Applied Linguistics: 31/3: 368–390

doi:10.1093/applin/amp038 Advance Access published on 16 October 2009

Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics

JENIFER LARSON-HALL and RICHARD HERRINGTON

University of North Texas

Move2 step1 In this article we introduce language acquisition researchers to two broad areas of applied statistics that can improve the way data are analyzed. First we argue that visual summaries of information are as vital as numerical ones, and suggest ways to improve them. Specifically, we recommend choosing boxplots over barplots and adding locally weighted smooth lines (Loess lines) to scatterplots. Second, we introduce the reader to robust statistics, a tool that can provide a way to use the power of parametric statistics without having to rely on the assumption of a normal distribution; robust statistics incorporate advances made in applied statistics in the last 40 years. Such types of analyses have only recently become feasible for the non-statistician practitioner as the methods are computer-intensive. Move3 step1 We acquaint the reader with trimmed means and bootstrapping, procedures from the robust statistics arsenal which are used to make data more robust to deviations from normality. We show examples of how analyses can change when robust statistics are used. Move2 step4 Robust statistics have been shown to be nearly as powerful and accurate as parametric statistics when data are normally distributed, and many times more powerful and accurate when data are non-normal.

INTRODUCTION

Statistics play an important role in analyzing data in all fields that employ empirical and quantitative methods, including the second language acquisition (SLA) field. This article is meant to address issues that are pertinent to the field of SLA, given our own constraints and parameters. For example, one statistical problem that we probably cannot avoid is the lack of truly random selection in experimental design, which Porte (2002) has noted. Given the populations we try to test and issues of validity versus reliability (do we use intact classrooms and get 'real' data, or use laboratory tests that can randomize better and get more 'reliable' data?) there is no simple way to always use true randomization in populations we test. However, there are other statistical issues in SLA that are amenable to improvement. For example, many SLA research designs use small sample sizes (generally less than 20 per group), meaning that the statistical power of a test of a normal distribution may be low (making it hard to reliably test whether data is normally distributed or not),

yet these studies use parametric statistics which assume a normal distribution. Another problem with any size group is reliably identifying outliers.

In this article we will put forward two broad types of techniques which researchers can use to improve the quality of their statistical analyses. The first suggestion is to use graphic techniques that are the most helpful in understanding data distributions in order to assess statistical relationships and differences between groups. The second suggestion is that researchers learn about and begin to incorporate statistics into their statistical analyses that are robust (or in other words, insensitive to) violations of assumptions of a normal distribution.

GRAPHICS

Introduction

Because doing a statistical analysis is as much an art as a science (Westfall and Young 1993: 20), researchers need to provide as much information about their data as possible to their reading audience.[1] The best kinds of visual information can help readers verify the assumptions about the data and the numerical results that are presented in the text and provide intuitions about relationships or group differences. The American Psychological Association (APA) Task Force on Statistical Information (Wilkinson 1999) recommends always including visual data when reporting on statistics.

Tufte (2001) claims that improving the resolution of our graphics by providing as much information as possible may lead to improvements in the science we perform. At present, most published articles in the field of SLA, if they present graphics, show a barplot if the data are distributed into groups, and a scatterplot if the data involves relationships between variables. We suggest that these graphics be improved by using boxplots instead of barplots for group-difference data and adding Loess lines to scatterplots for relational data.

Boxplots instead of barplots

Barplots are popular in the SLA field. In the five years of papers published in *Applied Linguistics*, *Language Learning* and *Studies in Second Language Acquisition* from 2003 to 2007 that we examined, 110 studies contained group difference quantitative data that could have been represented with boxplots. However, of those 110 studies, only one used a boxplot, while 46 used barplots. An additional 12 used line graphs (the remainder did not provide graphics). A novice to the field would assume that barplots were the graphic of choice for SLA researchers, and continue to follow this tradition. However, barplots (and line graphs) are far less informative than boxplots, providing only one or two points of data (depending on whether error bars are used) compared with the five or more points that boxplots provide. While both types of plots may be somewhat impoverished by Tufte's (2001) standards, boxplots

Table 1: A comparison of the information used to create the boxplot versus the barplot for the 'Late' group in Figure 1

	Boxplot	Barplot
Mean	–	3.10
First quartile	2.3	–
Median (second quartile)	2.9	–
Third quartile	3.8	–
Minimum score	1.6	–
Maximum score	4.9	–
Outliers labeled	Yes	No

should always be preferred over barplots unless the data are strictly frequency data, such as the number of times that one teacher uses recasts out of the total number of instances of negative evidence.[2] In fact, one reviewer of this article lauded the recommendation to use boxplots over barplots and said, 'If we had a contest on which graphical method conveys the least amount of information and has the best potential to mislead, barplots would win easily'. Table 1 shows the information that is used to calculate both types of graphics that are shown in Figure 1. Table 1 clearly shows how impoverished the data used in the barplot is.

Figure 1 gives an example of a barplot and a boxplot of the same data, compared side by side.

Notice that the data look different in the two kinds of graphics. The boxplot provides far more information about the *distribution* of scores than the barplot. One of the advantages of the boxplot (invented by Tukey, 1977) is that it is helpful in interpreting the differences between sample groups without making any assumptions regarding the underlying probability distribution, but at the same time indicating the degree of dispersion, skewness, and outliers in the given data set. For example, in looking at the boxplot in Figure 1 (the graph on the right) we notice that the range of scores is wide for the non-native speakers (as indicated by the length of the whiskers on either side of the box for the 'Non', 'Late', and 'Early' labels), but quite narrow for the native speakers (NS). We can also note an outlier in the NS scores. Boxplots are robust to outliers but barplots may change considerably if only one data point is added or removed. Lastly, we could note that the data for the NS is *not* symmetric, since there is only a lower whisker but no upper whisker. This means the distribution is skewed. The other distributions in Figure 1 are slightly skewed as well, as their medians are not perfectly in the center of the boxes and/or the boxes are not perfectly centered on the whiskers.

Because many readers may not be familiar with boxplots, Figure 1 labels the parts of the boxplot (which is notched in this case, although it doesn't have to be). While a barplot shows the mean score, the line in the middle of the

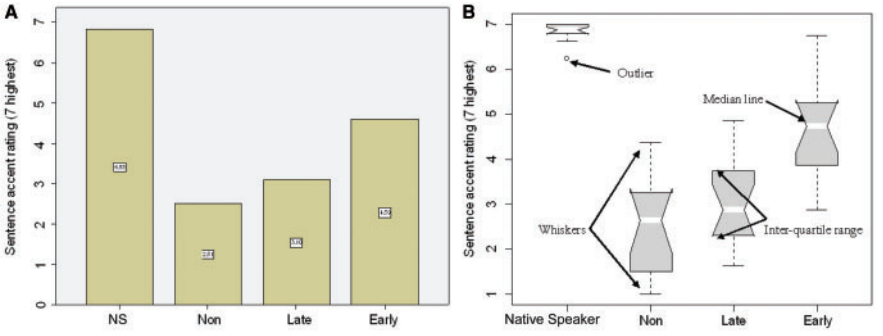

Figure 1: Comparison between a barplot (A) and a boxplot (B) of the same data

boxplot (here, in white) shows the median point. The length of the box contains all of the points that comprise the 25th to 75th percentile of scores (in other words, the first to third quartiles), and this is called the interquartile range (IQR). The ends of the box are called the hinges of the box. The whiskers of the boxplot extend out to the minimum and maximum scores of the distribution, unless these points are distant from the box. If the points extend more than 1.5 times the IQR above or below the box, they are indicated with a circle as outliers (there is one outlier in the NS group). The notches on the boxplot can be used to get a rough idea of the 'significance of differences between the values' (McGill *et al*. 1978). This is not exactly the same as the 95% confidence interval; the actual calculation in R is ± 1.58 IQR/sqrt(n) (see R help for 'boxplot.stats' for more information). If the notches lie outside the hinges (outside the box part), as they do just slightly for the Non and Early groups, this would indicate low confidence in the estimate (McGill *et al*. 1978).

Readers who have been convinced that boxplots are useful will find that it is easy to switch from barplots to boxplots since practically any program which can provide a barplot (SPSS, SAS, S-PLUS, R) can also provide a boxplot. Directions for making boxplots in SPSS and R are included in the online Appendix A.

Loess lines on scatterplots

A move from barplots to boxplots will improve visual reporting with group difference data. A way to improve visual reporting of relationships between variables is to include a smoother line along with the traditional regression line on a scatterplot (Wilcox 2001). Smoothers provide a way to explore how well the assumption of a linear association between two variables holds up. If the smoother line and regression line match fairly well, confidence is gained in assuming that the data are linear enough to perform a correlation

(Everitt and Dunn 2001). There are many kinds of smoothers (Hastie and Tibshirani 1990), but the one that is used often for fitting non-parametric curves through data by authors such as Wilcox (2001) and Crawley (2007) is Cleveland's smoother, commonly called the Loess line (Wilcox 2001). This line is a locally weighted running-line smoother, and it calculates lines over small intervals of the data using weighted least squares. In layman's terms, it is like regression lines are being calculated for small chunks of the data at a time. Clearly, if the concatenation of locally produced regression lines matches the regression line calculated over the entire data set, the assumption of linearity throughout the data set is upheld. Figure 2 shows four sets of data that contain both regression lines and Loess lines (note that these graphs are meant for illustrative purposes only, not for making actual inferences about relationships of the variables labeled).

Although the smoother line can be used as a guide, it is impossible to set out infallible guidelines for visually determining whether the regression line is

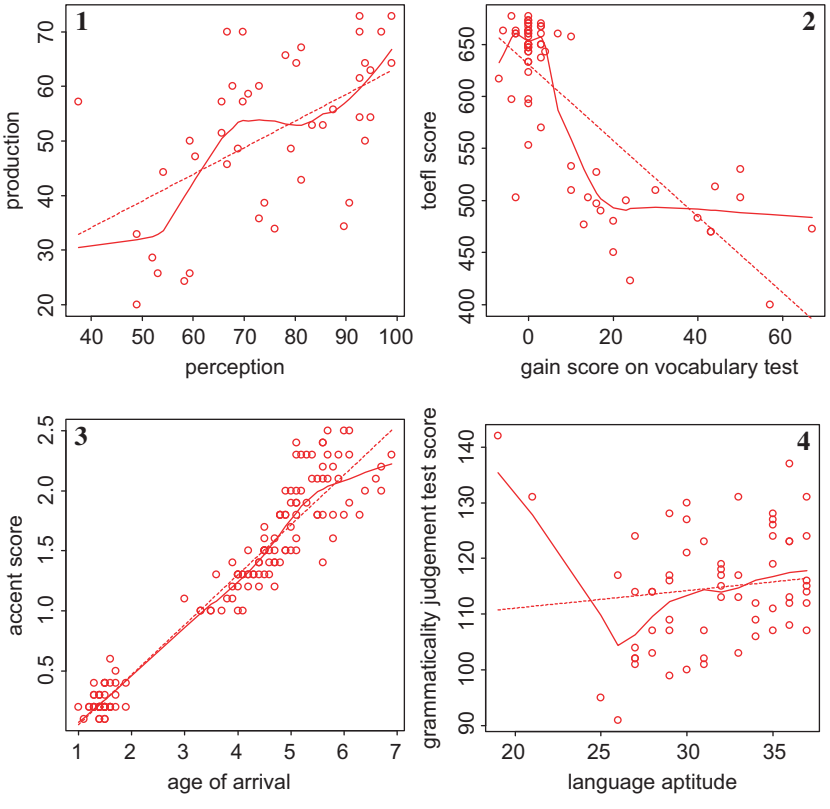

Figure 2: Four scatterplots with superimposed regression (dotted) and Loess lines (solid)

'close enough' to the Loess line to say that the data are linear (formal methods for testing curvature do exist however; see Wilcox 2005: 532–3). This is a matter of judgement that will improve with seeing more examples, which is why researchers who make claims about relationships between variables should provide scatterplots that contain both regression and Loess lines. Then, no matter what the author claims, readers will be able to make judgements for themselves on the appropriateness of assuming a linear relationship between the variables.

In Figure 2, we would say that the Loess lines in graphs 1 and 3 are 'close enough' to be considered linear. On the other hand, the Loess line in graph 2 shows a large deviation from a straight line, and it is likely the data should be analyzed as two different groups, as there seem to be two different patterns in the data. In graph 4, there appears to be a modest positive correlation between the variables, but the two outliers at the far left of the graph have skewed the regression line to be essentially flat. The smoother line shows a sharper angle in the non-outlier data.

Directions for creating a Loess line over a scatterplot in SPSS and R can be found in the online Appendix A. Other graphics that we don't discuss here, such as the relplot (which resembles the plot of ellipses shown later in this article in Figure 7; see Wilcox 2003 for more information) can help identify outliers in relationships between two variables. The kernel density estimator (g2plot using Wilcox's commands; see Wilcox 2003: 87 for an example) is an improvement on the histogram and can give a different perspective from boxplots. In addition, the shift function is a good graphic for comparing two groups (see Wilcox 2003: 276). A whole variety of exciting graphs that can be used with R can be viewed at addictedtor.free.fr/graphiques.

ROBUST STATISTICS

Introduction

In this section we explain to our reader why robust statistics are a desirable and useful tool to learn more about. What we call here robust statistics are not new; in fact, many of the robust alternatives to standard statistical estimates were proposed by scientists in the late 19th and early 20th centuries. However, the foundational works on robust statistics were published in the 1960s and early 1970s, with works such as Tukey (1960, 1962), Huber (1964) and Hampel (1968).[3] While work has continued vigorously on robust statistics since that time, practically speaking one needs statistical programs and adequate computational power in order to use robust statistics, and these requirements have only just come into view in the recent past[4] (we prefer the free R statistical program, see http://www.r-project.org; Maronna *et al.* 2006 assert that the most complete and user-friendly robust library is the one found in S-PLUS, which is also available in R; Rand Wilcox also has many robust

functions that can be incorporated into R or S+ and are available at http://www-rcf.usc.edu/~rwilcox/in the allfun or Rallfun files).

The programs are available, the computers are fast enough, and researchers can now begin to take advantage of the improvements that incorporating robust statistics into their own work will provide. Appendix A, found online, will provide some code to understand how we ran all of the robust statistics that are used in this article.

We will introduce below the concepts of trimmed means and bootstrapping, which are useful procedures that can help readers understand how robust statistics differ from classical statistics. Before we do that, however, readers will want to know why the use of robust statistics is desirable. Conventional wisdom has often promoted the view that standard analysis of variance (ANOVA) techniques are robust to non-normality, and that small deviations from the idealized assumptions of statistical tests (such as a normal distribution) would result in only minimal error in conclusions that were reached. Such is the view still of almost any book on statistics or research methods that you could lay hands on in the social sciences, which may make readers somewhat skeptical of our claim. This view is fairly accurate only with respect to Type I error (Wilcox 2001) (rejecting the null hypothesis when in reality it is true, and there actually is no difference between groups). When it is assumed that there are no differences between groups in a group difference testing setting (for example, one might want to show that a group of advanced non-native speakers do not differ from a native speaker group), then the probability level corresponding to the critical cut-off score, used to reject the null hypothesis, is found to be close to the nominal level of 0.05. However, statistical simulation studies have found that standard methods are not robust when differences exist (Tukey 1960; Hampel 1973), which is more often the situation that researchers are hoping for (such as, for example, when two treatments are applied and the researcher is hoping that one will result in more language learning).

Tukey (1960) found that one of the most problematic distributions was one he called a 'contaminated normal' distribution, which visually is quite close to a normal distribution. The contaminated normal is slightly longer-tailed than normal distributions (Huber 1981; Wilcox 2001), as can be seen in Figure 3. The contaminated normal is formed mathematically by taking two normal populations with the same mean, but with one that has a larger standard deviation than the other, and mixing data from the population with the wider standard deviation into the population with the narrower standard deviation (Tukey 1960).

The problem with the longer tails of the contaminated normal is that the extra data points in the tails means that the amount of variability is increased, and this makes it more likely that differences which are in fact statistical[5] are found to be non-statistical (Tukey 1960; Huber 1981; Wilcox 2001). The reason this is important to SLA data is that real data sets in Applied Linguistics are probably not *exactly* normally distributed (Micceri 1989 claims

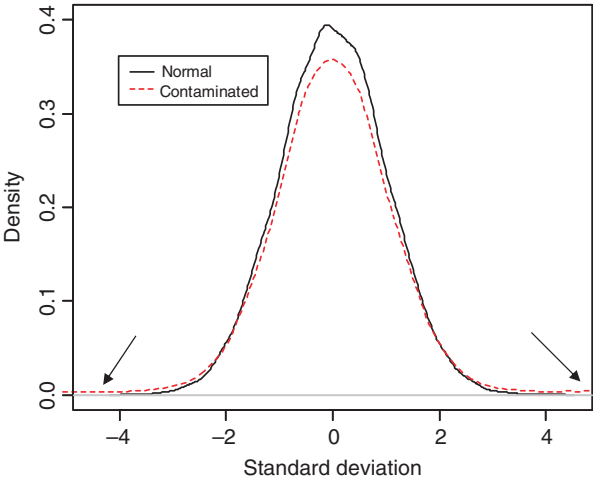

Figure 3: Density function of a normal distribution and a superimposed contaminated normal distribution

this for psychological data), and may demonstrate deviations from normality including heavier tails (as evidenced by outliers) or skewness. As readers can see in Figure 3, it would be quite difficult to tell the difference in a data set between data with an exactly normal distribution versus a distribution that is symmetric but heavy-tailed. Even small departures from normality (not to mention much larger ones such as obvious skewness) can have an effect on the statistical conclusions that can be drawn.

Wilcox (2001) notes that in a standard normal distribution the variance is 1, but in a contaminated normal like that in Figure 3 the variance has increased to 10.9. Such inflation of the variance means that the standard error will also be inflated, and since statistical tests divide by some measure of variability like variance or standard error, the resulting statistic will be smaller when the variance is larger (and less likely to be statistical).

An illustration from Wilcox (2003) can help clarify this point. Imagine we have 10 data points for 2 groups, shown in Table 2. For the sake of this article, let's say they represent scores on a test of how much vocabulary, out of a possible 25 points, was remembered after Group 1 received no treatment (the control group) and Group 2 received a special treatment (the treatment group). The mean score of the control group is 5.5 and the mean of the treatment group is 8.5. Is this difference statistical? To test the null hypothesis that there is no difference between the groups, apply an independent samples *t*-test. The *t*-test value is $t = -2.22$ and $p = 0.039$. The *p*-value is below the normal alpha level of $\alpha = 0.05$, and thus we may reject the null hypothesis, and conclude there is a statistical difference between groups. However, say that the score of the 10th participant in the treatment group is changed from 13 to 25. Now the

Table 2: Original scores for a fictional vocabulary retention experiment

Group 1: control	1	2	3	4	5	6	7	8	9	10
Group 2: treatment	4	5	6	7	8	9	10	11	12	13

average of the treatment group (Group 2) becomes 9.7. Logically, because the difference between sample means has increased, we would still want to conclude that there is a statistical difference between groups. However, because the score of 25 increases the variance (the distance from the mean) in the treatment group, this increases the denominator of the *t*-test equation,

$$t_{\mathrm{df}} = \frac{X_T - X_C}{\sqrt{\mathrm{var}_{\mathrm{pooled}}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}}, \quad 6$$

leaving us with a smaller *t*-value ($t = -1.99$) and a *p*-value larger than our alpha ($p = 0.07$). In other words, with the one changed value we now conclude that the groups are statistically *not* different! This goes counter to our sense of group differences, but shows that more data in the tails of the distribution, and thus more variance, can affect *p*-values and statistical conclusions.

To summarize thus far, while small deviations from normality in the distribution are fairly robust to Type I errors (rejecting the null hypothesis when in reality it is true, and there actually is no difference between groups), we are much more likely to make a Type II error (accepting the null hypothesis when in reality it is not true and there actually is a difference between groups) with such deviations (Hampel *et al*. 1986). Making Type II errors means that we are losing power to find true differences between groups or relationships between variables. Power is a technical statistical term, but can be understood here in layman's terms to mean the strength to find a result.

We will give an example of the kind of problems that have been found with small departures from normality. Wilcox (1995) reported on the power of the Welch procedure that is used in *t*-tests when variances are unequal. The power of this test to find the true results when the distribution is normal is 0.93 (where 1.00 is perfect power), but drops to 0.28 when the distribution is a contaminated normal with a standard deviation of 10, and to 0.16 when the contaminated normal has a SD of 20 (Wilcox 1995: 69). On the other hand, a test procedure based on 20% trimmed means (a robust method described in more detail below) yields power of .89 with the normal distribution, and only lowers to 0.78 for a contaminated normal with $K = 10$, and 0.60 with a contaminated normal of $K = 20$ (ibid.). Statisticians agree that robust statistics are even more necessary when statistical models more complex than *t*-tests are used (Hampel *et al*. 1986).

Statistical modeling has shown that robust methods work much better than parametric methods when the underlying distribution is not normal, and they

work nearly as well if the underlying distribution is in fact normal (Tukey 1960; Yuen and Dixon 1973; Huber 1981; Luh and Guo 2001; Wilcox 2001). As has been discussed above, the idea that statistical tests are robust to small deviations from normality should not be assumed. Additionally, rules of thumb, such as those which assert that if group sizes are 30 or more there is no reason to worry about meeting normality requirements (Pallant 2001; Weinberg and Abramowitz 2002), are also inaccurate. Westfall and Young (1993) performed a simulation study which found that with group sizes of $n = 160$, skewed distributions, even without outliers, could have very poor results. Using data from an actual study, Wilcox (2003: 123) found that even with $n = 105$, the *t*-test performed poorly and more than 300 subjects would have been necessary to get good results. Remember that poor results mean that although there may indeed be differences between groups or relationships between variables, traditional parametric statistics will not be able to detect that difference.

Huber (1981) states that robust methods are more similar to parametric methods than nonparametric or distribution-free methods, because they continue to use the same parametric models; the difference is that the parametric models 'are no longer supposed to be literally true, and . . . one is also trying to take this into account in a formal way' (Huber 1981: 6). Since robust methods can deal with non-normality, including skewness, and because it is nearly impossible for researchers to know with certainty that their distributions are normal ones, we know of no reason not to recommend that researchers learn more about robust methods and employ them in all cases.

Outliers

Many are familiar with Mark Twain's quote that there are 'lies, damn lies, and statistics' (see the August 2005 issue of *Statistical Science* for a sophisticated and sometimes tongue-in-cheek discussion of how such lying may be accomplished). One reason this aphorism may resonate with those who have used statistics in their own research is that the addition or subtraction of just one participant, or an incorrect data entry for one participant may result in a totally opposite conclusion to the one reached before the participant was added or subtracted or before the data entry was corrected. The kinds of non-robust estimators, such as the average, that are used in parametric statistics can be easily affected by just one extreme point.

Many researchers realize this, and perform their statistics with outliers removed, usually showing the reader a graph so that the outlier's 'outlying-ness' can be perceived, and sometimes performing statistics with the outlier both included and removed. Removing outliers is definitely an important step to take to make the data fit the assumptions of normality that are imposed by classical parametric statistics. There are several problems with this ad hoc basis for removing outliers, however. The first problem is that throwing away data points that seem to be outliers results in the non-independence of

the remaining data (Wilcox 1998), and independence of the data is one of the assumptions for all statistical tests. Huber affirms that 'classical normal theory is not applicable to cleaned samples' (1981: 4). The second problem is that the decision about what points to remove is personal and subjective. Robust methods provide objective and replicable ways of diagnosing outliers and then performing statistical inferences with these outliers removed (Hampel *et al*. 1986). The third problem is that what is often the problem in a distribution is not the obvious outlier but the 'outliers' to the normal distribution which reside in the heavier tails of the contaminated normal and are not easily dealt with. One way that has been devised to deal with the problem of outliers in robust statistics is by using trimmed means. Other, similar types of procedures (among them, M-estimates, L-estimates, and R-estimates) are more mathematically complicated but follow the same basic logic, so we introduce our reader to trimmed means as a general procedure which is widely employed in robust statistics.

TECHNIQUES USED IN ROBUST STATISTICS

Measures of location and trimmed means

The mean or average of the data is an example of a non-robust estimator. It can be highly influenced by just one outlier in the data. An alternative to the mean is the median score, which is quite robust to outliers. The problem with the median is that it effectively discards all data points except for one or two. We want the estimator we use to reflect what is typical of the data set without being distorted by outliers. The median is not distorted by outliers, but it also does not include much information from the data set.

The trimmed mean represents a compromise between the mean and the median, and between power and bias in the test statistics (Huber 1981). A trimmed mean captures the shape of the data without giving too much weight to outliers by trimming points off the ends of the data set. In theory, any amount could be trimmed, but Wilcox (2001, 2003) on the basis of simulation studies asserts that 20% is a good amount for general use.[7] The way to trim means is to first put the observations in numerical order. To trim by 20%, multiply .2 by the n. Thus, with a data set where $n = 10$, two points would be trimmed off both the lowest and highest end of the data (since $0.2n = 0.2(10) = 2$), resulting in a data set with six scores. This may seem unintuitive—if you have a small data set, you do not want to make it smaller by discarding data! However, robust statistics will in fact result in a more reliable description of the 'average' trend than if all of the data points had been left included.

In fact, one objection that might be raised to trimmed means is that they discard information. It is true that they do, but the idea is that they do not throw away as much information as the median, while being more resistant to outliers than the mean, yet still capturing the general trend of the data.

If the data set is not *exactly* normally distributed, and we would assume that most data sets in our field are not, the trimmed means will be a better reflection of what is typical in the data set.

Because the 20% trimmed means is mathematically quite easy to perform, we would like to note that researchers should not try to use this method without finding statistical programs which can evaluate data using complete arrays of robust techniques. For example, one cannot just plug the 20% trimmed mean into the equation for the sample variance in the same manner as for the untrimmed mean. Removing extreme data points from the set results in interdependence among the remaining points. We will need special equations now to calculate the variance if trimmed means or other robust estimators are used. Proper types of software which calculate trimmed means will ensure that these requirements are met, and can be found for free using the R statistical program and Wilcox's robust commands (recommended books for getting started with robust statistics in R are Crawley, 2007 and Wilcox, 2003).

Bootstrapping

Bootstrapping is another tool in the robust statistical toolbox that can help researchers make more accurate conclusions about their data. Bootstrapping is an approach to statistical inference that makes fewer assumptions about the underlying probability distribution that describes the data than the normal Gaussian distribution does (Efron and Tibshirani 1993). In this type of approach, as Westfall and Young (1993: 12) describe, 'the observed data are used repeatedly, in a computer-intensive simulation analysis, to provide inferences. In simple terms, resampling does with a computer what the experimenter would do in practice, if it were possible: he or she would repeat the experiment.'

It turns out that this process is exactly the same process that was used by the statistician Gosset as an empirical verification of his mathematical derivation of the null distribution of the Student's *t*-test (Student, 1908, as discussed in Wilcox, 2001). Gosset simulated the null distribution by sampling from a normal distribution, calculating the mean and standard deviation of each observation, and finding the resulting *t*-test statistic. Repeating the process over and over, critical values for *t* were then determined. Because Gosset did this without a computer, the process took over a year. Now resampling methods can do the same kinds of simulations in several seconds, except that in bootstrap resampling, the resampling is done from observed data and not from the hypothetical normal distribution.

Using this process bootstrapping generates a distribution (an empirically generated sampling distribution) that can be examined for the significance of the statistics in the same way that the critical value of a *t*-test, based on a normal distribution, can be examined for significance, just as Gosset did.[8] This approach assumes that the empirical distribution function is a reasonable

estimate of the unknown, population distribution function. Using the data as an approximation to the population density function, data are re-sampled with replacement[9] from the observed sample to create an empirical sampling distribution for the test statistic under consideration. This resampling is done thousands of times without regard to the original groupings of data, resulting in proxy samples. In the resampling method, the hypothesis testing is done by noting that for the proxy samples, any statistical differences between groups should be due merely to chance. The percentage of statistical tests for the proxy samples which are as large or larger than the observed statistical difference determines the observed *p*-value for the data. Accordingly, when the observed *p*-value is less than 0.05 (or any other threshold we may care to set, but this is the generally accepted level in the field, although there are good arguments for setting it to 0.10, see Kline 2004), we reject the assumed null hypothesis of no difference in the population.

The number of bootstrap samples that should be performed should also be considered when doing a bootstrap. Although it is quite easy to ask for a very high number of samples, work by Wasserman and Bockenholt (1989) shows that in many cases, no more than 1000 bootstrap samples are required to obtain accurate confidence intervals for a location estimate.

An example of how this would work for a *t*-test is that the original *p*-value of the *t*-test done with the original data is compared with the *p*-values of the *t*-tests performed for all of the groups created by replacement sampling. The test statistic (the *t* value in the *t*-test) generated by the proxy samples are then compared with the test statistic generated by the original *t*-test; consequently, '[t]he resampling-based *p*-value is then the proportion of resampled data sets yielding a *t*-statistic as extreme as the original *t*-statistic' (Westfall and Young 1993: 13).[10] This *p*-value is used in the same way as the familiar *p*-value: if it is less than 0.05, the difference between groups is assumed to be statistical.

The reader should be able to see then, that the logic by which the *p*-value is generated in the resampling case is the same logic as that behind the 'classical' parametric tests that are used, but in the resampling case, the empirical cumulative distribution function (whatever that turns out to be, given the data) is used to make inferences about the likelihood of the *p*-value given the data instead of the normal distribution. Indeed, the middle 95% of the ordered means of the bootstrap sample will comprise the 95% confidence interval of the data. The great value of resampling is thus that researchers do not need to assume that the data are normally distributed. Although the bootstrap does not eliminate problems due to skewness (Wilcox 2003: 220), the combination of 20% trimming and the bootstrap does make a practical difference[11] (in some cases smaller confidence intervals for skewed data can be achieved by using a more refined bootstrap method referred to as the 'abc percentile' method, see Efron and Tibshirani 1993). Simulation studies run by Wilcox (reported on in Wilcox 2003: 220) show that with skewed, heavy-tailed distributions, bootstrap methods can reduce Type I error probabilities compared with Student's *t*, although they are still substantially higher than $\alpha = 0.05$. Although larger

sample sizes are still desirable because they increase the precision of the confidence interval, the bootstrap is able to function well with symmetrically distributed moderately small samples, such as $n = 10$ (Westfall and Young 1993; Chernick 1999).

Combining bootstrapping with trimmed means has been shown, in a variety of papers by Wilcox and colleagues (Keselman *et al*. 2000, 2003; Wilcox 2001) to further reduce problems with skewed distributions. Problems with skewed distributions are not entirely erased in all cases but robust methods certainly provide a better way of dealing with skewed distributions than using traditional parametric methods.

Now we will illustrate the concept of the bootstrap by using data from a real experiment. The data come from an unpublished study of how accurately various groups of Japanese learners of English produced words beginning with /r/ and /l/ (data available upon request to first author; see Appendix B online for more details about this study). First we will show how the bootstrap operates on one group of data, just to illustrate the idea of bootstrapping. The scores of the group who lived in the US at an early age are given in Table 3, arranged in a numerically ascending order.

The distribution of scores is clearly non-normal and skewed, as shown on the histogram in Figure 4. The mean of the scores is 95.2, but there is one outlier whose score was much lower than this.

A bootstrap sample using 1000 randomly generated samples might include samples like those found in Table 4.

The bootstrapped sampling distribution (called a percentile or uncorrected bootstrap; this is different from the percentile-t bootstrap; see Wilcox 2003 for more information) contains a set of mean scores for the entire group calculated from the 1000 random samples sampled with replacement. In other words, each sample, such as Sample 1, is averaged to give a mean score. Because there were 1000 random samples, 1000 mean scores were thus generated. These new mean scores range from 87.86 to 99.14. As can be seen from the histogram for the bootstrapped sample (Figure 5), the array of mean scores now forms a distribution. This is the empirical distribution by which the mean score of the original data set will be judged.

Note that the bootstrap distribution in Figure 5 was done using all 14 scores in the original sample data. Using the 20% trimmed means would be a way to eliminate the skewing influence of the low score of 72. In the case of $n = 14$,

Table 3: Scores of 'early immersionists' on an accuracy of initial /r/ and /l/ measure

Participants	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
Accuracy	72	90	90	96	96	97	98	98	99	99	99	99	100	100

$0.2n = 2.8$ (when the result is a decimal then round down), so we would eliminate the lowest and highest two scores from the distribution.

Next we will compare the results of a parametric analysis and a robust analysis of the language accuracy task with all groups of participants included. There are three groups of Japanese users of English in this study: (i) the 'early immersionists' lived in the USA as children but returned to Japan by age 7; (ii) the 'late immersionists' lived in the USA as young adults; (iii) the 'non-immersionists' had never lived in an English-speaking country but were majoring in English at their university. Additionally, there was a group of native speakers of English who produced words beginning with /r/ and /l/. The measure being compared here is how accurately what the participants produced aligned with how native speakers of English perceived the initial sound to be (again, more details about the entire study can be found in the online Appendix B; also, specific code used to generate the robust analysis is found in online Appendix A).

Because there are four groups in all, a parametric analysis would use a one-way ANOVA. A one-way ANOVA returns a statistical main effect, $F_{3,55} = 5.27$, $p = 0.003$. Tukey post hoc tests among the groups found that the NS were statistically different from the non-immersionists ($p = 0.002$) but not the late immersionists ($p = 0.407$) or the early immersionists ($p = 0.834$). Using robust statistics, 20% means-trimmed bootstrapped multiple comparisons between the NS and the three non-native groups finds substantially different *p*-values

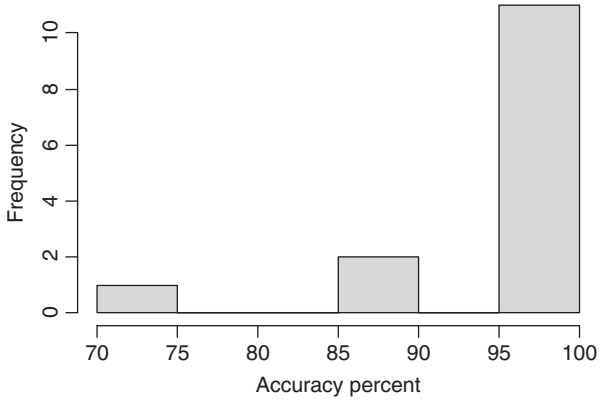

Figure 4: Histogram of original scores on the accuracy measure by early immersionists

Table 4: Two possible bootstrapped samples of the original accuracy measure

Sample 1	72	72	90	98	98	98	99	100	100	100	100	100	100	100
Sample 2	98	98	98	98	98	98	98	99	99	99	99	99	99	100

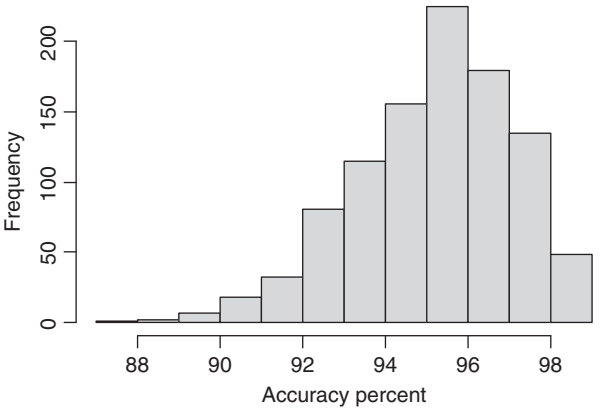

Figure 5: Histogram of bootstrapped mean scores on the accuracy measure by early immersionists

(non, $p = 0.0000$; late, $p = 0.01$; early, $p = 0.01$).[12] Because the sample sizes used were small, robust statistics provide a weight of evidence that with repeated testing, differences would be found between all of the non-native groups and the NS.

At this point, our eager reader may be wondering how to begin using robust statistics in his or her own work. The ideas presented in this article are merely a sampling of the wide variety of robust methods that are available, and for further information we recommend starting by looking at books by Crawley (2007) and Wilcox (2003). Information about the use of the bootstrap in a robust test of bivariate correlation can be found in Wasserman and Bockenholt (1989). Robust tests have been extended to virtually all of the ANOVA methods including repeated measures designs (see Wilcox 2003, 2005). Performing resampling methods is possible using many different statistical programs; MacKinnon *et al*. (2004) remark that resampling methods are available without further modification in AMOS (used in SPSS) and SAS. We note that the terminology of statistics changes very little when robust methods are used. In other words, when you use a robust *t*-test, you will still report the value of the test, a *p*-value associated with it, and confidence intervals and effect sizes as per parametric tests. The only difference is that you will report that you used the 20% trimmed means, or bootstrapping, and name the robust method that was applied, such as Yuen's (1974) method for comparing two independent groups (more information about names can be found in Wilcox, 2003 and 2005).

A NEW PERSPECTIVE ON DATA ANALYSIS

The example given above showed that statistical conclusions about differences between groups can change when robust techniques such as bootstrapping

and trimmed means are used. This section will give further examples of how robust statistics can provide a new perspective on data analysis, using examples with language acquisition data. The first example uses data from a study made by the first author (Larson-Hall 2008) of the language abilities of 200 Japanese users of English, some of whom began studying English at a young age, and others who began their study in junior high (see Appendix B online for more information about this study). One of the research questions examined was whether the age that students began studying English affected their scores on an oral phonemic discrimination test and a grammaticality judgement test when total amount of input was factored out. Conventional analysis of covariance (ANCOVA) analysis found that there was no effect of group (earlier or later starters) for the grammaticality judgement test ($F_{1,197} = 1.69$, $p = 0.20$), but the effect of group was statistical for the phonemic discrimination test ($F_{1,197} = 6.55$, $p = 0.01$). The problem with conventional ANCOVA analysis is that it compares groups assuming a linear association, and if the data are not linear, the ANCOVA will generally not be statistical. On the other hand, a robust ANCOVA (we used the ancboot command, found in Wilcox, 2005, p. 529, which uses the 20% trimmed means and bootstrapping, and performs well with heteroscedasticity) does not require a linear association. The ancboot method of ANCOVA compares linear models along a running-interval smooth (similar to the Loess line), finding the tendency of the data instead of forcing it to be along a straight line. This analysis indicates when there are group differences at specific points along the *x*-axis. In the case of the Larson-Hall (2008) data, a robust ANCOVA found a statistical advantage for later starters on the GJT at 800 hours of input, but an advantage for earlier starters at 1833 and 2000 hours of input. For the phonemic discrimination test, a robust ANCOVA found a statistical advantage for the earlier starters at 1300, 1555, 1833 and 2000 hours. The results of the ANCOVA can be more clearly understood by looking at scatterplots with the two groups separated, as in Figure 6. The scatterplots are overlaid with the smooth lines, indicating the trend of the data in Figure 6 at specific hours of input on the *x*-axis.

What the robust statistics do is give a more nuanced picture of the combined influences of age and input on test scores, and in fact provide a way to integrate the results of previous studies which found no beneficial effects for a younger starting age with the results of this study which did find beneficial effects (previous studies were looking only at the very low end of hours of input, where advantages for earlier starters did not appear).

Another example comes from a reanalysis of raw data provided by DeKeyser (2000). DeKeyser gave 57 Hungarian immigrants to the USA a grammaticality judgement test and examined the correlation between their age of arrival (AOA) and their scores. Like many other studies examining the relationship between age of acquisition and ultimate language ability, his data showed a statistical and negative correlation between AOA and scores across the entire range of children and adults ($r = -0.63$). DeKeyser did not focus on this overall

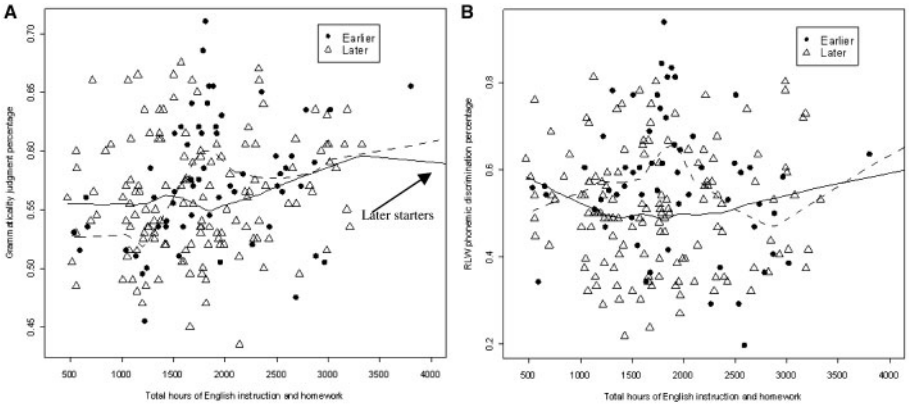

Figure 6: Scatterplots showing results for groups of earlier and later learners of English on a grammaticality judgement test (A) and phonemic discrimination test (B) as a function of hours of input. Smooth lines are calculated for both groups (dashed line for earlier starters and unbroken line for later starters)

score much, as he wanted to show there were different patterns between younger and older arrivals, and he somewhat arbitrarily chose a cut-off point of 15 to split the groups.

Robust estimates of location can result in different but more valid statistical results than classical parametric tests when the researcher is interested in making inference about the majority of the observations in a population. A robust correlation using the cor.plot command from the mvoutlier library in R with DeKeyser's data (using an algorithm for outlier detection; R code for this command is illustrated in online Appendix A) reveals that there is no statistical correlation across the part of the data in the sample that excludes outliers ($r = 0.03$, ns). Figure 7 shows ellipses containing the data used in each of the correlations (the classical correlation and the robust correlation). The figure shows that the robust correlation excludes most of the data from the youngest learners in order to find the data which best represents the overall trend. It can be seen that robust correlation could even provide a principled reason for splitting the data (although at a different point than the one DeKeyser used), and this example shows again that data can be seen in a new light when robust statistics are used.

These two examples serve to show that robust statistics can make a difference in the statistical analyses that are done in the field. We would like SLA researchers to become aware of some of the most important and enduring changes taking place in the field of modern statistics because we feel they can profitably be applied to improve the accuracy and reliability of our own studies.

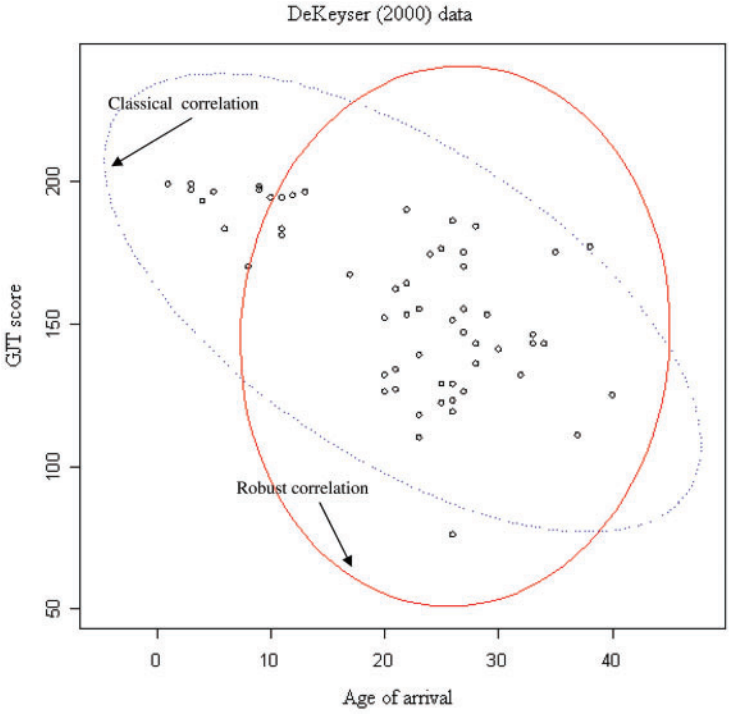

Figure 7: Comparison of data included in classical and robust correlation of the DeKeyser (2000) data

CONCLUSION

Quantitative articles which use statistical methods are the kind of studies most often published in the SLA field (Lazaraton 2000). As such, methods that improve the accuracy of statistical inference should be highly important to the SLA field. Our main purpose in this article is the introduction of effective graphical procedures and robust statistics to researchers in the language acquisition field. Small changes in the way researchers analyze and present data can make large differences in the way research is comprehended.

Work in modern statistics has shown that parametric tests which have been assumed to be robust to slight deviations from normality are in fact not. Robust statistics have been formulated to deal with real, applied data that does not necessarily conform to a normal distribution, and using robust statistics routinely will lead to more power to discover real differences and more accuracy in estimating the statistics involved. Robust statistics also provide objective and replicable ways of dealing with outliers, and deal better with small and non-normally distributed data sets than parametric statistics.

We have also suggested that researchers always include graphics in their research reports (along with raw data, if possible), and that these graphics should be as informative as possible. In practice, we suggested that researchers should use boxplots instead of barplots for group difference data, and add Loess (smooth) lines along with regression lines to scatterplots.

We want to emphasize that modern statistical methods do not solve every problem that may arise when conducting statistical analyses, but they do offer a much better way of approaching quantitative analysis. Although there may not always be one easy and best way to solve every problem using robust methods, years of research in statistics have shown that parametric statistics, which depend on the assumption of normality, are not nearly as accurate as robust statistics in almost every case (Tukey 1960; Yuen and Dixon 1973; Huber 1981; Luh and Guo 2001; Wilcox 2001). If we are to be responsible researchers we need to find out about the advances that have been taking place in the field of statistics for the last 40 years and incorporate these methods, which are much more practical for authentic data sets, into our analyses.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

NOTES

1 Various statisticians have called for providing raw data (Fienberg *et al*. 1985; Westfall and Young 1993) but as far as we know, this is not required for any journals in the field of SLA. Raw data can help others verify that statistical procedures have been used correctly and that conclusions are based on solid statistical reasoning. An example will serve to illustrate our point. An article by Hirata (2004) erroneously concluded that groups in the study were statistically different. We know this because the author provided the raw data for her eight participants. In reporting on the differences between groups for the perception tasks, Hirata apparently used the significance value (the *p*-value) from Levene's test for equality of variances, not from the *t*-test. Hirata reported that the *p*-value for the difference between the experimental and control group was $p = 0.004$ in the post-test condition, while in fact the *p*-value was $p = 0.20$, meaning that the groups were *not* statistically different in the post-test.

2 It should be noted that if the frequency of use of recasts were compared over 10 different teachers, a boxplot might be entirely appropriate to show the range of frequencies. In any case where an average could be computed a boxplot could be used.

3 The term 'robust statistics' is applied to a whole range of techniques that are meant to make data more robust to violations of assumptions of classical techniques.

4 As recently as 1993, Westfall and Young stated that the major impediment to using robust statistics in applied work was that they were time-consuming with the computational power available. This is no longer the case.

5 The use of the term 'statistical (difference)' here is deliberate. Although some statisticians believe it is more accurate to say 'statistically significant difference' or 'significant difference', we have chosen to follow Kline's (2004) recommendation to return the use of the word 'significant' to its ordinary meaning of 'important' (which is does NOT necessarily mean when it modifies 'statistical') and simply call differences statistical.

6 Definitions for the variables in this equation: t_{df} = the t-value at the given degrees of freedom; X_T = mean of the treatment group; X_C = mean of the control group; var_{pooled} = the pooled variance, which is equal to $(n_T - 1)$(standard deviation of the treatment group)^2 + $(n_C - 1)$(standard deviation of the control group)^2 all divided by $n_T + n_C - 2$; n_T = number in the treatment group; n_C = number in the control group.

7 It should be noted that this type of trimming is symmetric trimming, where an equal number of points is removed from both ends of the distribution prior to the computation of an estimate of the location of the distribution. More recently, methods for asymmetric trimming have been proposed (Keselman *et al.* 2007).

8 To remind readers, hypothesis testing using parametric statistics calculates a test statistic using various formulas but mostly involves using the average or average differences, within-group variances of the data set, and then returns a probability (or p-value) for the observed test statistic. This probability indicates the probability with which the same or even more extreme results would be found if the null hypothesis were true (Klein 2004: 63–4). The p-value is the probability of the data given the hypothesis, written ($p(D|H_o)$), not the probability of the hypothesis given the data, written ($p(H_o|D)$) (Nickerson 2000). Hypothesis-testing is a procedure that relies on the theoretical sampling distribution to determine whether the data are probable given the null hypothesis. The theoretical sampling distribution, assumed to be a Gaussian or normal curve, produces the p-values for the null hypothesis.

9 Resampling with replacement means that as each number from the original data set is randomly drawn, it is returned to the original set and may be chosen again in future draws. Each computer-generated sample is the same size as the original data set.

10 Westfall and Young created the *multtest* package in R which performs these types of bootstraps.

11 When resampling is done in combination with bootstrapping, samples are taken from the entire data set, not the trimmed set. Trimming is done on the bootstrap sample that is generated from all of the data.

12 Although by a rubric of $p < 0.05$ all of these values are statistical, because there are multiple comparisons, an adjustment is made that sets the cut-off p-value lower, to $p = 0.009$ in fact. By this cut-off value, the differences between the late and early groups with the NS are still not statistical, although of course at $p = 0.01$ they are much closer to that point to be statistical than in the non-robust version.

REFERENCES

Chernick, M. 1999. *Bootstrap Methods: A Practitioner's Guide*. Wiley-Interscience.

Crawley, M. J. 2007. *The R Book.* Wiley.

DeKeyser, R. M. 2000. 'The robustness of critical period effects in second language acquisition,' *Studies in Second Language Acquisition* 22: 499–533.

Efron, B. and **R. J. Tibshirani.** 1993. *An Introduction to the Bootstrap.* Chapman & Hall.

Everitt, B. and **G. Dunn.** 2001. *Applied Multivariate Data Analysis.* 2nd edn. Hodder Arnold.

Fienberg, S. E., M. E. Martin, and **M. L. Straf.** 1985. *Sharing Research Data.* National Academy Press.

Hampel, F. R. 1968. *Contributions to the Theory of Robust Estimation*. Unpublished PhD thesis, University of California, Berkeley.

Hampel, F. R. 1973. 'Robust estimation: A condensed partial survey,' *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* 27: 87–104.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and **W. A. Stahel.** 1986. *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

Hastie, T. J. and **R. J. Tibshirani.** 1990. *Generalized Additive Models.* Chapman and Hall.

Hirata, Y. 2004. 'Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts,' *Computer Assisted Language Learning* 17/3–4: 357–76.

Huber, P. J. 1964. 'A robust version of the probability ratio test,' *Annuals of Mathematical Statistics* 36/6: 1753–8.

Huber, P. J. 1981. *Robust Statistics.* John Wiley & Sons.

Keselman, H. J., J. Algina, R. Wilcox, and **R. K. Kowalchuk.** 2000. 'Testing repeated measures hypotheses when covariance matrices are heterogeneous: revisiting the robustness of the Welch-James test again,' *Educational and Psychological Measurement* 60: 925–38.

Keselman, H. J., R. R. Wilcox, and **L. M. Lix.** 2003. 'A generally robust approach to hypothesis testing in independent and correlated groups designs,' *Psychophysiology* 40: 586–96.

Keselman, H. J., R. Wilcox, L. M. Lix, J. Algina, and **K. Fradette.** 2007. 'Adaptive robust estimation and testing,' *British Journal of Mathematical and Statistical Psychology* 60: 267–93.

Klein, R. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* American Psychological Association.

Kline, R. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* American Psychological Association.

Larson-Hall, J. 2008. 'Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation,' *Second Language Research* 24/1: 35–63.

Lazaraton, A. 2000. 'Current trends in research methodology and statistics in applied linguistics,' *TESOL Quarterly* 34/1: 175–81.

Luh, W.-M. and **J.-H. Guo.** 2001. 'Using Johnson's transformation and robust estimators with heteroscedastic test statistics: an examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design,' *British Journal of Mathematical and Statistical Psychology* 54: 79–94.

MacKinnon, D. P., C. M. Lockwood, and **J. Williams.** 2004. 'Confidence limits for the indirect effect: Distribution of the product and resampling methods,' *Multivariate Behavioral Research* 39/1: 99–128.

Maronna, R. A., R. D. Martin, and **V. J. Yohai.** 2006. *Robust Statistics: Theory and Methods.* Wiley.

McGill, R., J. W. Tukey, and **W. A. Larsen.** 1978. 'Variations of box plots,' *The American Statistician* 32/1: 12–16.

Micceri, T. 1989. 'The unicorn, the normal curve, and other improbable creatures,' *Psychological Bulletin* 105/1: 156–66.

Nickerson, R. S. 2000. 'Null hypothesis significance testing: a review of an old and continuing controversy,' *Psychological Methods* 5/2: 241–301.

Pallant, J. 2001. *SPSS Survival Manual.* Open University Press.

Porte, G. K. 2002. *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research*. John Benjamins.

Student. 1908. 'The probable error of a mean,' *Biometrika* 6/1: 1–25.

Tufte, E. R. 2001. *The Visual Display of Quantitative Information.* 2nd edn. Graphics Press.

Tukey, J. W. 1960. 'A survey of sampling from contaminated distributions' in I. Olkin, S. G. Ghwyne, W. Hoeffding, W. G. Madow, and H. B. Mann (eds): *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling*. Stanford University Press, pp. 448–485.

Tukey, J. W. 1962. 'The future of data analysis,' *The Annals of Mathematical Statistics* 33: 1–67.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Addison-Wesley.

Wasserman, S. and **U. Bockenholt.** 1989. 'Bootstrapping: applications to psychophysiology,' *Psychophysiology* 26/2: 208–21.

Weinberg, S. L. and **S. K. Abramowitz.** 2002. *Data Analysis for the Behavioral Sciences Using SPSS.* Cambridge University Press.

Westfall, P. H. and **S. S. Young.** 1993. *Resampling Based Multiple Testing.* Wiley.

Wilcox, R. 1995. 'ANOVA: A paradigm for low power and misleading measures of effect size?,' *Review of Educational Research* 65/1: 51–77.

Wilcox, R. 1998. 'How many discoveries have been lost by ignoring modern statistical methods?,' *American Psychologist* 53/3: 300–14

Wilcox, R. 2001. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. Springer.

Wilcox, R. 2003. *Applying Contemporary Statistical Techniques.* Elsevier Science.

Wilcox, R. 2005. Introduction to robust estimation and hypothesis testing. 2nd edn. Elsevier Science.

Wilkinson, L. and **Task Force on Statistical Inference, A. P. A., Science Directorate, Washington, DC, US.** 1999. 'Statistical methods in psychological journals: guidelines and explanations,' *American Psychologist* 54/8: 594—604.

Yuen, K. K. 1974. 'The two-sample trimmed t for unequal population variances,' *Biometrika* 61: 165–70.

Yuen, K. K. and **W. J. Dixon.** 1973. 'The approximate behaviour and performance of the two-sample trimmed *t*.' *Biometrika* 60/2: 369–7.