Best practices in data cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.

Book · January 2013

Jason W Osborne
Clemson University

Data Cleaning Basics: Best Practices in Dealing with Extreme Scores

Jason W. Osborne, PhD

In quantitative research, it is critical to perform data cleaning to ensure that the conclusions drawn from the data are as generalizable as possible, yet few researchers report doing so (Osborne JW. *Educ Psychol*. 2008;28:1-10). Extreme scores are a significant threat to the validity and generalizability of the results. In this article, I argue that researchers need to examine extreme scores to determine which of many possible causes contributed to the extreme score. From this, researchers can take appropriate action, which has many laudatory effects, from reducing error variance and improving the accuracy of parameter estimates to reducing the probability of errors of inference.

Keywords: Data cleaning; Extreme scores; Outliers; Parameter estimates

Most authors of peer-reviewed journal articles go to great lengths to describe their study, the research methods, the sample, the statistical analyses used, results, and conclusions based on those results. However, few seem to mention data cleaning (which can include screening for extreme scores, missing data, normality, etc). To be sure, some of the researchers do check their data for these things (and may neglect to report having done that), but Osborne[1] examined 2 years' worth of empirical articles in top-tier *Educational Psychology* journals, none explicitly discussed any data cleaning. There is no reason to believe that the situation is different in other disciplines.

The goal of this article is to discuss the issue of extreme scores, which can dramatically increase risk for errors of inference, problems with generalizability (biased estimates), and suboptimal power (some “robust” procedures and nonparametric tests are incorrectly considered to be immune from these sorts of issues; however, even robust and nonparametric tests benefit from clean data[2,3]).

The goal of this article is to highlight why it is critical to screen data for extreme scores and specific suggestions for how to deal with them.

What Are Extreme Scores and Why Do We Care About Them?

An extreme score, or data point far outside the normal distribution for a variable or population,[4-6] is also described as an observation that “deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.”[7] Arguably, if an extreme score has origins in a different mechanism or population, it does not belong in your analysis. Outliers have also been defined as values that are “dubious in the eyes of the researcher”[8] and contaminants,[9] all of which lead to the same conclusion.

So Why Do We Care About Extreme Values?

Extreme values can cause serious problems for statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests. Second, if nonrandomly distributed, they can substantially alter 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 because they may not be generated by the population of interest.[2,5,10]

What Is an Extreme Score?

There is as much controversy over what constitutes an extreme score as whether to remove them or not. It is always a good idea to visually inspect data before any other analysis. Simple rules of thumb (eg, data points 3 or more SDs from the mean) are good starting points, unless the sample is particularly small.[11,12] I recommend examining scores at or beyond 3 SDs from the mean, as in a normally distributed population, the probability of an individual being more than 3 SDs from the mean by random chance alone is 0.26%. Because of this, we have a strong basis for suspecting data points beyond ±3 SD from the mean are *not generated by the population of interest* and as such should be dealt with in some fashion.

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 (influence) that specific cases may exert on the predicted value of the regression line.[13] Standardized or studentized residuals in regression and analysis of variance (ANOVA)-type analyses can also help

From North Carolina State University.

Address correspondence to Jason W. Osborne, PhD, North Carolina State University, Curriculum and Instruction and Counselor Education, Poe 602c, Campus Box 7801, NCSU, Raleigh, NC 27695-7801. E-mail: jason_osborne@ncsu.edu.

1527-3369/09/1001-0343$36.00/0

doi:10.1053/j.nainr.2009.12.009

identify within-group outliers. The $z = \pm 3$ rule works well for standardized residuals as well.

What Causes Extreme Scores and What Should We Do About Them?

Extreme scores can arise from (at least) six possible reasons for data points that may be suspect. First note that not all extreme scores are illegitimate contaminants, and not all illegitimate scores show up as extreme scores.[14] It is therefore important to consider the range of causes that may be responsible for extreme scores. Inferred cause can then inform what action a researcher should take with a given extreme score.

Extreme Scores From Data Errors

Extreme scores are often caused by errors in data collection, recording, or entry. Data from an interview or survey can be recorded incorrectly, or mis-keyed upon data entry (eg, a survey respondent reporting *yearly* wage rather than *hourly* wage). Errors of this nature can often be corrected by returning to the original documents, recalculating or inferring the correct response, or recontacting the original participant. This can save important data and eliminate an problematic extreme score.

Extreme Scores From Intentional or Motivated Misreporting

Motivated misreporting by research participants is a long-discussed source of bias in data. A participant may make a conscious effort to sabotage the research,[15] or may be acting from social desirability or self-presentation motives. Identifying and reducing this issue is difficult, unless researchers take care to triangulate or validate data in some manner. Osborne and Blanchard[16] summarizes several approaches to identifying response sets such as this. If you suspect motivated mis-responding in your data, you should probably remove that participant, because the data are being influenced by more than the phenomena you wish to examine.

Extreme Scores From Sampling Error or Bias

No sampling framework is perfect, and sampling error or bias can produce extreme scores by erroneously including individuals from populations not intended to be sampled.

For example, some colleagues and I[17] randomly sampled registered nurses from licensure rolls for a survey on organizational commitment. As part of this survey, we asked nurses to report their salary. Upon examining some very extreme scores, we discovered we had inadvertently surveyed some registered nurses who had moved into hospital administration (with a much higher salary) but who had also maintained their nursing license. These cases, being extreme and not of the population of interest (floor nurses) were removed.

Extreme Scores From Standardization Failure

Unexpectedly, extreme scores can be caused by research methodology, particularly if something anomalous happened during a particular subject's experience. Unusual phenomena such as construction noise outside a research laboratory or an experimenter feeling particularly grouchy, or even events outside the context of the research laboratory, 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 noncalibrated equipment is another common cause of extreme scores.

Let us consider two possible cases in relation to this source of outliers. In the first case, we might have a piece of equipment in our laboratory that was miscalibrated, yielding measurements that were extremely different from other days' measurements. If the miscalibration results in a fixed change to the score that is consistent or predictable across all measurements (eg, all measurements are off by 100), then adjustment of the scores is appropriate. If there is no clear way to defensibly adjust the measurements, they must be discarded.

Extreme Scores From Faulty Distributional Assumptions

Incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers.[18] Blood sugar levels, disciplinary referrals, scores on classroom tests where students are well-prepared, and self-reports of low-frequency behaviors (eg, number of times a student has been suspended or held back a grade) may give rise to highly nonnormal distributions. These distributions may look like they have a substantial number of extreme scores, but after transformation (s) to improve normality,[19] it might be the case that few, if any, of the data points are subsequently identified as outliers.

The data presented in Fig 1 on 180 students taking an examination in an undergraduate psychology class shows a highly skewed distribution with a mean of 87.50 and an SD of 8.78. Although one could argue that the lowest scores on this test are outliers because they are more than 3 SDs below the

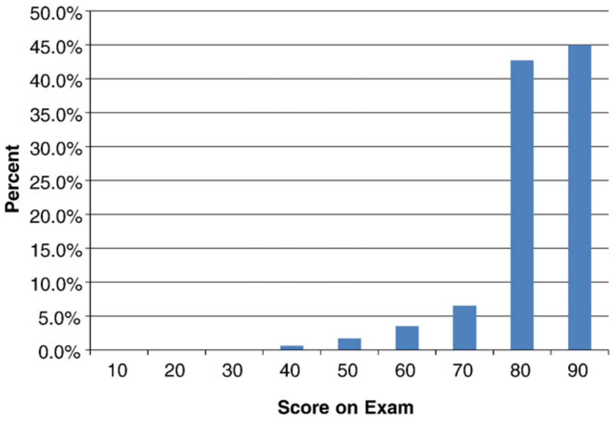

Fig 1. Performance on class unit examination, Undergraduate Education Psychology Course.

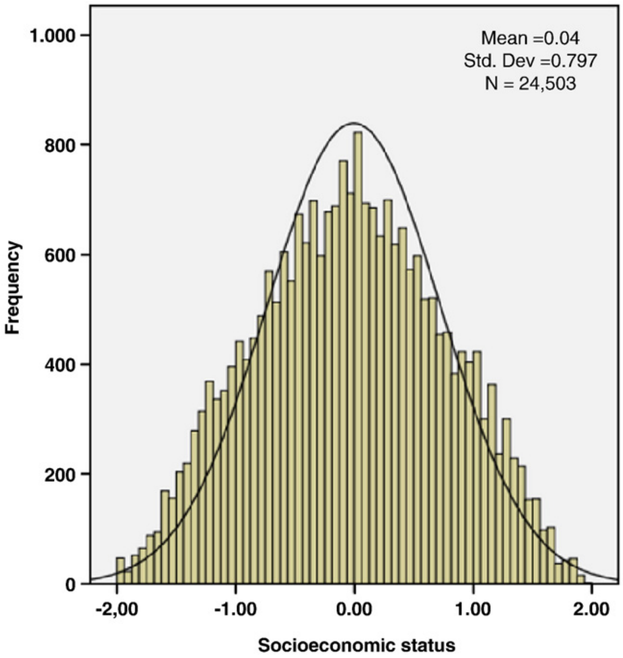

Fig 2. Distribution of SES.

mean, a better interpretation is that the data are not normally distributed. In this case, a transformation should be used to normalize the data before analysis of extreme scores should occur or analyses appropriate for nonnormal distributions should be used.

Extreme Scores 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.[20,21] 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 legitimate individual outlying values becomes greater, although as a percentage of the sample, they become less significant overall.

When extreme scores occur as a function of the inherent variability of the data, opinions differ widely on what to do. When legitimate extreme scores are in a data set, they can have deleterious effects on power, accuracy, and type I/II error rates. One way to deal with them is to use *truncation*, in which you specify an upper reasonable limit to your data and recode higher scores to that number (eg, in a study of adolescents one indicated he had 99 *close* friends, yet by our definition that would be impossible; thus, we recoded all responses above 15 (the highest reasonable number of close friends) to 15). This keeps all data in the sample while at the same time reducing the influence of these scores. Data transformations (eg, square root and log) also have the effect of reducing the effect of extreme scores when used appropriately.[19]

Alternatively, extreme scores can present an opportunity for inquiry. When researchers in Africa discovered some women who had been repeatedly exposed to human immunodeficiency virus over several years but remained uninfected,[22] they represent potential for an important advance in understanding. Thus, before discarding outliers, researchers need to consider whether those data contain valuable information that may not necessarily relate to the intended study but have importance in a more global sense.

To be clear on this point, no matter the inferred cause of the extreme score, it must be dealt with in some fashion and that decision should be reported and defended in any research reports that involve the data. Extreme scores should be corrected, removed, truncated, reduced in importance through data transformation, or separated from the rest of the sample for separate study. This affords the most replicable, honest estimate of the population parameters possible.[23,24] Not only are basic parameter estimates closer to population values when illegitimate extreme values are removed, but inferential statistics (correlations, *t* tests, etc) have substantially lower error rate.[24]

Advanced Techniques for Dealing With Extreme Scores: 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. Certain parameter estimates, especially the mean and least squares estimations, are particularly vulnerable to outliers, or have "low breakdown"

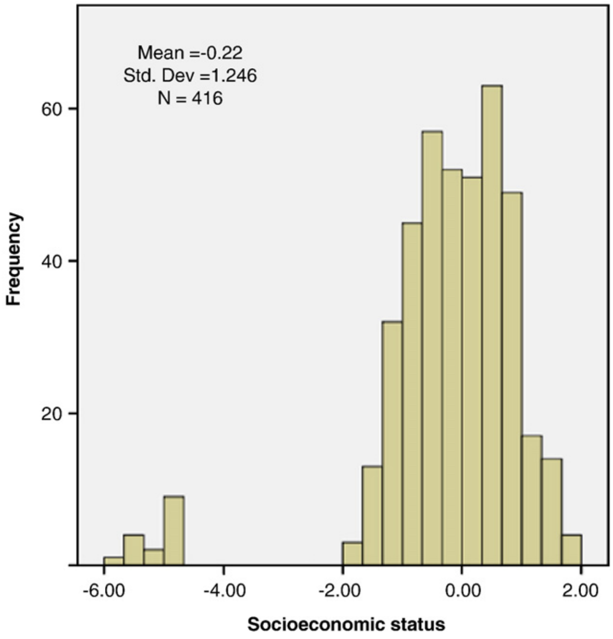

Fig 3. Distribution of SES with 4% outliers.

Table 1. The effects of outliers on correlations

		Average	Average		% More	% Errors	% Errors after	
Population, (r)	N	initial r	cleaned r	t	accurate	before cleaning	cleaning	T
−0.06	52	0.01	−0.08	2.5*	95	78	8	13.40†
	104	−0.54	−0.06	75.44†	100	100	6%	39.38†
	416	0	−0.06	16.09†	70	0	21	5.13†
0.46	52	0.27	0.52	8.1†	89	53	0	10.57†
	104	0.15	0.50	26.78†	90	73	0	16.36†
	416	0.30	0.50	54.77†	95	0	0	–

One hundred samples were randomly 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 0.274 is significant at $P < .05$. With n = 104, a correlation of 0.196 is significant at $P < .05$. With n = 416, a correlation of 0.098 is significant at $P < .05$, two tailed.

*$P < .01$.

†$P < .001$.

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.[25] 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.[14]

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 and the least median of squares 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,[26] whereas Windsorized regression smoothes the Y-data by replacing extreme residuals with the next closest value in the dataset.[27]

Many options exist for analysis of nonideal variables. In addition to the abovementioned options, analysts can choose from nonparametric analyses, because these types of analyses have few if any distributional assumptions, although research by Zimmerman[3,28] do point out that even nonparametric analyses suffer from outlier cases.

The Effects of Extreme Scores and Their Removal on Individual Variables

Extreme scores have several specific effects on variables that are otherwise normally distributed. To illustrate this, we will use *socioeconomic status* (SES)* that represents a composite of family income and social status based on parent occupation. In this data set, the scores were transformed to z scores. This variable shows strong normality, with a skew of −0.001 (0.00 is perfectly symmetrical; as depicted in Fig 2).

Samples from this distribution should also share these distributional traits, especially large samples. For example, a relatively large sample of n = 416 that included 4% extreme scores on one side of the distribution (high-poverty students), the distribution properties changed substantially (as depicted in Fig 3):

The skew is now −2.18. Substantial error has been added to the variable (SD is increased 56%), and it is clear that those 16 students at the very bottom of the distribution do not belong to the normal population of interest. Removal of these outliers returned the distribution to a mean of −0.02, SD = 0.78, skew = 0.01, not significantly different from the original population of over 24000.

Osborne and Overbay[24] performed similar simulations of the effects of small numbers of outliers on repeated samples from a known population in the context of correlation and ANOVA-type analyses. The effects were striking.

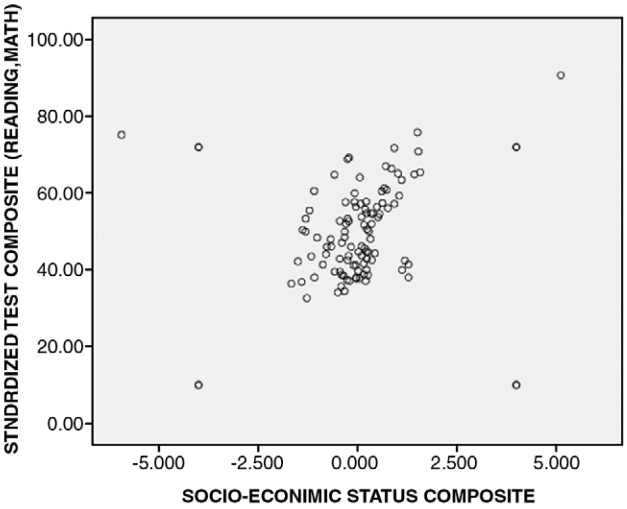

Fig 4. Correlation of SES and achievement, with 4% outliers.

* From the National Centers for Educational Statistics NELS 88 data set.

Table 2. The effects of outliers on *t* tests

Outliers	n	Initial mean difference	Cleaned mean difference	*t*	% more accurate mean difference	Average initial *t*	Average cleaned *t*	*t*	% Type I or II errors before cleaning	% Type I or II errors after cleaning	*t*
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	<1
	104	0.22	0.14	5.36 ‡	67.0	0.05	−0.08	1.27	3.0	3.0	<1
	416	0.09	0.06	4.15 ‡	61.0	0.14	0.05	0.98	2.0	3.0	<1
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	<1
	104	0.20	0.14	3.98 ‡	54.0	0.02	−0.07	0.93	3.0	3.0	<1
	416	0.15	0.11	2.28 *	68.0	0.26	0.09	2.14 *	3.0	2.0	<1
Unequal group means, outliers in one cell	52	4.72	4.25	1.64	52.0	0.99	1.44	−4.70 ‡	82.0	72.0	2.41 †
	104	4.11	4.03	0.42	57.0	1.61	2.06	−2.78 †	68.0	45.0	4.70 ‡
	416	4.11	4.21	−0.30	62.0	2.98	3.91	−12.97 ‡	16.0	0.0	4.34 ‡
Unequal group means, outliers in both cells	52	4.51	4.09	1.67	56.0	1.01	1.36	−4.57 ‡	81.0	75.0	1.37
	104	4.15	4.08	0.36	51.0	1.43	2.01	−7.44 ‡	71.0	47.0	5.06 ‡
	416	4.17	4.07	1.16	61.0	3.06	4.12	−17.55 ‡	10.0	0.0	3.13 ‡

One hundred samples were drawn for each row. Outliers were actual members of the population who scored at least $z = \pm 3$ on the relevant variable.

**P* < .05.

†*P* < .01.

‡*P* < .001.

The Effect of Extreme Scores on Correlations and Regression

As Table 1 demonstrates, outliers had adverse effects on correlations. Removal of the outliers produced more accurate (ie, closer to the known "population" correlation) estimates of the population correlation 70% to 100% of the time. Furthermore, in most cases errors of inference were significantly less common (between 89.7%–100% of errors of inference were eliminated for all but the largest data sets, which had few errors of inference). with cleaned than uncleaned data.

As Fig 4 shows, a few randomly chosen outliers in a sample of 100 can cause substantial mis-estimation of the population correlation. In the sample of almost 24 000 students, these two variables were correlated very strongly, r = 0.46. In this particular example, the correlation with 4% outliers in the analysis was r = 0.16 and was not significant, whereas after removal of the extreme scores, the correlation closely estimated the expected magnitude (r = 0.48).

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 easily generalize to more complex analyses such as ANOVA. For these analyses, two different conditions were examined: when there were no significant differences between the groups in the population (sex differences in SES produced a mean group difference of 0.0007 with an SD of 0.80 and with 24 501 *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 an SD of 9.75 and 24 501 *df* produced a *t* of 10.69, $P < .0001$).

The results in Table 2 again illustrate the expected effects of outliers on *t* test analyses designs. Removal of outliers had beneficial effects, in that the results tended to become more like the population: for both groups, differences and *t* statistics became more accurate in most the samples.

Missing Data as a Special Case of Extreme Score

Missing data can be thought of as another potential type of extreme score. As such, much of the previous discussion applies. There are multiple reasons why data might be missing, and it is important to attempt to ascertain the reason for missingness, just as extremeness. Cole gives a much more thorough treatment of how to analyze and deal with missing data, for those interested.[29]

However, one underutilized technique is analyzing differences between those with missing data and those with complete data. For example, researchers can code a variable that represents which category each subject falls into and then can analyze other data as a function of missingness to determine if missingness is associated with particular subgroups or other variables. This can shed important light onto whether missingness can be causing significant bias.

Summary

In sum, the best, most sophisticated analyses must be considered flawed if quantitative researchers do not take the time to thoroughly understand and examine their data to ensure the best possible outcome (ie, the most accurate, generalizable representation of the population). Although over a century of writings on quantitative methods has yielded a very diverse set of opinions about this topic, analyses and principles summarized herein should convince the readers that it is in their best interest to thoroughly clean their data before analysis.

References

1. Osborne JW. Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educ Psychol.* 2008;28: 1-10.
2. Zimmerman DW. A note on the influence of outliers on parametric and nonparametric tests. *J Gen Psychol.* 1994; 121:391-401.
3. Zimmerman DW. Increasing the power of nonparametric tests by detecting and downweighting outliers. *J Exper Educ.* 1995;64:71-78.
4. Jarrell MG. A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. *Res Sch.* 1994;1:49-58.
5. Rasmussen JL. Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behav Res.* 1988;23:189-202.
6. Stevens JP. Outliers and influential data points in regression analysis. *Psychol Bull.* 1984;95:334-344.
7. Hawkins DM. Identification of Outliers. New York: Chapman and Hall; 1980.
8. Dixon WJ. Analysis of extreme values. *Ann Math Stat.* 1950; 21:488-506.
9. Wainer H. Robust statistics: a survey and some prescriptions. *J Educ Stat.* 1976;1:285-312.
10. Schwager SJ, Margolin BH. Detection of multivariate outliers. *Ann Stat.* 1982;10:943-954.
11. Miller J. Reaction time analysis with outlier exclusion: bias varies with sample size. *Q J Exp Psychol.* 1991;43:907-912.
12. Van Selst M, Jolicoeur P. A solution to the effect of sample size on outlier elimination. *Q J Exp Psychol.* 1994;47: 631-650.
13. Newton RR, Rudestam KE. Your Statistical Consultant: Answers to Your Data Analysis Questions. Thousand Oaks, CA: Sage.; 1999.

14. Barnett V, Lewis T. Outliers in Statistical Data. New York: Wiley; 1994.

15. Huck SW, Sutton CO. Some comments concerning the use of monotonic transformations to remove the interaction in two-factor ANOVA's. *Educ Psychol Meas.* 1975;35:789-791.

16. Osborne JW, Blanchard MR. Random responding from students is a threat to the validity of educational research results. *Educational Psychology*. in press.

17. Brewer CS, Nauenberg E, Osborne JW. Differences among hospital and non-hospital RNs participation, satisfacton, and organizational committment in western New York. Paper presented at: National meeting of the Association for Health Service Research; June, 1998; Washington DC; 1998.

18. Iglewicz B, Hoaglin DC. How to Detect and Handle Outliers. Wilwaukee, WI: ASQC Quality Press; 1993.

19. Osborne JW. Notes on the use of data transformations. Practical assessment, research, and evaluation; 2002. p. 8. Available online at http://ericae.net/pare/getvn.asp?v=8&n=6.

20. Evans VP. Strategies for detecting outliers in regression analysis: an introductory primer. In: Thompson B, editor. Advances in Social Science Methodology, Vol. 5. Stamford, CT: JAI Press.; 1999. p. 213-233.

21. Sachs L. Applied Statistics: A Handbook of Techniques. 2nd ed. New York: Springer-Verlag; 1982.

22. Rowland-Jones S, Sutton J, Ariyoshi K, et al. HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women. *Nat Med.* 1995;1:59-64.

23. Judd CM, McClelland GH. Data analysis: A Model Comparison Approach. San Diego, CA: Harcourt Brace Jovanovich; 1989.

24. Osborne JW, Overbay A. The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research, and Evaluation; 2004. p. 9.

25. Anscome FJ. Rejection of outliers. *Technometrics.* 1960;2: 123-147.

26. Rousseeuw P, Leroy A. Robust Regression and Outlier Detection. New york: Wiley; 1987.

27. Lane K. What Is Robust Regression and How Do You Do It? Annual meeting of the southwest educational research association. Austin, TX; 2002.

28. Zimmerman DW. Invalidation of parametric and nonparamteric statistical tests by concurrent violation of two assumptions. *J Exp Educ.* 1998;67:55-68.

29. Cole JC. How to deal with missing data. In: Osborne JW, editor. Best Practices in Quantitative Methods. Thousand Oaks, CA: Sage Publishing; 2008.