Applied Linguistics 2015: 36/5: 591–610
doi:10.1093/applin/amu001 Advance Access published on 14 February 2014

Bootstrapping in Applied Linguistics: Assessing its Potential Using Shared Data

*LUKE PLONSKY, JESSE EGBERT and GEOFFREY T. LAFLAIR

Northern Arizona University
*E-mail: luke.plonsky@nau.edu

Parametric analyses such as *t* tests and ANOVAs are the norm—if not the default—statistical tests found in quantitative applied linguistics research (Gass 2009). Applied statisticians and one applied linguist (Larson-Hall 2010, 2012; Larson-Hall and Herrington 2010), however, have argued that this approach may not be appropriate for small samples and/or nonnormally distributed data (e.g. Wilcox 2003), both common in second language (L2) research. They recommend instead 'robust statistics' such as bootstrapping, a nonparametric procedure that randomly resamples from an observed data set to produce a simulated but more stable and statistically accurate outcome. The present study tests the usefulness of bootstrapping by reanalyzing raw data from 26 studies of applied linguistics research. Our results found no evidence of Type II error (false negative). However, 4 out of 16 statistically significant results were not replicated (i.e. a Type I error 'misfit' five times higher than an alpha of .05). We discuss empirically justified suggestions for the use of bootstrapping in the context of broader methodological issues and reforms in applied linguistics (see Plonsky 2013, 2014).

The field of applied linguistics is currently in the early stages of a methodological reform. This movement has produced calls for and evidence of gradual but substantial improvements in quantitative research practices such as more thorough data reporting and interest in practical as well as statistical significance (Norris and Ortega 2000, 2006; Plonsky and Gass 2011; Plonsky 2013, 2014; Loewen *et al*. 2014).

Among calls for reforms such as these, Larson-Hall (2010, 2012) and Larson-Hall and Herrington (2010) have made a case for a statistical procedure called 'bootstrapping'. This technique, one of a set of 'robust' statistics, is a Monte Carlo resampling procedure designed to simulate sampling distribution that could otherwise only be obtained from a much larger *N* (Efron 1979; Lee and Rogers 1998; Keselman *et al*. 2008; Beasley and Rogers 2009). By doing so, bootstrapping is argued to enable analyses that are not adversely affected by (i.e. are robust to) low statistical power and nonnormal distributions (Wilcox 2001), both of which are common in L2 research (Phakiti 2010; Plonsky 2013). Thus, the rationale for proposing this procedure is motivated by both general statistical theory as well as the nature of quantitative data found in most L2 research. Although the arguments in favor of this technique are compelling, its

potential benefits are an empirical matter. This article examines the potential value of bootstrapping in this field by applying the technique to raw data obtained from 26 published studies of applied linguistics research.

The literature review that follows begins with a brief introduction to and rationale for using bootstrapping in the context of applied linguistics. We focus on the procedure's potential to mitigate the effects of two limitations commonly observed in the field: (i) small samples/low statistical power and (ii) nonnormal data. We then review simulation studies from other fields that, like the present study, sought to explore the potential of this analytical procedure to provide more accurate results. Throughout our review and the study that follows, we have sought to minimize the technical aspects of bootstrapping in order to provide a more conceptual and accessible discussion.

LITERATURE REVIEW

Jenifer Larson-Hall and Richard Herrington's (2010) paper in *Applied Linguistics* introduced and recommended bootstrapping as an alternative to traditional parametric analyses such as *t* tests and ANOVAs. Their rationale for proposing this procedure was based on both statistical theory and the potential of bootstrapping to overcome several problems facing quantitative data analysis in applied linguistics research. This recommendation was also timely, coinciding with accumulating evidence and concern in the field for other methodological issues such as practical vs. statistical significance, instrument reliability, and reporting practices (e.g. Plonsky 2013).

Another such issue and problem, which is mitigated using bootstrapping and which has received increased attention in recent years, is the lack of statistical power resulting from the small samples typical of L2 research. For example, in a methodological synthesis of 174 studies of research in the interactionist tradition of second language acquisition, Plonsky and Gass (2011) observed an average sample of only 22 participants. Plonsky (2013), likewise, found group *N*s to average only 19 across 606 primary L2 studies published in *Language Learning* and *Studies in Second Language Acquisition*. Both studies also extracted effect sizes from primary reports and calculated post hoc power at only 0.56 and 0.57, respectively.

In an ideal world, a priori power would be calculated and used to gauge and obtain a sufficiently powered sample (i.e. one that is likely to detect a statistically significant effect, if present). However, this practice is not always feasible and is extremely rare in L2 research (see Plonsky 2013). Bootstrapping provides an alternative to relying on underpowered samples. The procedure randomly resamples with replacement—typically thousands of times—from the already-observed data set, producing an estimated sampling distribution from which descriptive and test statistics may be calculated. Rather than relying on a single set of observations, the researcher is then able to estimate the level of statistical significance for numerous samples. In this way, the data obtained via bootstrapping provides greater power by simulating a data set

we might obtain if we had replicated an experiment or resampled numerous times (Lee and Rogers 1998; Larson-Hall and Herrington 2010).[1]

A second problem commonly observed in L2 research and potentially mitigated via bootstrapping is nonnormally distributed data. Quantitative L2 researchers rely heavily on means-based analyses such as *t* tests and ANOVAs (Lazaraton 2005; Gass 2009; Plonsky 2013). Like other parametric procedures, these tests assume that the data being analyzed conform to a normal distribution, a condition often violated (e.g. Phakiti 2010) or left unchecked (Plonsky 2013). By resampling repeatedly, bootstrapped results are less sensitive to irregularities such as outliers, thus providing descriptive and test statistics that are robust to deviations from normality in the original sample.

Further supporting the use of robust statistics described here is the finding that, given a normally distributed data set, robust statistics such as bootstrapping have been found to approximate their parametric equivalents in power and accuracy; given a nonnormal distribution, bootstrapped analyses are much more powerful and therefore more likely to either detect statistical significance when present or to reveal a statistical relationship as spurious (Tukey 1960; Lee and Rogers 1998; Lansing 1999; Wilcox 2001; Larson-Hall and Herrington 2010; Tongbai *et al*. 2010).

A counter argument to the need to employ bootstrapped analyses could be made based on the claim that ANOVA and other means-based analyses are robust to violations of assumptions such as nonnormally distributed data, unbalanced *N*s, and unequal variance across groups (Lansing 2004; but cf. Keselman *et al*. 2008). Statements to this effect are commonly found in introductory statistics textbooks in applied linguistics (Bachman 2004) as well as other domains (Field 2005). According to Larson-Hall and Herrington (2010), however, these claims are only valid with respect to Type I error (i.e. incorrectly rejecting the null hypothesis). If this is true, we would expect the bootstrapped results of the present study to uncover possible Type II errors (i.e. incorrectly failing to reject the null hypothesis) in the results of published studies that have used conventional *t* tests and ANOVAs.

Small samples/low power and nonnormal data have been described as two common problems that bootstrapping may partially overcome. These related conditions result in a threat to the validity of quantitative L2 research and further highlight the need for alternative procedures such as bootstrapping, which is argued to mitigate the negative effects of all these conditions simultaneously.

Simulation research

The proposed advantages of bootstrapping as described in the previous section have been tested in a number of empirical studies. This type of 'simulation' research, found in a variety of fields ranging from nutrition to neuroscience, can be one of two types: primary or secondary. In a primary simulation, the

researcher collects data to address one or more research questions. However, in these studies both parametric and bootstrapped statistical tests are run and then compared. In secondary simulations such as the current study, researchers collect raw data from previous studies that have used parametric tests. The researchers then run bootstrapped versions of the original tests to compare the results of the two techniques considering, again, the different parameters of the sample and test results. The findings in this area are both limited and mixed, providing a lack of compelling evidence in favor of or against the use of bootstrapping for research in applied fields (e.g. Lansing 2004; Wolfe and McGill 2011).

We know of only one simulation study in applied linguistics. Larson-Hall and Herrington (2010) used real, unpublished data to compare the results of parametric and bootstrapped analyses (i.e. a primary simulation). The study compares native speakers' ($n = 15$) pronunciation with that of three groups of learners: A ($n = 14$), B ($n = 15$), and C ($n = 15$). Following an ANOVA, which revealed a statistically significant main effect across groups, Tukey post hoc tests showed a statistically significant difference between the native speakers and Group A ($p = .002$) but not Group B ($p = .407$) or Group C ($p = .834$). However, bootstrapped post hoc tests yielded a statistically significant difference for all three groups (A, $p < .0001$; B, $p = .01$; C, $p = .01$). These results indicate that the nonstatistical p values resulting from the parametric tests were due to a lack of power and that these (true) differences could only be found with increased statistical power simulated via bootstrapping. In this case, the bootstrapped analyses provide evidence of a Type II error in the original, parametric analysis. (See Wolfe and McGill 2011 for another example of a primary simulation study that found a lower Type II error rate for bootstrapped analyses.)

Bootstrapping has also been shown to reduce erroneous rejection of the null hypothesis (Type I error). In the field of cognitive neuroscience, Di Nocera and Ferlazzo (2000), for example, examined within-subject reliability for event-related potentials, running both traditional and bootstrapped ANOVAs. Whereas the results for parametric ANOVAs found a statistically significant difference between conditions, bootstrapped ANOVAs did not find this effect to be reliable across the sample of participants.

Despite evidence in favor of bootstrapping for reducing Type I and II errors, such results have not been obtained uniformly across simulation studies. Lansing's (2004) bootstrapped results based on data collected from very small samples (ns = 6–8) replicated those obtained for conventional ANOVAs. She concluded that, despite its potential to reduce Type I error, bootstrapping is not appropriate for very small samples and is not necessary for larger samples which are more likely to meet or approximate the assumptions required for parametric tests. A similar finding was also obtained by Welch *et al.* (1998). Their study found virtually no difference between bootstrapped and parametrically derived descriptives (mean, standard error). Finally, like the two previous studies, the findings of Tongbai *et al.* (2010) do not support the benefits of

bootstrapping either. Their study is also the only (quasi-)secondary simulation study we are aware of. The researchers reanalyzed MANOVA results from one primary data set and 12 additional test data sets created to simulate different conditions/distributions. Under nonnormal conditions, bootstrapped analyses lacked power and did not replicate the statistically significant effect obtained by the parametric test. For normally distributed data sets, bootstrapping and conventional tests achieved the same result. Neither case provides evidence in favor of bootstrapping as a useful alternative to parametric analyses.

Sharing data

One final issue bears discussing before moving on to our study. It is not so much statistical in nature but, rather, has to do with the culture (or the lack thereof) of data sharing in applied linguistics. It was not our original intention to examine this issue. However, it took on greater prominence as the study developed, largely because of our reliance on the willingness of primary researchers to share their data sets with us. Furthermore, we see this matter tying closely to concerns over transparency, raised recently in the context of meta-analysis where researchers also rely heavily, if not exclusively, on thorough reporting practices.

Responding to missing data in primary reports, a number of synthesists and meta-analysts have called for improved reporting of basic descriptive statistics such as standard deviations, which are used to calculate the Cohen's *d* effect size (e.g. Norris and Ortega 2000; Oswald and Plonsky 2010; Plonsky 2011, 2013). Others have contacted primary authors directly to request missing data, a practice met with mixed results. For example, approximately one-third of the requests made by both Plonsky (2011) and Lee *et al.* (2014) resulted in the provision of missing data. This response rate is disappointing considering most requests simply asked for descriptive statistics. The analyses for the present study, by contrast, required primary authors to submit raw data sets.

For a variety of reasons, many researchers in applied linguistics may be reluctant to sharing their raw data. We are not unique in this respect. Wicherts *et al*. (2006) were interested in reanalyzing data from published research in psychology to examine the sensitivity of reported findings to outliers. They sent data requests to 141 authors of papers published in four major journals of the American Psychological Association (APA) and received raw data sets from only 26 per cent. [Interestingly, subsequent analysis revealed that researchers who were more willing to share their data produced stronger evidence and were less likely to have committed errors in reporting their results (Wicherts *et al*. 2011).] Similar findings have been observed in other fields as well including medicine, which is generally considered methodologically rigorous and favorable toward data sharing (Reidpath and Allotey 2001). In addition to nonconformation to policy, these cases also contradict two of the

five Mertonian norms of science: Communalism and Organized Skepticism (see Merton 1973).

Despite apparent resistance to data sharing, there seems to be a gradual shift underway toward greater transparency in the social sciences (Firebaugh 2007; Wicherts *et al*. 2011). In applied linguistics we see signs that researchers are beginning to view their work 'as a shared community endeavor' (Abbuhl 2012: 146), such as the IRIS Database, the improved status of replication research (see Porte 2012), freely accessible and searchable corpora [e.g. The Corpus of Contemporary American English (COCA), Davies 2008; Spanish Learner Language Oral Corpora 2 (SPLLOC2); MacWhinney's TalkBank; the Multimedia Adult ESL Learner Corpus (MAELC), Reder *et al*. 2003], and the provision of online space for publishing supplementary materials by journals such as *Applied Linguistics* and *Language Learning*. Perhaps most encouraging is the recent resolution of the Linguistic Society of America (LSA) on data sharing (see Puschmann 2010), which stated:

> '. . . that the Linguistic Society of America encourages members and other working linguists to:
>
> - make the full data sets behind publications available, subject to all relevant ethical and legal concerns;...
> - when serving as reviewers, expect full data sets to be published (again subject to legal and ethical considerations) and expect claims to be tested against relevant publicly available datasets.'

Outside of applied linguistics, we see even stronger indicators of progress in this area. Most notably, according to Article 8.14 of the APA's (2010) Ethical Principles of Psychologists and Code of Conduct, a document signed by all authors of APA journal articles, 'After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis . . .' See also similar policies in political science (Meier 1995), personality/psychology (Lucas and Donnellan 2013), the journal *Science* (http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail), and the American Economic Association (http://www.aeaweb.org/aer/data.php).

Regardless of any evidence of and policies favoring increasing transparency, the extent to which applied linguists are willing to share their data has yet to be examined systematically. Therefore, in addition to our assessment of bootstrapping, this study represents a first step toward examining transparency among applied linguists with respect to data sharing.

Research questions

RQ1: To what extent do the results of bootstrapped *t* tests and ANOVAs differ from those of traditional tests?

RQ2: To what extent are applied linguists willing to share their raw data for the purpose of reanalysis?

METHOD

Data collection

The data for the following analyses were recruited from quantitative L2 research. Following Plonsky (2013), candidate studies published 2000–12 were identified from two L2 journals: *Language Learning* and *Studies in Second Language Acquisition*. In addition to the year and journal, we considered only those studies that conducted one or more means-based statistical tests (*t* tests and univariate ANOVAs). These criteria identified 255 potential studies.

Data from all 255 studies were solicited via emails sent to the corresponding author. The emails we sent described the goals of the project and identified which publication was of interest. We asked the authors if they would be willing to participate by sending us the raw data for the first means-based statistical test in their study.[2] Table 1 displays the wide variety of responses to our email request. From the 255 studies that were initially identified as possible participants, 37 data sets (14.5 per cent) were received from the authors of those studies and one of the studies reported the raw data in the publication.

In order to confirm that we had the correct data, we compared the descriptive statistics and statistical test results from each of the 37 data sets to the results reported in the published study. During this phase of the project, 11 studies were eliminated for one of two reasons (see Table 1). First, the results of the reanalyses of the summary statistics and the statistical tests that were provided were not always the same as in the original study. In many cases, this occurred because the shared data set contained cases that had been eliminated from the original analysis. Secondly, a number of shared data sets were formatted in a way that was indecipherable, and the data for group comparisons could not be readily identified. The results of this preliminary screening left 26 data sets (10 per cent of the original 255) that could be used for the bootstrapping analysis. The final sample of 26 studies consisted of eight independent samples *t* tests, six paired samples *t* tests, and 12 one-way ANOVAs.

Because *t* and *F* tests assume a normal distribution, we examined each sample in the 26 data sets for normality and for the presence of outliers. The former was assessed using both the Kolmogorov–Smirnov and the Shapiro–Wilk tests. Data sets were coded as nonnormal when indicated as such by one or both tests. Outliers were identified as cases equal or greater than 2.58 standard deviations from the mean (99th percentile; Field 2005) and were calculated as a percentage of the total sample. For example, if two values had a z-score of more than 2.58 in a sample of 40, the study was given an outlier score of 5 per cent. In addition, the effect sizes of the differences between the two groups (Cohen's *d*) were calculated if they were not reported in

Table 1: Data requests and response types

Step	Studies	Per cent
No response	159	62
Response, no data provided	60	24
Could not find or no longer has data	35	14
Not comfortable sharing data	3	1
Promised to send data (but failed to do so)	5	2
No time to find and send data	3	1
Other or no reason provided	15	6
Response, data sets provided	36	14
Inaccurate comparisons of summary or test statistics	6	2
Data provided, incomprehensible coding or formatting	5	2
Usable data sets	25[a]	10

[a]One additional data set reported in a candidate article was also included for a total of 26 usable data sets.

the original study. In cases where more than two groups were being compared, eta-squared was calculated and converted to Cohen's *d* for ease of comparison using the following formula:

$$d = \sqrt{\eta^2 \frac{n_1 + n_2 - 2}{1 - \eta^2} \left(\frac{n_1 + n_2}{n_1 n_2} \right)}$$

We used these effect sizes, along with sample size and an alpha level (α) of .05, to calculate and record the post hoc statistical power of each analysis.

Analysis

This section will give a brief overview of the bootstrapping procedures we used, the statistical tests that were analyzed, and examples of the code we used for the bootstrapping procedure. While any number of statistical analyses can be bootstrapped (e.g. means and standard errors, correlation coefficients), this study is limited to bootstrapped analyses of independent and paired samples *t* tests, and *F* tests for one-way ANOVAs. The results of bootstrapped analyses give researchers an indication of the degree to which the results of their original tests may contain misfits, or cases in which Type I or Type II error might be present.

There are two types of misfits: one is associated with Type I error and the other with Type II error. A Type I misfit occurs when bootstrapped results suggest that the null hypothesis has been falsely rejected, and a Type II misfit occurs when the bootstrapped results suggest that the null hypothesis has been falsely retained. However, it is important to point out that while bootstrapping

may produce more accurate estimations of population distributions based on samples, these are only estimations. Thus, what we label Type I and Type II misfits are not actually Type I and Type II errors. Rather, misfits give us evidence for or against the validity of nonbootstrapped statistical inferences.

In order to investigate the shared data sets for misfits, bootstrapping was used to produce simulated test statistics. One method of bootstrapping consists of generating a simulated *p* value by randomly resampling from the entire data set (i.e. ignoring the original groupings of the experimental units) and examining the ratio of bootstrapped test statistics that are as extreme, or more extreme, than the original (Chernick 2008; Larson-Hall and Herrington 2010). Similar to a permutation test, with this method the results from the original sample are interpreted as significant if the simulated *p* value is below .05, indicating that the original test statistic is significantly larger than what we would expect by chance alone.

A second method for coming to conclusions about bootstrapped test statistics is to randomly resample from within each of the groups, calculate a test statistic at each iteration, and use the bootstrapped test statistics to create a confidence interval (Efron and Tibsharani 1993; Chernick 2008). Using this approach, the researcher rejects the null hypothesis if the mean of the bootstrapped test statistics is so large that the hypothesized null value (0 for *t* tests, 1 for *F* tests) falls outside of the corresponding confidence interval. The resulting confidence intervals from the bootstrapped analyses can then be compared with the test statistic in the original study. These comparisons give insight as to whether or not a misfit has occurred. According to Efron and Tibsharani (1993), simulated *p* values and bootstrapped confidence intervals will return similar results. However, one benefit of using the confidence interval approach is the information that can be extracted from confidence intervals, such as the amount of variance.

The bootstrapped confidence intervals in this study use the second method described above. They were calculated by (i) randomly resampling with replacement from the 2+ groups in a given dataset; (ii) calculating the appropriate test statistic (*t* or *F*); and (iii) repeating steps (i) and (ii) 10,000 times, recording the resulting test statistic at each resampling. A confidence interval was then constructed for each test statistic which contained the middle 95 per cent of the 10,000 test statistics. The goal of constructing these bootstrapped confidence intervals is to answer this question: 'If we repeatedly and randomly resample from two or more groups, will we see large differences between them at least 95 per cent of the time?'

The bootstrapped analyses were computed using R: A Language and Environment for Statistical Computing version 2.15.1 (R Core Team 2012). The analyses were conducted using the boot package version 1.3-4 (Davison and Hinkley 1997; Canty and Ripley 2012). The boot package allows for a large variety of bootstrapping procedures. The procedures that best fit our needs were the ordinary nonparametric bootstrap and the bias-corrected and accelerated (BCa) confidence intervals, which are calculated using the boot.ci

function in the boot package. The results of these analyses are reported below. (See LaFlair, Egbert, and Plonsky, forthcoming, for a tutorial on using bootstrapping in L2 research.)

RESULTS

In this section of the article, we summarize the descriptive statistics for the data sets included in the analysis and present the results of the bootstrapping procedure. Details for each of the 26 studies that met our inclusion criteria are found in Table 2. In order to ensure the anonymity of the authors who contributed data, we report original sample sizes and *p* values in ranges rather than exact values.

The overall sample sizes for these data sets ranged from less than 15 to several hundred, with a median of 46. An α of .05 was either stated or assumed for all studies based on the convention of the field. The *p* values for the full data set ranged from <.001 to .999, and approximately two-thirds of the original test statistics ($k = 16$) achieved significance at $p < .05$.

Cohen's *d* was used to measure the standardized mean difference in each study. The effect sizes (Cohen's *d*) for the 10 nonsignificant tests ranged from 0.003 to 0.63; the significant tests had *d* values that ranged from 0.47 to 3.29.

Post hoc statistical power was calculated based on the observed *N*, effect size, and α of .05. Observed power ranged from 0.03 to 0.99, with a median of 0.70 and a mean of 0.62 (standard deviation = 0.39, 95 per cent confidence interval [0.46, 0.78]). Although overall power was fairly high, the likelihood of obtaining a statistically significant result in individual studies ranged widely across the sample.

The tests for normality revealed that 18 of the 26 data sets were not normally distributed. This phase of the analysis also found outliers in 11 studies with the portion of outliers ranging from approximately 1 to 6 per cent of their respective samples.

Table 2 contains the bootstrapped confidence intervals for each of the 26 test statistics. It is important to note that confidence intervals are interpreted differently for *t* tests and *F* tests. Whereas statistical significance for a *t* test is manifested when '0' falls outside of the confidence interval, we determine significance for *F* tests by observing that '1' is outside of the confidence interval. The *t* statistic is calculated by dividing the mean difference between two sample means by the standard error of the mean difference, so the expected value is 0 if H_0 is true. The *F* statistic, on the other hand, is the ratio of the variance between treatments to the variance within treatments. Therefore, the expected value for *F* is 1 if there is no difference between the variances (i.e. H_0 is true).

The final column in Table 2 labels the cases where a 'misfit' was identified. For the purposes of this study, misfits occur when the results of the original hypothesis test and the bootstrapped results do not agree. A Type I misfit suggests the rejection of a true null, and a Type II misfit suggests the retention of a false null. As can be seen from the results in Table 2, none of the nine

Table 2: Descriptive statistics, original significance levels, and bootstrapped test results

Study	Test type	N	Distribution	Outliers[a] (per cent)	Power[b]	Significance[c]	d[d]	Bootstrapped confidence interval[e]	Misfit
1	Two sample *t* test	30–59	Normal	0	0.65	*	0.76	[−4.26, −0.47]	—
2	Two sample *t* test	>90	Nonnormal	2	0.03	n.s.	0.02	[−2.10, 1.88]	—
3	Two sample *t* test	30–59	Normal	0	0.32	n.s.	0.52	[−0.56, 3.66]	—
4	Two sample *t* test	30–59	Nonnormal	0	0.1	n.s.	0.25	[−2.67, 1.32]	—
5	Two sample *t* test	>90	Nonnormal	3	0.99	**	0.83	[−4.81, −0.98]	—
6	Two sample *t* test	30–59	Nonnormal	0	0.51	n.s.	0.63	[−4.21, 0.08]	—
7	Two sample *t* test	30–59	Nonnormal	0	0.99	***	1.43	[3.32, 6.42]	—
8	Two sample *t* test	<30	Nonnormal	5	0.08	n.s.	0.28	[−2.69, 1.87]	—
9	Paired samples *t* test	<30	Normal	0	0.46	**	1.08	[0.36, 6.09]	—
10	Paired samples *t* test	60–89	Normal	0	0.75	***	0.88	[−5.74, −1.89]	—
11	Paired samples *t* test	30–59	Nonnormal	6	0.09	n.s.	0.22	[−1.21, 3.18]	—
12	Paired samples *t* test	30–59	Nonnormal	3	0.04	n.s.	0.06	[−2.57, 1.53]	—
13	Paired samples *t* test	<30	Normal	0	0.5	**	1.06	[−6.51, −0.24]	—
14	Paired samples *t* test	<30	Normal	0	0.62	**	1.63	[1.42, 7.19]	—
15	One-way ANOVA	>90	Nonnormal	1	0.99	*	0.47	[0.03, 17.06]	Type I
16	One-way ANOVA	60–89	Nonnormal	0	0.05	n.s.	0.003	[$1.21e^{-10}$, $2.28e^{-5}$]	—
17	One-way ANOVA	>90	Normal	1	0.99	**	0.54	[0.91, 24.46]	Type I
18	One-way ANOVA	>90	Nonnormal	1	0.99	***	3.29	[384.60, 591.40]	—
19	One-way ANOVA	>90	Nonnormal	3	0.99	n.s.	0.39	[0.01, 13.69]	—
20	One-way ANOVA	30–59	Nonnormal	0	0.99	*	0.72	[0.25, 21.03]	Type I
21	One-way ANOVA	30–59	Nonnormal	2	0.99	***	2.43	[9.73, 54.53]	—
22	One-way ANOVA	60–89	Nonnormal	3	0.99	***	1.82	[11.59, 61.24]	—
23	One-way ANOVA	30–59	Nonnormal	0	0.99	***	2.98	[12.54, 73.19]	—
24	One-way ANOVA	30–59	Nonnormal	0	0.13	n.s.	0.11	[0.00, 2.20]	—
25	One-way ANOVA	60–89	Nonnormal	0	0.99	**	0.7	[0.74, 24.50]	Type I
26	One-way ANOVA	60–89	Normal	0	0.99	***	1.05	[7.87, 48.96]	—

[a]Outliers as a percentage of the entire sample. [b]Calculated post hoc based on the sample, observed effect size, and $\alpha = .05$. [c]Level of statistical significance resulting from the original analysis: $*p < .05$, $**p < .01$, $***p < .001$, n.s. = $p > .05$. [d]Effect size (Cohen's d) based on the study's original results. [e]95 per cent confidence interval of t or F.

nonsignificant original test statistics resulted in Type II misfits. On the other hand, only 12 of the 16 (75 per cent) significant differences reported in the original studies were also found to be significant using bootstrapped methods. In the other four cases, the bootstrapped confidence intervals suggest that we should fail to reject the null hypothesis when the original reported p value was below the α criterion of 0.05. Based on these results, the use of bootstrapping revealed a Type I misfit rate of 25 per cent for the data that were reported as significant and an overall misfit rate of 15 per cent for the entire data set. We cannot assume that this sample of studies or these bootstrapped results might generalize to the entirety of L2 research. However, these error/misfit rates are much higher than expected assuming an a priori α of 0.05 wherein the expected Type I error should be approximately 5 per cent. The evidence from this admittedly small data set suggests that the actual Type I error rate may be as much as five times higher than we would expect using traditional parametric statistics such as t tests and ANOVAs.

A closer look at the Type I misfit rate across the three statistical test types reveals that all four of the misfits occurred with one-way ANOVAs, even though ANOVAs represent just over half of the originally significant tests. It is not initially apparent why the bootstrapped results contradict the original ANOVA results more than the t test results. We might have expected that the four misfit ANOVAs were originally based on small sample sizes. However, this does not appear to be a likely explanation, as the overall sample size for each of the four misfit ANOVAs was greater than 55, with only one sample group of less than 30.

These results are not as surprising, however, once we examine the original p values and effect sizes or the four misfit ANOVA results. Although the original test statistics from these four studies resulted in statistically significant results at $p < .05$, the original p values for three of the four misfit tests were among the five highest p values from the originally significant data. Additionally, while the effect sizes for the four misfit tests ranged from $d = 0.54$ to $d = 0.72$, with a mean of $d = 0.66$, which many would interpret as a medium-to-large effect size (Cohen 1988; Plonsky and Oswald under review), the original effect sizes for these results were among the six lowest effect sizes from the original statistical results. In other words, in the case of the four misfit ANOVAs, it seems that the *a priori* α levels in these studies were not sufficient to control for Type I error rates. Furthermore, although the effect sizes in the four misfit studies might be interpreted by some as medium-to-large, these effects are not large enough to achieve statistical significance when subjected to the robust statistical method of bootstrapping.

DISCUSSION

Bootstrapping

As the most frequent techniques for analyzing quantitative data in applied linguistics, the validity of study findings based on t tests and ANOVAs are of

great importance to the field. The potential of bootstrapping, proposed to help overcome threats to such findings such as small samples and nonnormal distributions, is therefore worth exploring. To that end and to address our first research question, we compared the results of 26 bootstrapped *t* tests and ANOVAs with their parametric equivalents as published in two top-tier journals in applied linguistics.

We found four instances of bootstrapped results contradicting those of the original reports. None of these bootstrapped analyses suggested an incorrect failure to reject the null hypothesis when it should have been rejected (i.e. Type II misfit). All four misfits, rather, were cases where the statistically significant result in the original report was not replicated in the bootstrapped analysis (i.e. possible evidence of a Type I error). The Type I misfit rate among the 16 studies with statistically significant results is therefore 25 per cent, five times what we would expect with α set to .05. This finding is also noteworthy in light of claims of ANOVA to be robust to violations of assumptions leading to Type I errors (e.g. Larson-Hall and Herrington 2010).

Looking further at the misfits, we see that all four occurred with ANOVAs and that none occurred with samples less than 30 (see Table 3). This finding contradicted our expectation of bootstrapping to improve on the limitations of small samples and low power as shown in the example above from Larson-Hall and Herrington (2010). Rather, this finding aligns with Lansing (2004) who found nearly identical results for parametric and bootstrapped ANOVAs based on very small samples. The present study supports her conclusion that because bootstrapping resamples with replacement from the originally observed values, bootstrapped analyses based on very small samples are similarly limited in their ability to represent the population (Mooney and Duval 1993). To be sure, bootstrapping is not a replacement or cure-all for inadequate sampling.

We would modify Lansing's conclusions, though, proposing that this finding may only hold true when the effect sizes are larger and therefore more easily detected in tests of statistical significance. We also see that all four of the misfits occurred among larger samples (>30) and two of the four occurred among the largest *N*s (>90). Furthermore, as shown in Table 3, two of the six studies in the largest sample size grouping resulted in Type I misfits. The relationship between sample size, power, and statistical significance may help in explaining this finding: given a large enough sample, a statistically significant effect can *always* be obtained no matter how small the mean difference between groups. Or, as Tukey (1991: 100) stated, 'The effects of A and B are always different—in some decimal place—for any A and B. Thus asking ''are the effects different?'' is foolish'. This point is clearly exemplified in our post hoc power analyses: observed power for all four Type I misfits was 0.99.

These findings suggest a tension as to how an appropriate sample size should be determined. The results of this study demonstrate that very large samples may overestimate the importance of an effect by simply yielding a *p* value of less than .05. At the same time, however, very small *N*s, which are common in L2 research, lack power and are therefore less likely to reliably identify effects

Table 3: Frequency of sample sizes and misfits

Total *N*	No misfit	Misfit
>30 ($k=4$)	4	0
30–59 ($k=11$)	10	1
60–89 ($k=5$)	4	1
>90 ($k=6$)	4	2

as statistically significant. Consider, for example, the following rank-ordered effect sizes of the 10 nonstatistically significant results as observed in the original reports: $d=0.003$, 0.02, 0.06, 0.11, 0.22, 0.25, 0.28, 0.39, 0.52, 0.63. Although we do not know about the variables, instruments, or participants behind these effect sizes, we would consider at least six of them to be substantial and practically significant according to either Cohen's (1988) or Oswald and Plonsky's (2010) benchmarks for interpreting *d* values. By relying exclusively or even primarily on *p*, these six studies have likely all overlooked potentially meaningful and important results. More concerning still is the accumulation of such results and misinterpretations across multiple studies on a common topic, leading to a blunt underestimation of effects. Gelman and Weakliem (2009) refer to this practice as a Type M (for magnitude) error, a notion closely related to the author and editorially rooted bias in favor of publishing statistically significant results usually discussed in the context of meta-analysis (see also Norris and Ortega 2006; Ioannidis 2008; Plonsky and Oswald 2012).

Based on the concerns brought to light in this study and throughout the meta-analytic literature, we echo the now-growing chorus of applied linguists who have considered the relationship between sample size, power, and effect size along with the inherent limitations of *p* values (Crookes 1991; Lazaraton 1991; Norris and Ortega 2000, 2006; Plonsky 2009; Larson-Hall 2010; Oswald and Plonsky 2010; Brown 2011; Plonsky 2011; Plonsky and Gass 2011; Plonsky 2012; Plonsky and Oswald 2012; Plonsky 2013). As an alternative and way forward, we provide the following brief recommendations for researchers in applied linguistics: (i) at the design stage of a study, an appropriate sample size should be informed by an a priori power analysis based on effect sizes in previous studies or meta-analyses; (ii) at the interpretation stage, the outcome of statistical tests should be considered as a function of the relationship between their observed effect, sample size, and the presence of sampling error; and (c) rather than relying on the flawed notion of statistical significance, the focus of quantitative results should be on the effect sizes and their practical significance with respect to theory and/or practice.

To summarize our position here, the field has much to gain from reducing our reliance on the use of null hypothesis significance testing; a greater focus on descriptive statistics—namely means, standard deviations, confidence

intervals, and effect sizes—along with visual presentations of data would move the field forward much more efficiently by providing more reliable and accurate estimates of relationships.

We now return to the larger question posed by this study as to whether and under what conditions bootstrapping should be employed in applied linguistics. Our findings indicate that, yes, the field stands to gain from the addition of bootstrapping to the researchers' repertoire of quantitative analyses. We are not suggesting that *t* tests and ANOVAs be uniformly replaced by their bootstrapped equivalents. Rather, there are conditions under which applied linguistics research may benefit from bootstrapping in addition to more thorough data analytic practices such as those described in the preceding paragraphs.

Like Larson-Hall and Herrington (2010), we recommend the use of bootstrapping when one or more assumptions of parametric tests are violated. Of course doing so requires researchers to examine their data for these features. Plonsky's (2013) review of 606 studies published in the same journals sampled for the present study found only 17 per cent reporting to have checked statistical assumptions. And only 3 per cent were found to have done so in the 174 studies of L2 interaction reviewed by Plonsky and Gass (2011). We also recommend bootstrapping in addition to parametric analyses when a study has either very low or high statistical power. In the case of the former, bootstrapping, though not a replacement for adequate sampling, may reveal statistical effects that were not revealed by parametric tests (a Type II misfit); bootstrapping the latter, as the results of the present study show, can help control for Type I error by showing statistical results in the original analysis to be spurious and the result of a large sample rather than a large effect or mean difference. Recall that this pattern was observed in 15 per cent of the entire sample and 25 per cent of the sample when $p < .05$ in original reports. When no misfit is observed, the results can be considered more reliable than when based on a single test; when a misfit is observed, it is the author's responsibility to explain why this might be and to reconcile the apparent contradiction with respect to the observed data, the relationships under investigation, and the study's procedures and instruments. In either case, and whenever bootstrapping is employed, the results of both analyses must be presented. The results from multiple reanalyses can then be combined synthetically to produce secondary assessments of the value of bootstrapping in L2 research.

Data sharing

The secondary aim of this study was to examine applied linguists' willingness to make their data available for reanalysis. Based on the response rates reported in L2 meta-analyses and from similar studies in other fields, we did not anticipate receiving data from a large portion of the authors we contacted. Our results confirmed these expectations. Data sets were received from 36 authors (14 per cent of the original 255 authors contacted), 25 of which (10 per cent of the original 255) were usable.

Many of the data sets we requested were collected over a decade ago and we recognize both their limited shelf life and the effort required to annotate them sufficiently for reanalysis by secondary researchers. These impediments were made clear to us in authors' responses and indirectly in the rate of nonresponse (see Table 1). By comparison, Wicherts *et al.* (2006) received data sets from 26 per cent of the 141 authors they contacted in psychology. Neither of these results comes close to meeting the requirement to share data for the purpose of reanalysis as established by the APA and LSA, and there are several steps the field might take to improve practices related to transparency and data sharing.

Based on our results, we agree with Wicherts *et al.*'s (2006) recommendation that primary data sets be published as online supplements to primary studies. Not only would such a standard enhance replication and meta-analytic research, but it would also promote the synthetic ethic embodied by both of these approaches (see Norris and Ortega 2006; Abbuhl 2012; Plonsky 2012; Porte 2012). Furthermore, published data sets might also lead to more justifiable analyses and data reporting practices, both of which lack transparency (see Plonsky and Gass 2011; Plonsky 2013). Finally and most central to the present study, requiring publication of raw data sets would greatly facilitate the work of methodologists interested in reanalyzing primary data to better understand the nature of L2 data or explore alternative techniques such as bootstrapping.

Future directions and conclusion

As an initial foray into the application of bootstrapping in applied linguistics, the results of this study point to the need for additional research in several areas. First, there is a need for further bootstrapped analyses of primary data to determine if and under what conditions misfits occur. This research can be carried out in much the same fashion as the present study. As seen in the results to our second research question, however, this approach is limited by researchers' willingness to share their data. Nonetheless, we are hopeful that the field will continue to move toward greater transparency. Future secondary simulations would demonstrate the extent to which a shift toward openness has occurred. As primary simulations accumulate, the need to solicit data from individual researchers may also diminish and published primary simulations could be combined via secondary analysis.

This study was limited in the range of statistical analyses addressed. Although *t* tests and ANOVAs are by far the most frequent tests employed in applied linguistics (Gass 2009), procedures exist for bootstrapping other statistics (e.g. correlation coefficients, multiple regression) and merit empirical attention as well (e.g. Beasley *et al.* 2007).

The extent to which applied linguistics research is able to benefit from this type of inquiry depends on individual researchers as well as journal editors. The former must continue to devote their time and effort to addressing issues such as bootstrapping and other larger methodological concerns such as statistical versus practical significance and reporting practices.

The latter group, journal editors, can encourage and enhance methodological reform in at least two ways. First, editors can make space for empirical and position papers on methods, both of which are currently scarce in applied linguistics journals. And secondly, editorial policy can be used to shape the methods used by researchers. The inclusion of effect sizes in *Language Learning* following Ellis' (2000) editorial is a prime example of the impact that editorial policy can have. Likewise, if raw data sets are to be made available via online supplements, it is the responsibility of editors to require authors to submit them before publication or, ideally, along with the first submission of a manuscript. Finally, recognizing that editors and regular manuscript reviewers cannot be expected to possess expertise in all methodological practices, journals might consider following Magnan's (1994) now-expired policy at *The Modern Language Journal* of soliciting a methodological review for all papers that advance beyond the initial external review. Top-down reforms such as these have enormous potential to improve the means by which applied linguistics research is carried out and, thus, our ability to accurately inform theory and practice.

Finally and as always, no degree of statistical sophistication should ever take the place of principled analysis and interpretation based on an understanding of the data and the constructs they represent. It will always be important for researchers to take a step back from the statistical analysis to evaluate the degree to which a particular technique is practically significant/useful in moving forward our knowledge of a given set of constructs.

ACKNOWLEDGEMENTS

This article would not have been possible without the assistance of the authors who took the time and effort to locate and share their raw data with us. We are extremely grateful for their generosity. We would also like to thank Jenifer Larson-Hall for introducing the field of applied linguistics to bootstrapping and for her very thoughtful comments on an earlier draft of this manuscript.

NOTES

1 We do not mean to propose bootstrapping as a means to eliminate the need for rigor at the sampling phase of a study. Likewise, bootstrapping is not to be seen as an alternative to collecting large (i.e. sufficiently powered) samples, random sampling, or random assignment, particularly when a study seeks to draw experimental inferences. It is all too true that 'you can't fix through analysis what you bungled by design' (Light *et al.* 1990, p. viii). Our position in this article is simply that, when ideal sampling conditions cannot be met, bootstrapping may assist researchers in making the most of their data.

2 In a number of cases, the first test was conducted to ensure comparability of groups.

REFERENCES

Abbuhl, R. 2012. 'Practical methods for teaching replication to applied linguistics students' in G. Porte (ed.): *Replication Research in Applied Linguistics*. Cambridge, pp. 135–50.

APA. 2010. Ethical principles of psychologists and code of conduct. Washington, DC: Retrieved 18 April 2013 from http://www.apa.org/ethics/code/index.aspx.

Bachman, L. F. 2004. *Statistical Analyses for Language Assessment*. Cambridge University Press.

Beasley, W. H., L. DeShea, L. E. Toothaker, J. L. Mendoza, D. E. Bard, and **J. L. Rodgers.** 2007. 'Bootstrapping to test for nonzero population correlation coefficients using univariate sampling,' *Psychological Methods* 12: 414–33.

Beasley, W. H. and **J. L. Rogers.** 2009. 'Resampling Methods' in R. E. Millsap and A. Maydeu-Olivares (eds): *The SAGE Handbook of Quantitative Methods in Psychology*. Sage.

Brown, J. D. 2011. 'Quantitative research in second language studies' in E. Hinkel (ed.): *Handbook of Research on Second Language Teaching and Learning*. vol. 2. Routledge.

Canty, A. and **B. Ripley.** 2012. boot: Bootstrap R (S-Plus) Functions. R package version 1. 3–4.

Chernick, M. R. 2008. *Bootstrap Methods: A Guide for Practitioners and Researchers,* 2nd edn. John Wiley & Sons.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences,* 2nd edn. Erlbaum.

Crookes, G. 1991. 'Power, effect size, and second language research. Another researcher comments,' *TESOL Quarterly* 25: 762–5.

Davies, M. 2008. 'The Corpus of Contemporary American English: 450 million words, 1990-present,' available at http://corpus.byu.edu/coca/.

Davison, A. C. and **D. V. Hinkley.** 1997. *Bootstrap Methods and Their Applications*. Cambridge University Press.

Di Nocera, F. and **F. Ferlazzo.** 2000. 'Resampling approach to statistical inference: Bootstrapping from event-related potentials,' *Behavior Research Methods, Instruments, & Computers* 32: 111–19.

Efron, B. 1979. 'Bootstrap methods: Another look at the jackknife,' *Annals of Statistics* 7: 1–26.

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

Ellis, N. C. 2000. 'Editorial statement,' *Language Learning* 50: xi–xiii.

Field, A. 2005. *Discovering Statistics Using SPSS,* 2nd edn. Sage.

Firebaugh, G. 2007. 'Replication data sets and favored-hypothesis bias: Comment on Jeremy Freese (2007) and Gary King (2007),' *Sociological Methods & Research* 36: 200–9.

Gass, S. 2009. 'A survey of SLA research' in W. Ritchie and T. Bhatia (eds): *Handbook of Second Language Acquisition*. Emerald, pp. 3–28.

Gelman, A. and **D. Weakliem.** 2009. 'Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects,' *American Scientist* 97: 310–16.

Ioannidis, J. P. A. 2008. 'Why most discovered true associations are inflated,' *Epidemiology* 19: 640–8.

IRIS. A digital repository of data collection instruments for research in second language learning and teaching, available at http://www.iris-database.org/iris/app/home/index.

Keselman, H. J., J. Algina, L. M. Lix, R. R. Wilcox, and **K. N. Deering.** 2008. 'A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes,' *Psychological Methods* 13: 110–29.

LaFlair, G. T., J. Egbert, and **L. Plonsky.** forthcoming. 'Bootstrapping' in L. Plonsky (ed.): *Advancing Quantitative Methods in Second Language Research*. Routledge, pp. 135–50.

Lansing, L. 1999. *Bootstrapping versus the Student's t: The problems of type I error and power*. Unpublished master's thesis, Lehigh University.

Lansing, L. 2004. *A self-report and experimental study of situational factors involved in academic dishonesty using standard and bootstrapping analyses*. Unpublished doctoral dissertation, Lehigh University.

Larson-Hall, J. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS*. Routledge.

Larson-Hall, J. 2012. 'Our statistical intuitions may be misleading us: Why we need robust statistics,' *Language Teaching* 45: 460–74.

Larson-Hall, J. and **R. Herrington.** 2010. 'Improving data analysis in second language acquisition by utilizing modern developments in applied statistics,' *Applied Linguistics* 31: 368–90.

Lazaraton, A. 1991. 'Power, effect size, and second language research. A researcher comments,' *TESOL Quarterly* 25: 759–62.

Lazaraton, A. 2005. 'Quantitative research methods' in E. Hinkel (ed.): *Handbook of Research in Second Language Teaching and Learning*. Erlbaum, pp. 109–224.

Lee, W.-C. and **J. L. Rogers.** 1998. 'Bootstrapping correlation coefficients using univariate

and bivariate sampling,' *Psychological Methods* 3: 91–103.

Lee, J., J. Jang, and **L Plonsky.** 2014. 'The effectiveness of second language pronunciation instruction: A meta-analysis,' *Applied Linguistics.* doi:10.093/applin/amu040.

Light, R. J., J. D. Singer, and **J. B. Willet.** 1990. *By Design: Planning Research on Higher Education*. Harvard University Press.

Loewen, S., E. Lavolette, L. A. Spino, M. Papi, J. Schmidtke, S. Sterling, and **D. Wolff.** 2014. 'Statistical literacy among applied linguists and second language acquisition researchers,' *TESOL Quarterly* 48: 360–88.

Lucas, R. E. and **M. B. Donnellan.** 2013. 'Improving the replicability and reproducibility of research published in the Journal of Research in Personality,' *Journal of Research in Personality* 47: 453–4.

Magnan, S. S. 1994. 'From the editor: The MLJ tradition and the challenges ahead,' *The Modern Language Journal* 78: 7–9.

Meier, K. J. 1995. 'Replication: A view from the streets,' *PS: Political Science and Politics* 28: 456–8.

Merton, R. K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Mooney, C. Z. and **R. D. Duval.** 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage.

Norris, J. M. and **L. Ortega.** 2000. 'Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis,' *Language Learning* 50: 417–528.

Norris, J. M. and **L. Ortega.** 2006. 'The value and practice of research synthesis for language learning and teaching' in J. M. Norris and L. Ortega (eds): *Synthesizing Research on Language Learning and Teaching*. Benjamins, pp. 3–50.

Oswald, F. L. and **L. Plonsky.** 2010. 'Meta-analysis in second language research: Choices and challenges,' *Annual Review of Applied Linguistics* 30: 85–110.

Phakiti, A. 2010. 'Analysing quantitative data' in B. Paltridge and A. Phakiti (eds): *Continuum Companion to Research Methods in Applied Linguistics*. Continuum, pp. 39–49.

Plonsky, L. 2009. 'Nix the null: Why statistical significance is overrated,' Paper presented at the Second Language Research Forum (SLRF), East Lansing, MI.

Plonsky, L. 2011. 'The effectiveness of second language strategy instruction: A meta-analysis,' *Language Learning* 61: 993–1038.

Plonsky, L. 2012. 'Replication, meta-analysis, and generalizability' in G. Porte (ed.): *Replication Research in Applied Linguistics*. Cambridge University Press, pp. 116–32.

Plonsky, L. 2013. 'Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research,' *Studies in Second Language Acquisition* 35: 655–87.

Plonsky, L. 2014. 'Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform,' *Modern Language Journal* 98: 450–70.

Plonsky, L. and **S. Gass.** 2011. 'Quantitative research methods, study quality, and outcomes: The case of interaction research,' *Language Learning* 61: 325–66.

Plonsky, L. and **F. L. Oswald** (2012). 'How to do a meta-analysis' in A. Mackey and S. M. Gass (eds): *Research methods in second language acquisition: A practical guide*. Basil Blackwell, pp. 275–95.

Plonsky, L. and **F. L. Oswald.** under review. 'How big is 'big'? Interpreting effect sizes in L2 research,' Manuscript under review.

Porte, G. (ed.) 2012. *Replication Research in Applied Linguistics*. Cambridge University Press.

Puschmann, C. 2010. Blog, available at http://blog.ynada.com/184. Accessed 18 April, 2013.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, available at http://www.R-project.org/.

Reder, S., K. Harris, and **K. Setzler.** 2003. 'The multimedia adult ESL learner corpus,' *TESOL Quarterly* 37: 546–57.

Reidpath, D. D. and **P. A. Allotey.** 2001. 'Data sharing in medical research: An empirical investigation,' *Bioethics* 15: 125–34.

Spanish Learner Language Oral Corpora 2 (SPLLOC2) Available at http://www.splloc.soton.ac.uk/splloc2/index.html.

Tongbai, R. R., F. Yu, and **K. M. Miller.** 2010. 'Multivariate nonparametric techniques for astigmatism analysis,' *Journal of Cataract & Refractive Surgery* 36: 594–602.

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. 448–85.

Tukey, J. W. 1991. 'The philosophy of multiple comparisons,' *Statistical Science* 6: 100–16.

Welch, W. W., D. Huffman, and **F. Lawrenz.** 1998. 'The precision of data obtained in large-scale science assessments: An investigation of bootstrapping and half-sample replication methods,' *Journal of Research in Science Teaching* 35: 697–704.

Wicherts, J. M., M. Bakker, and **D. Molenaar.** 2011. 'Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results,' *PLoS One* 6: 1–7.

Wicherts, J. M., D. Borsboom, J. Kats, and **D. Molenaar.** 2006. 'The poor availability of psychological research data for reanalysis,' *American Psychologist* 61: 726–8.

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

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

Wolfe, E. W. and **M. T. McGill.** 2011. 'Comparison of asymptotic and bootstrap item fit indices in identifying misfit to the Rasch model,' Paper presented at the National Conference on Measurement in Education, New Orleans, LA.