

□ *Author notes:* We would like to thank Amy Holt and Shannon Wildman for their assistance with various aspects of data analysis. Parts of this manuscript were developed while the first author was on the faculty of the University of Oklahoma. Correspondence relating to this manuscript should be addressed to Jason W. Osborne via email at jason_osborne@ncsu.edu.

What causes outliers and what should we do about them?

Outliers can arise from several different mechanisms or causes. Anscombe (1960) sorts outliers into two major categories: those arising from errors in the data, and those arising from the inherent variability of the data. Not all outliers are illegitimate contaminants, and not all illegitimate scores show up as outliers (Barnett & Lewis, 1994). It is therefore important to consider the range of causes that may be responsible for outliers in a given data set. What should be done about an outlying data point is at least partly a function of the inferred cause.

Outliers from data errors. Outliers are often caused by human error, such as errors in data collection, recording, or entry. Data from an interview can be recorded incorrectly, or miskeyed upon data entry. One survey the first author was involved with (reported in Brewer, Nauenberg, & Osborne, 1998) gathered data on nurses' hourly wages, which at that time averaged about $12.00 per hour with a standard deviation of about $2.00. In our data set one nurse had reported an hourly wage of $42,000.00. This figure represented a data collection error (specifically, a failure for the respondent to read the question carefully). Errors of this nature can often be corrected by returning to the original documents--or even the subjects if necessary and possible--and entering the correct value. In cases like that of the nurse who made $42,000 per hour, another option is available--recalculation or re-estimation of the correct answer. We had used anonymous surveys, but because the nature of the error was obvious, we were able to convert this nurse's salary to an hourly wage because we knew how many hours per week she worked and how many weeks per year she worked. Thus, if sufficient information is available, recalculation is a method of saving important data and eliminating an obvious outlier. If outliers of this nature cannot be corrected they should be eliminated as they do not represent valid population data points.

Outliers from intentional or motivated mis-reporting. There are times when participants purposefully report incorrect data to experimenters or surveyers. A participant may make a conscious effort to sabotage the research (Huck, 2000), or may be acting from other motives. Social desirability and self-presentation motives can be powerful. This can also happen for obvious reasons when data are sensitive (e.g., teenagers under-reporting drug or alcohol use, mis-reporting of sexual behavior). If all but a few teens under-report a behavior (for example, the frequency of sexual fantasies teenage males experience…), the few honest responses might appear to be outliers when in fact they are legitimate and valid scores. Motivated over-reporting can occur when the variable in question is socially desirable (e.g., income, educational attainment, grades, study time, church attendance, sexual experience).

Environmental conditions can motivate over-reporting or mis-reporting, such as if an attractive female researcher is interviewing male undergraduates about attitudes on gender equality in marriage. Depending on the details of the research, one of two things can happen: inflation of all estimates, or production of outliers. If all subjects respond the same way, the distribution will shift upward, not generally causing ouliers. However, if only a small subsample of the group responds this way to the experimenter, or if multiple researchers conduct interviews, then outliers can be created.

Outliers from sampling error. Another cause of outliers or fringeliers is sampling. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample. For example, in the previously described survey of nurse salaries, RNs who had moved into hospital administration were included in the database we sampled from, although we were particularly interested in floor nurses. In education, inadvertently sampling academically gifted or mentally retarded students is a possibility, and (depending on the goal of the study) might provide undesirable outliers. These cases should be removed as they do not reflect the target population.

Outliers from standardization failure. Outliers can be caused by research methodology, particularly if something anomalous happened during a particular subject's experience. One might argue that a study of stress levels in schoolchildren around the country might have found some significant outliers if it had been conducted during the fall of 2001 and included New York City schools. Researchers experience such challenges all the time. Unusual phenomena such as construction noise outside a research lab or an experimenter feeling particularly grouchy, or even events outside the context of the research lab, such as a student protest, a rape or murder on campus, observations in a classroom the day before a big holiday recess, and so on can produce outliers. Faulty or non-calibrated equipment is another common cause of outliers. These data can be legitimately discarded if the researchers are not interested in studying the particular phenomenon in question (e.g., if I were not interested in studying my subjects' reactions to construction noise outside the lab).

Outliers from faulty distributional assumptions. Incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers (e.g., Iglewicz & Hoaglin, 1993). Blood sugar levels, disciplinary referrals, scores on classroom tests where students are well-prepared, and self-reports of low-frequency behaviors (e.g., number of times a student has been suspended or held back a grade) may give rise to bimodal, skewed, asymptotic, or flat distributions, depending upon the sampling design. Similarly, the data may have a different structure than the researcher originally assumed, and long or short-term trends may affect the data in unanticipated ways. For example, a study of college library usage rates during the month of September may find outlying values at the beginning and end of the month, with exceptionally low rates at the beginning of the month when students have just returned to campus or are on break for Labor Day weekend (in the USA), and exceptionally high rates at the end of the month, when mid-term exams have begun. Depending upon the goal of the research, these extreme values may or may not represent an aspect of the inherent variability of the data, and may have a legitimate place in the data set.

Outliers as legitimate cases sampled from the correct population. Finally, it is possible that an outlier can come from the population being sampled legitimately through random chance. It is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails (Evans, 1999; Sachs, 1982). As a researcher casts a wider net and the data set becomes larger, the more the sample resembles the population from which it was drawn, and thus the likelihood of outlying values becomes greater.

In other words, there is only about a 1% chance you will get an outlying data point from a normally-distributed population; this means that, on average, *about 1% of your subjects should be 3 standard deviations from the mean.*

In the case that outliers occur as a function of the inherent variability of the data, opinions differ widely on what to do. Due to the deleterious effects on power, accuracy, and error rates that outliers and fringeliers can have, it might be desirable to use a transformation or recoding/truncation strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference (for more on transformations, see Osborne, 2002)

Outliers as potential focus of inquiry. We all know that interesting research is often as much a matter of serendipity as planning and inspiration. Outliers can represent a nuisance, error, or legitimate data. They can also be inspiration for inquiry. When researchers in Africa discovered that some women were living with HIV just fine for years and years, untreated, those rare cases are outliers compared to most untreated women, who die fairly rapidly. They could have been discarded as noise or error, but instead they serve as inspiration for inquiry: what makes these women different or unique, and what can we learn from them? In a study the first author was involved with, a teenager reported 100 *close* friends. Is it possible? Yes. Is it likely? Not generally, given any reasonable definition of "close friends." So this data point could represent either motivated mis-reporting, an error of data recording or entry (it wasn't), a protocol error reflecting a misunderstanding of the question, or something more interesting. This extreme score might shed light on an important principle or issue. Before discarding outliers, researchers need to consider whether those data contain valuable information that may not necessarily relate to the intended study, but has importance in a more global sense.

Identification of Outliers

There is as much controversy over what constitutes an outlier as whether to remove them or not. Simple rules of thumb (e.g., data points three or more standard deviations from the mean) are good starting points. Some researchers prefer visual inspection of the data. Others (e.g., Lornez, 1987) argue that outlier detection is merely a special case of the examination of data for influential data points.

Simple rules such as $z = 3$ are simple and relatively effective, although Miller (1991) and Van Selst and Jolicoeur (1994) demonstrated that this procedure (nonrecursive elimination of extreme scores) can produce problems with certain distributions (e.g., highly skewed distributions characteristic of response latency variables) particularly when the sample is relatively small. To help researchers deal with this issue, Van Selst and Jolicoeur (1994) present a table of suggested cutoff scores for researchers to use with varying sample sizes that will minimize these issues with extremely non-normal distributions. We tend to use a $z = 3$ guideline as an initial screening tool, and depending on the results of that screening, examine the data more closely and modify the outlier detection strategy accordingly.

Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices include Mahalanobis' distance and Cook's *D* are both frequently used to calculate the leverage that specific cases may exert on the predicted value of the regression line (Newton & Rudestam, 1999). Standardized or studentized residuals in regression can also be useful, and often the z=3 rule works well for residuals as well.

For ANOVA-type paradigms, most modern statistical software will produce a range of statistics, including standardized residuals. In ANOVA the biggest issue after screening for univariate outliers is the issue of within-cell outliers, or the distance of an individual from the subgroup. Standardized residuals represent the distance from the sub-group, and thus are effective in assisting analysts in examining data for multivariate outliers. Tabachnick and Fidell (2000) discuss data cleaning in the context of other analyses.

How to deal with outliers

There is a great deal of debate as to what to do with identified outliers. A thorough review of the various arguments is not possible here. We argue that what to do depends in large part on why an outlier is in the data in the first place. Where outliers are illegitimately included in the data, it is only common sense that those data points should be removed. (see also Barnett & Lewis, 1994). Few should disagree with that statement.

When the outlier is either a legitimate part of the data or the cause is unclear, the issue becomes murkier. Judd and McClelland (1989) make several strong points for removal even in these cases in order to get the most honest estimate of population parameters possible (see also Barnett & Lewis, 1994). However, not all researchers feel that way (see Orr, Sackett, & DuBois, 1991). This is a case where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions.

Keeping legitimate outliers and still not violating your assumptions. One means of accommodating outliers is the use of transformations (for a more thorough discussion of best practices in using data transformations, see Osborne, 2002). By using transformations, extreme scores can be kept in the data set, and the relative ranking of scores remains, yet the skew and error variance present in the variable(s) can be reduced (Hamilton, 1992).

However, transformations may not be appropriate for the model being tested, or may affect its interpretation in undesirable ways. Taking the log of a variable makes a distribution less skewed, but it also alters the relationship between the original variables in the model. For example, if the raw scores originally related to a meaningful scale, the transformed scores can be difficult to interpret (Newton & Rudestam, 1999; Osborne 2002). Also problematic is the fact that many commonly used transformations require non-negative data, which limits their applications. For this reason, many researchers turn to other methods to accommodate outlying values.

One alternative to transformation is truncation, wherein extreme scores are recoded to the highest (or lowest) reasonable score. For example, a researcher might decide that in reality, it is impossible for a teenager to have more than 15 close friends. Thus, all teens reporting more than this value (even 100) would be re-coded to 15. Through truncation the relative ordering of the data is maintained, and the highest or lowest scores remain the highest or lowest scores, yet the distributional problems are reduced.

Robust methods. Instead of transformations or truncation, researchers sometimes use various "robust" procedures to protect their data from being distorted by the presence of outliers. These techniques "accommodate the outliers at no serious inconvenience—or are *robust* against the presence of outliers" (Barnett & Lewis, 1994, p. 35). Certain parameter estimates, especially the mean and Least Squares estimations, are particularly vulnerable to outliers, or have "low breakdown" values. For this reason, researchers turn to robust or "high breakdown" methods to provide alternative estimates for these important aspects of the data.

A common robust estimation method for univariate distributions involves the use of a trimmed mean, which is calculated by temporarily eliminating extreme observations at both ends of the sample (Anscombe, 1960). Alternatively, researchers may choose to compute a Windsorized mean, for which the highest and lowest observations are temporarily censored, and replaced with adjacent values from the remaining data (Barnett & Lewis, 1994).

Assuming that the distribution of prediction errors is close to normal, several common robust regression techniques can help reduce the influence of outlying data points. The least trimmed squares (LTS) and the least median of squares (LMS) estimators are conceptually similar to the trimmed mean, helping to minimize the scatter of the prediction errors by eliminating a specific percentage of the largest positive and negative outliers (Rousseeuw & Leroy, 1987), while Windsorized regression smoothes the Y-data by replacing extreme residuals with the next closest value in the dataset (Lane, 2002).

Many options exist for analysis of non-ideal variables. In addition to the above-mentioned options, analysts can choose from non-parametric analyses, as these types of analyses have few if any distributional assumptions, although research by Zimmerman and others (e..g, Zimmerman, 1995) do point out that even non-parametric analyses suffer from outlier cases.

The effects of outlier removal

The rest of this paper is devoted to a demonstration of the effects of outliers and fringeliers on the accuracy of parameter estimates, and Type I and Type II error rates.

In order to simulate a real study where a researcher samples from a particular population, we defined our

Table 1
The effects of outliers on correlations

		Average	Average		% more	% errors	% errors	*t*
Population	*N*:	initial *r*	cleaned *r*	*t*	accurate	before	after	
r:						cleaning	cleaning	
r = -.06	52	.01	-.08	2.5**	95%	78%	8%	13.40***
	104	-.54	-.06	75.44***	100%	100%	6%	39.38***
	416	0	-.06	16.09***	70%	0%	21%	5.13***
r = .46	52	.27	.52	8.1***	89%	53%	0%	10.57***
	104	.15	.50	26.78***	90%	73%	0%	16.36***
	416	.30	.50	54.77***	95%	0%	0%	--

Note: 100 samples were drawn for each row. Outliers were actual members of the population who scored at least $z = 3$ on the relevant variable. With $N = 52$, a correlation of .274 is significant at $p < .05$. With $N = 104$, a correlation of .196 is significant at $p < .05$. With $N = 416$, a correlation of .098 is significant at $p < .05$, twotailed. ** $p < .01$, *** $p < .001$.

population as the 23,396 subjects in the data file from the National Education Longitudinal Study of 1988 produced by the National Center for Educational Statistics with complete data on all variables of interest. For the purposes of the analyses reported below, this population was sorted into two groups: “normal” individuals whose scores on relevant variables was between z =-3 and z = 3, and “outliers,” who scored at least z = 3 on one of the relevant variables.

In order to simulate the normal process of sampling from a population, but standardize the proportion of outliers in each sample, one hundred samples of N=50, N=100, and N=400 each were randomly sampled (with replacement between each sampling) from the population of “normal” subjects. Then an additional 4% were randomly selected from the separate pool of outliers bringing each sample to N=52, N=104, or N=416, respectively. This procedure produced samples that could easily have been drawn at random from the full population.

The following variables were calculated for each of the analyses below:

Accuracy was assessed by checking whether the original or cleaned correlation was closer to the population correlation. In these calculations the absolute difference was examined.

Error rates were calculated by comparing the outcome from a sample to the outcome from the population. If a particular sample yielded a different conclusion than was warranted by the population, that was considered an error of inference.

The effect of outliers on correlations. The first example looks at simple zero-order correlations. The goal was to see the effect of outliers on two different types of correlations: correlations close to zero (to demonstrate the effects of outliers on Type I error rates), and correlations that were moderately strong (to demonstrate the effects of outliers on Type II error rates). Toward this end, two different correlations were identified for study in the NELS data set: the correlation between locus of control and family size (r = -.06), and the correlation between composite achievement test scores and socioeconomic status (r = .46). Variable distributions were examined and found to be reasonably normal.

Correlations were then calculated in each sample, both before removal of outliers and after. For our purposes, r = -.06 was not significant at any of the sample sizes, and r = .46 was significant at all sample sizes. Thus, if a sample correlation led to a decision that deviated from the “correct” state of affairs, it was considered an error or inference.

As Table 1 demonstrates, outliers had adverse effects upon correlations. In all cases, removal of the outliers had significant effects upon the magnitude of the correlations, and the cleaned correlations were more accurate (i.e., closer to the known population correlation) 70 - 100% of the time. Further, in most cases the incidence of errors of inference was lower with cleaned than uncleaned data.

The effect of outliers on t-tests and ANOVAs. The second example deals with analyses that look at group mean differences, such as t-tests and ANOVA. For the purpose of simplicity, these analyses are simple t-tests, but these results would generalize to any ANOVA. For these analyses two different conditions were examined: when there were no significant differences between the groups in the population (sex differences in socioeconomic status (SES) produced a mean group difference of 0.0007 with a SD of 0.80 and with 24501 *df* produced a t of 0.29), and when there were significant group differences in the population (sex differences in mathematics achievement test scores produced a mean difference of 4.06 and SD of 9.75 and 24501 *df* produced a t of 10.69, $p < .0001$). For both variables the effects of having outliers in only one cell as compared to both cells were examined. Distributions for both dependent variables were examined and found to be reasonably normal.

For these analyses, t-tests were calculated in each sample, both before removal of outliers, and after. For our purposes, t-tests looking at SES should not produce significant group differences, whereas t-tests looking at mathematics achievement test scores should. Two different issues were examined: mean group differences and the magnitude of the *t*. If an analysis from a sample led to a different conclusion it was considered an error.

The results in Table 2 illustrate the effects of outliers on t-tests and ANOVAs. Removal of outliers produced a significant change in the mean differences between the two groups when the groups were equal in the population, but tended not to when there were strong group differences. Removal of outliers produced significant change in the *t* statistics primarily when there were strong group differences. In both cases the tendency was for both group differences and *t* statistics to become more accurate in a majority of the samples. Interestingly, there was little evidence that outliers produced Type I errors when group means were equal, and thus removal had little discernable effect. But when there were strong group differences, outlier removal tended to have a significant beneficial effect on error rates, although not as substantial an effect as seen in the correlation analyses.

The presence of outliers in one or both cells, surprisingly, failed to produce any differential effects. The expectation had been that the presence of outliers in a single cell would increase the incidence of Type I errors.

Why this effect was not shown could have to do with the type of outliers in these analyses, or other factors, such as the absolute equality of the two groups on SES, which may not reflect the situation most researchers face.

To remove, or not to remove?

Although some authors argue that removal of extreme scores produces undesirable outcomes, they are in the minority, especially when the outliers are illegitimate. When the data points are suspected of being legitimate, some authors (e.g., Orr, Sackett, & DuBois, 1991) argue that data are more likely to be representative of the population as a whole if outliers are not removed.

Conceptually, there are strong arguments for removal or alteration of outliers. The analyses reported in this paper also empirically demonstrate the benefits of outlier removal. Both correlations and t-tests tended to show significant changes in statistics as a function of removal of outliers, and in the overwhelming majority of analyses accuracy of estimates were enhanced. In most cases errors of inference were significantly reduced, a prime argument for screening and removal of outliers.

Although these were two fairly simple statistical procedures, it is straightforward to argue that the benefits of data cleaning extend to simple and multiple regression, and to different types of ANOVA procedures. There are other procedures outside these, but the majority of social science research utilizes one of these procedures. Other research (e.g., Zimmerman, 1995) has dealt with the effects of extreme scores in less commonly-used procedures, such as nonparametric analyses.

References

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics, 2,* 123-147.

Barnett, V, & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Brewer, C. S., Nauenberg, E., & Osborne, J. W. (1998, June). *Differences among hospital and non-hospital RNs participation, satisfaction, and organizational commitment in western New York.* Paper presented at the National meeting of the Association for Health Service Research, Washington DC.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics, 21,* 488-506.

Evans, V.P. (1999). Strategies for detecting outliers in regression analysis: An introductory primer. In B. Thompson (Ed.), *Advances in social science methodology*: (Vol. 5, pp. 213-233). Stamford, CT.: JAI Press.

Hamilton, L.C. (1992). *Regressions with graphics: A second course in applied statistics.* Monterey, CA.: Brooks/Cole.

Hawkins, D.M. (1980). *Identification of outliers.* London: Chapman and Hall.

Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Longman.

Iglewicz, B., & Hoaglin, D.C. (1993). *How to detect and handle outliers.* Milwaukee, WI.: ASQC Quality Press.

Jarrell, M. G. (1994). A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. *Research in the schools, 1,* 49-58.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA.: Harcourt Brace Jovanovich.

Lane, K. (2002, February). *What is robust regression and how do you do it?* Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Lornez, F. O. (1987). Teaching about influence in simple regression. *Teaching Sociology, 15*(2), 173-177.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology, 43*(4), 907-912.

Newton, R.R., & Rudestam, K.E. (1999). *Your statistical consultant: Answers to your data analysis questions.* Thousand Oaks, CA.: Sage.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation., 8*, Available online at http://ericae.net/pare/getvn.asp?v=8&n=6.

Osborne, J. W., Christiansen, W. R. I., & Gunter, J. S. (2001). *Educational psychology from a statistician's perspective: A review of the quantitative quality of our field.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research, 23*(2), 189-202.

Rousseeuw, P., & Leroy, A. (1987). *Robust regression and outlier detection*. New York: Wiley.

Sachs, L. (1982). *Applied statistics: A handbook of techniques* (2nd ed). New York: Springer-Verlag.

Schwager, S. J., & Margolin, B. H. (1982). Detection of multivariate outliers. *The annals of statistics, 10*, 943-954.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334-344.

Tabachnick, B.G., & Fidell, L. S. (2000). Using multivariate statistics, 4th edition. Pearson Allyn & Bacon.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics, 1*(4), 285-312.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, 121*(4), 391-401.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64*(1), 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*(1), 55-68.

Table 2
The effects of outliers on t-tests

	N	Initial mean difference	Cleaned mean difference	*t*	% more accurate mean difference	Avg Initial *t*	Average Cleaned *t*	*t*	% Type I or II errors before cleaning	% Ty or erro aft clean
OUTLIERS										
Equal group means, outliers in one cell	52	0.34	0.18	3.70***	66.0%	-0.20	-0.12	1.02	2.0%	1.0
	104	0.22	0.14	5.36***	67.0%	0.05	-0.08	1.27	3.0%	3.0
	416	0.09	0.06	4.15***	61.0%	0.14	0.05	0.98	2.0%	3.0
Equal group means, outliers in both cells	52	0.27	0.19	3.21***	53.0%	0.08	-0.02	1.15	2.0%	4.0
	104	0.20	0.14	3.98***	54.0%	0.02	-0.07	0.93	3.0%	3.0
	416	0.15	0.11	2.28*	68.0%	0.26	0.09	2.14*	3.0%	2.0
Unequal group means, outliers in one cell	52	4.72	4.25	1.64	52.0%	0.99	1.44	-4.70***	82.0%	72
	104	4.11	4.03	0.42	57.0%	1.61	2.06	-2.78**	68.0%	45
	416	4.11	4.21	-0.30	62.0%	2.98	3.91	-12.97***	16.0%	0.0
Unequal group means, outliers in both cells	52	4.51	4.09	1.67	56.0%	1.01	1.36	-4.57***	81.0%	75
	104	4.15	4.08	0.36	51.0%	1.43	2.01	-7.44***	71.0%	47
	416	4.17	4.07	1.16	61.0%	3.06	4.12	-17.55***	10.0%	0.0

Note: 100 samples were drawn for each row. Outliers were actual members of the population who scored at least $z = 3$ on the relevant variable.
* $p < .05$, ** $p < .01$, *** $p < .001$