International Journal of Listening
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijl20

The Statistical and Methodological Acuity of Scholarship Appearing in the International Journal of Listening (1987-2011)

Shaughan A. Keaton [a] & Graham D. Bodie [b]
[a] Communication Studies Department , Young Harris College
[b] Department of Communication Studies , The Louisiana State University
Published online: 11 Sep 2013.

To cite this article: Shaughan A. Keaton & Graham D. Bodie (2013) The Statistical and Methodological Acuity of Scholarship Appearing in the International Journal of Listening (1987-2011), International Journal of Listening, 27:3, 115-135, DOI: 10.1080/10904018.2013.813206

To link to this article: http://dx.doi.org/10.1080/10904018.2013.813206

The Intl. Journal of Listening, 27: 115–135, 2013

ISSN: 1090-4018 print / 1932-586X online
DOI: 10.1080/10904018.2013.813206

Routledge
Taylor & Francis Group

ARTICLES

The Statistical and Methodological Acuity of Scholarship Appearing in the *International Journal of Listening* (1987–2011)

Shaughan A. Keaton
Communication Studies Department
Young Harris College

Graham D. Bodie
Department of Communication Studies
The Louisiana State University

This article investigates the quality of social scientific listening research that reports numerical data to substantiate claims appearing in the *International Journal of Listening* between 1987 and 2011. Of the 225 published articles, 100 included one or more studies reporting numerical data. We frame our results in terms of eight recommendations to improve future listening scholarship. In particular, the results suggest needed variation in demographics and added attention to psychometric properties of scores. Standards for reporting and inspecting data should also be followed with more regularity, and tests of statistical assumptions along with information about missing data are urged. Effect sizes are rarely included in results, and no studies reported confidence intervals, suggesting overreliance on null hypothesis statistical testing when drawing implications for practice. Lastly, there were some noteworthy misappropriations of statistical techniques that are discussed.

Research is conducted to advance knowledge. Without research, the field of listening (like other fields) cannot progress, and our ability to inform practice is diluted. One way to advance knowledge and improve the practice of listening is through social scientific research, goals of which include the accurate description and explanation of how and why people listen to particular

Correspondence concerning this article should be addressed to either author: Shaughan A. Keaton, Communication Studies Department, Young Harris College, 1 College St., Young Harris, GA 30582. E-mail: findingmulder@gmail.com; Graham D. Bodie, Department of Communication Studies, The Louisiana State University, 136 Coates Hall, Baton Rouge, LA 70803. E-mail: gbodie@lsu.edu

others in specified contexts and relationships (see Bodie, 2009).[1] Such research has the *potential* to contribute greatly to theory building and practical efforts. We stress potential here because the larger body of social scientific research on listening is only as good as the individual studies constituting its existence. As such, this manuscript investigates the quality of one type of social scientific listening research important to the development of the field, namely studies reporting numerical data to substantiate claims.[2]

While narrative reviews outlining the implications of quantitative listening research for theory and practice exist (e.g., Bodie & Fitch-Hauser, 2010), no know assessment of its quality has been undertaken.[3] Periodically assessing issues of quality is vital to an accurate assessment of the larger field, and a systematic investigation can only benefit future work. Because the *International Journal of Listening* (*IJL*) is arguably the most notable scholarly outlet for the publication of listening research, we take the recent celebration of its silver jubilee as a chance to assess issues of quality in quantitative social scientific listening research.

Like other modes of inquiry, quantitatively oriented social scientific research is highly conventional. There are rather straightforward, formulaic procedures that, regardless of discipline, have been implemented, usually upon the informed recommendation of some scholarly body (Levine, 2011). Although blind adherence to convention is not encouraged, consistency in reporting is. Stability in reporting allows for comparability of results and a more coherent body of knowledge that can lead to broader claims than any one individual study can ever hope to accomplish. The leading authority on "best practices" for making principled arguments with numerical data is the American Psychological Association (APA); the *IJL* suggests following the recommendations of the APA publication manual, now in its sixth edition. Therefore, this study was undertaken to assess the degree to which the reporting practices of articles in the IJL conform to (and vary from) standard practices as defined by the APA as a function of time.

It is important to note that best practices for reporting data fluctuate as a function of time. For instance, only within the last two decades has it been standard to report effect sizes and confidence intervals. Likewise, standard practices shift due to advancements in statistical knowledge and the development and refinement of new techniques like structural equation modeling and multilevel modeling. Thus, when appropriate, we report trend data to assess whether the *IJL* has increased its keenness in statistical and methodological reporting over the time as these standards have evolved.

METHOD

Sample

The data for this project consisted of studies published in the 32 issues of the *International Journal of Listening* distributed between 1987 and 2011. Of the 225 total articles, 100 (44.4%)

[1] We wish to point out here the focus of this article is on one branch of research on listening, and although large, it in no way constitutes the entirety of the tree. As such, this article essentially "ignores" other, equally valid conceptualizations (Gehrke, 2009). The interested readers are directed to the following sources for alternative conceptualizations of listening: Beard (2009), Ihde (2007), Lipari (2009, 2010), and Purdy (2000).

[2] Our focus on research reporting numbers in no way asserts a preference for this type of research; it merely reflects the expertise of the authors. We encourage others to engage in similar projects focusing on separate classes of research.

[3] We cautiously utilize the term "quantitative" here and urge readers not to interpret us as supporting the quantitative-qualitative dichotomy (Bodie, 2011).

utilized some sort of quantitative reporting method. Of those articles, 12 presented more than one study for a total of 112 studies. In these 112 studies, two were meta-analyses, which were removed for a total of 110 studies in this examination.[4] Because studies and not articles report statistics, it is these 110 studies that are the focus of our analyses.

Procedure

Two independent coders decided first whether each of the 240 published studies in these issues reported numbers ($n = 110$; 45.8%) or not ($n = 130$; 54.2%); the former were then randomly assigned to these two coders for analysis. After each coder finished all assigned analyses, all data were double checked for accuracy by the second author; disagreements were resolved through discussion.

Due to low sample sizes for individual issues (which for the first 20 years corresponded exactly with volume/year), we split the data into five, five-year intervals for analysis: α) 1987–1991 ($n = 16$; 33.3%); β) 1992–1996 ($n = 29$; 70.7%); γ) 1997–2001 ($n = 27$; 77.1%); δ) 2002–2006 ($n = 15$; 35.7%); and ε) 2007–2011 ($n = 23$; 31.1%). Per convention, alpha was set to .05, and our total sample size was 110 out of 240 possible studies (45.8%). For differences among the five, five-year groups, power to detect small effects ($f = .10$) was .11, power to detect moderate effects ($f = .25$) was .51, and power to detect large effects ($f = .40$) was .93. For bivariate correlations, power to detect small effects ($r = .10$) was .18, power to detect moderate effects ($r = .30$) was .90, and power to detect large effects ($r = .50$) was above .99. For differences between two independent means and degrees of freedom at 57, the sample had power to detect small effects ($d = .20$) of .12, power to detect medium effects ($d = .50$) of .47, and power to detect large effects ($d = .80$) of .86. Power to detect effects with the Wilcoxon rank sum test was assessed with the same levels for small, medium, and large effect sizes and exceeded .95 for all three. Therefore, if the effects discovered in this study are small, Type II error is probable in the case of nonsignificant findings when using bivariate correlations and t-tests.

The publication manual of the American Psychological Association, currently in its sixth edition (2010), outlines a host of best practices for the reporting of research in article form. We chose to focus on the variables listed in Table 1 in order to compare our results with those offered by assessments of other sets of journals likely of interest to the readership of the *IJL* (Cohen, 1994; Levine & Hullett, 2002; Meline & Wang, 2004; Sun & Fan, 2010).

RESULTS

Of the 110 studies included in our analyses, 23 developed principled arguments by reporting only descriptive statistics; an additional 87 also utilized inferential statistics. Because practices for reporting methods are pertinent to all of the studies, we will discuss those procedures first, followed by practices for reporting results, which will examine the uses of descriptive and inferential statistics, respectively.

[4]Reporting conventions for meta-analytic reviews are remarkably different from those for individual studies.

TABLE 1
Type of Data Collected as Recommended by APA

Category	*Variables*
Methods reporting	Race/ethnicity, level of education, location, age, sex, participants
Research design	Post-facto, experimental, longitudinal
Measurements and measurement types	Self-report, assessment, evaluation, other-report, observational coding
Psychometrics	Reliability estimates, measurement models, inter-rater reliability estimates
Sampling	Random, non-random, intact classroom, convenient, simple, snowball, stratified, sample size
Results reporting	All relevant significant and nonsignificant results, effect sizes
Descriptive	
Univariate	Central tendency (mean, median, mode), range, variability (variance, standard deviation), shape (kurtosis, skew)
Bivariate	Cross-tabs, scatterplots, measures of dependence (correlation), covariance, slope
Inferential	
Generalized linear models	Assumptions of normality to determine parametric or nonparametric statistics, missing data, effect sizes, confidence intervals
Bivariate	Correlation (*Spearman's ρ, Pearson's r), regression
Simple multivariate	Multiple regression, logistic regression, ANOVA, MANOVA, *Kruskal-Wallis, ANCOVA, discriminant analysis, main effects, interactions, post-hoc tests
Full multivariate	Canonical correlation, MANOVA, MANCOVA, multivariate regression, EFA, CFA, PCA, SEM
Differences in central tendency or expected/observed	t-test, *Mann-Whitney-Wilcoxon rank-sum test, $^{*}\chi^2$

Note. Common non-parametric estimations are designated with an asterisk.

Methods Reporting

APA recommends reporting sample characteristics as specifically as possible, listing important descriptors such as race/ethnicity, level of education, location, age, and sex. The research design should be described in detail, discussing the operational procedures for collecting data and the design for doing so and all measurements and measurement types including psychometric properties of scores.

Demographics

Demographics reported across studies are presented in Table 2. Of the 12 studies that reported race, the mean numbers of Caucasian participants per study (237.46, $SD = 268.83$) far outnumbered African American ($M = 28.63$, $SD = 16.89$) and Hispanic participants ($M = 16.5$, $SD = 13.47$). Likewise, the location distribution suggested studies favored U.S. samples ($n = 91$, 82.73%). Other locations utilized included Germany ($n = 8$), Taiwan ($n = 3$), South Korea ($n = 2$), Japan ($n = 2$), Australia ($n = 2$), China ($n = 1$), Malaysia ($n = 1$), Pakistan ($n = 1$), Hong Kong ($n = 1$), Saudi Arabia ($n = 1$), Canada ($n = 1$), and Indonesia ($n = 1$).

TABLE 2
Demographic Reporting

	Yes		*No*	
Characteristic	*n*	%	*n*	%
Race[α]	12	10.9	93	84.6
College students[α]	74	67.3	31	28.2
Class rank[β]	12	16.2	62	83.8
U.S. sample[γ]	91	82.7	17	15.5
Age[α]	39	34.5	66	60.0
Biological sex[α]	57	51.8	48	43.6

[α]Five studies were not applicable to discussions involving human participants.

[β]Only 74 studies reported class rank.

[γ]In two studies it was not apparent from where the participants originated.

The ratio of studies that used college students to those that did not is graphically displayed in Figure 1, showing a relative imbalance during each five-year interval. Of the 74 studies that used college student samples, only 12 reported class with freshmen ($M = 157.75$; $SD = 179.67$) and sophomore ($M = 116.00$; $SD = 167.89$) students outnumbering, on average, juniors ($M = 48.75$; $SD = 54.94$) and seniors ($M = 59.00$; $SD = 61.14$).

Reflecting a bias toward college student samples, the mean age of the samples reported ($n = 39$) was 22.9 ($SD = 6.57$) with reported ranges ($n = 26$) between 17.5 and 43.6 years. For the 11 studies reporting a standard deviation with mean age, the average value was 3.67 ($SD = 2.49$). Across the five-year intervals, neither the means for the age, $F(4, 20) = 0.98$, $p < .44$, $\eta^2 = .16$, nor the standard deviations, $F(2, 8) = 0.07$, $p < .93$, $\eta^2 = .02$, were significantly different (Figure 2).[5]

On average, across the samples, there was not a significant disparity in the sex composition of participants ($M_{female} = 149.05$, $SD = 167.13$; $M_{male} = 130.24$, $SD = 194.04$; $t(57) = 0.56$, $p < .58$, Cohen's $d = .10$). The differences in the means of biological sex across five-year intervals can be observed in Figure 3. Upon visual inspection, it appears that the variation is most apparent in the last ten years. Therefore, we estimated differences across the intervals with the Mann-Whitney-Wilcoxon rank sum test due to small, nonnormal samples: 1987–1991 ($n = 8$, $z = 1.47$, $p < .14$); 1992–1996 ($n = 15$, $z = 0.63$, $p < .53$); 1997–2001 ($n = 11$, $z = 0.62$, $p < .53$); 2002–2006 ($n = 10$, $z = 2.80$, $p < .01$); and 2007–2011 ($n = 14$, $z = 2.08$, $p < .04$). These data imply that differences in the sex distribution of participants remained roughly equal for the first 15 years of the journal and deviated in the last 10 years, with female participants outnumbering male participants during this latter time period and a particularly large discrepancy during the five-year period between 2002 and 2006.

[5]Because the sample size of those studies that reported age and standard deviation did not meet the requirements for detecting small effects, a Kruskal-Wallis test was performed in each instance indicating that the mean ranks of each category did not vary significantly across five-year intervals, supporting the non-significant ANOVA results: average age means, $H(4) = 2.35$, $p < .67$, and mean standard deviations, $H(2) = 0.614$, $p < .73$. The listing of age descriptives varied with 25 (64.1%) reporting means, 26 (39.4%) age ranges, and 11 (16.7%) including standard deviations.

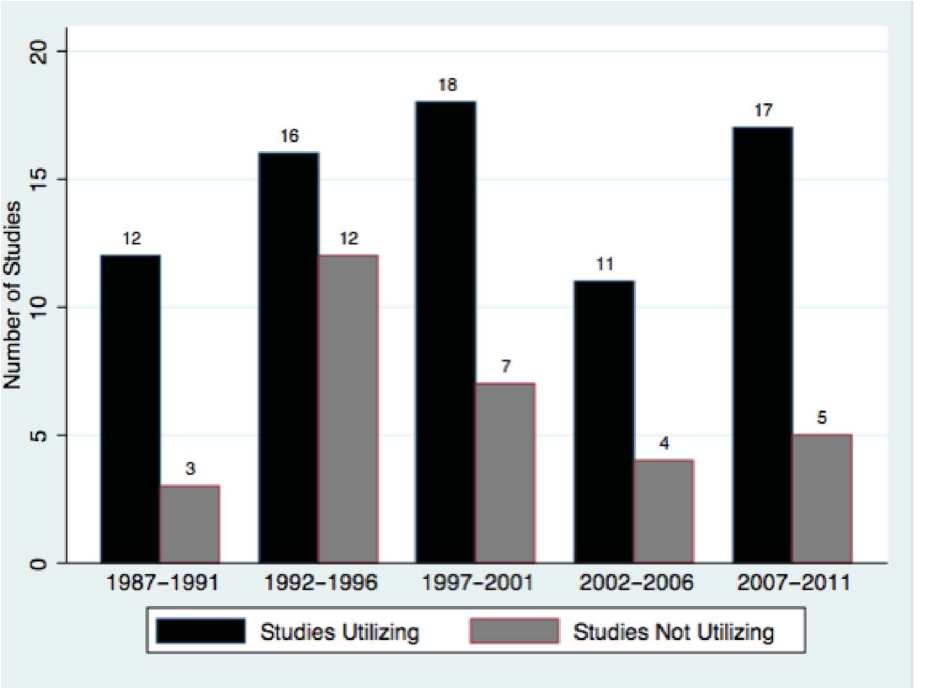

FIGURE 1 Use of college students per five-year interval (color figure available online).

Research designs and measures

The majority of studies utilized cross-sectional, post-facto ($n = 61$, 55.5%) designs with experimental research constituting approximately one-third of studies ($n = 38$, 34.6%). Even less prominent were longitudinal studies ($n = 7$, 6.4%).

Operational measure types included self-report ($n = 67$, 60.9%), assessment ($n = 37$, 33.6%), evaluation ($n = 17$, 15.5%), other-report ($n = 13$, 11.8%), and behavioral coding ($n = 5$, 4.6%). Reporting psychometrics of scores was not consistent for any of the designs: Self-report (measurement models: $n = 15$, 22.4%; reliability estimates: $n = 33$, 49.3%), assessment (measurement models: $n = 6$, 16.2%; reliability estimates: $n = 18$, 48.7%), evaluation (measurement models: $n = 8$, 47.1%; reliability estimates: $n = 8$, 47.1%), and other-report (measurement models: $n = 6$, 46.2%; reliability estimates: $n = 7$, 53.9%). Psychometrics should have been reported in each of these instances. It is clear that more consistency is needed in this area. In addition, the average reported alpha value also suggests measures are barely acceptable by commonly utilized criteria, only appropriate for beginning stages of research (total factors/components/constructs reported = 312, mean alpha = .71, $SD = .19$; see Nunnally, 1978). Finally, three out of five studies that utilized observational coding reported a necessary measure of inter-rater reliability.

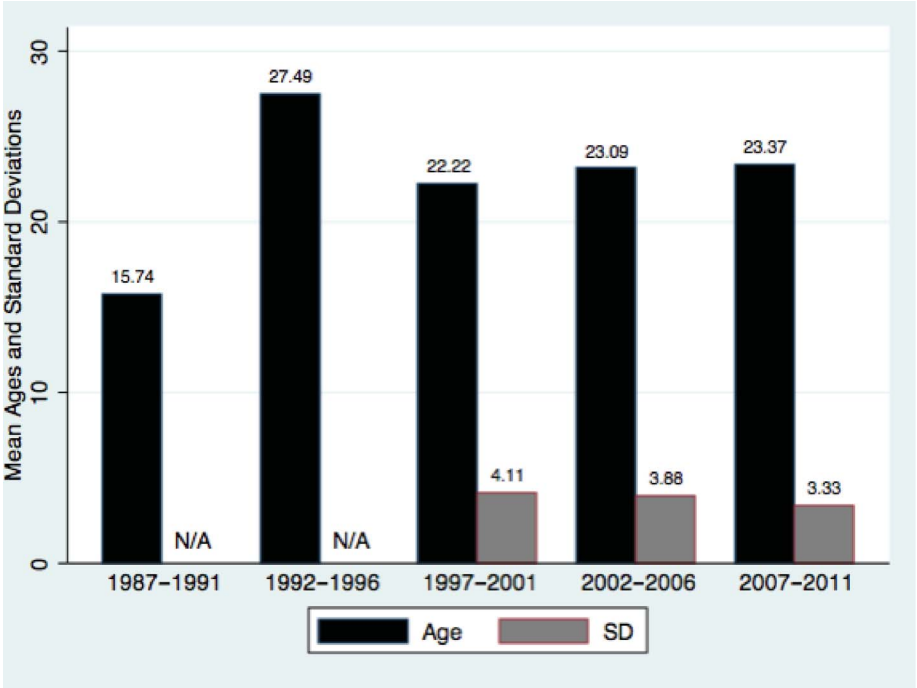

FIGURE 2 Mean age and standard deviations per five-year interval. The mean age for 1987–1991 is skewed because of a single study that utilized third grade students and reported a mean age of 8.8 years. With that study removed, the average age is 22.4 years (color figure available online).

Sample characteristics

The APA manual recommends that the type of sample should be reported (e.g., random, nonrandom, intact classroom). The types of data samples documented included convenient, simple, random, snowball, stratified, and the use of intact classes. Of the 110 studies, 100 (90.9%) used convenience samples. Other techniques included simple sample ($n = 1$, 0.9%), random sample ($n = 7$, 6.4%), snowball sampling ($n = 1$, 0.9%), and stratified sampling ($n = 2$, 1.8%). In addition, 28 studies utilized intact classes (25.5%), with 12 of those used in experimental designs (31.6% of all experimental studies).

When using inferential statistics and some types of descriptive statistics, the size of a researcher's sample can alter the outcome of the statistical analysis. It is especially crucial to report sample size to contribute to a study's ability to be replicated and to report the capacity to detect various effects—especially in the occasion of results not found to be statistically significant. In some instances, it may be that there are effects in the population but the sample is not large enough to detect them. In our total sample of studies ($n = 110$), sample size was not reported twice so they were removed for this section. A test for normality was performed with the mean distribution positively skewed (9.93, $p < .000$). Upon examination of a box plot for statistical

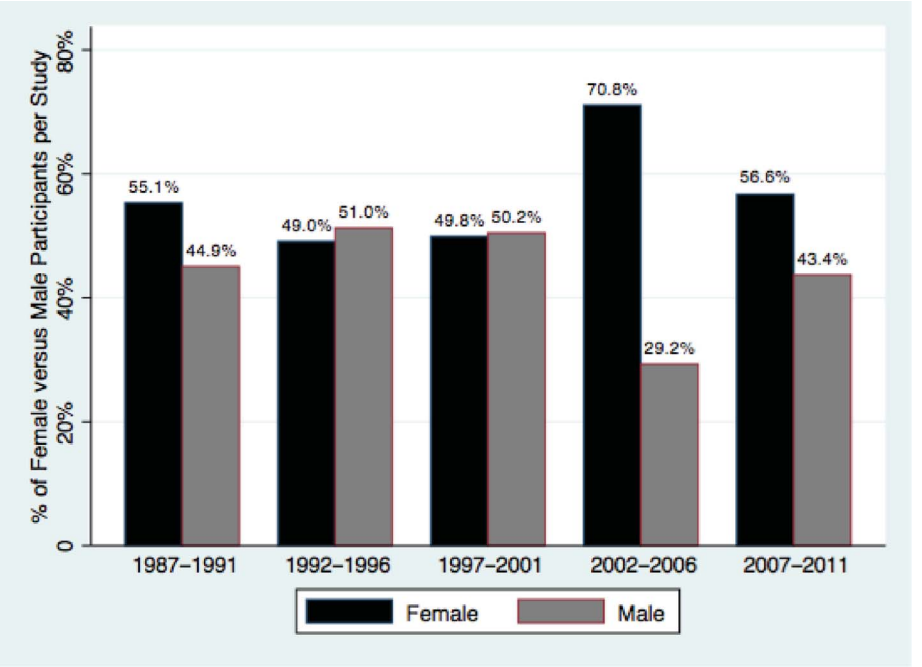

FIGURE 3 Sex distribution across five-year intervals (color figure available online).

outliers—which is a technique applicable to nonnormal samples because of its utilization of the median instead of a mean—an additional 10 were deleted for a total of 98 (overall mean sample size = 159.87, SD = 119.93).[6] These outliers were deleted for the purposes of this particular section because there were several samples that numbered in the tens of thousands and grossly inflated mean values (and for these studies, power is not an issue, at least for committing Type II error). An analysis of variance was estimated to investigate whether or not the mean sample sizes differed significantly across five-year intervals, $F(4, 91) = 0.84$, $p < .50$, $\eta^2 = .03$. The results suggest that the means did not significantly fluctuate over time (see Figure 4). Out of all of the remaining 98 studies, 12 reported statistical power (12.2%), but APA did not recommend reporting statistical power until the fifth edition in 2001. Out of the studies appearing in IJL that use numbers to support claims from 2002 on ($n = 37$), only five reported power (13.5%) after the APA manual's fifth edition was published.

Results Reporting

APA recommends summarizing data and the analyses used to stimulate discourse. All relevant results, statistically nonsignificant results, and effect sizes should be included. There were two main types of results reported using numbers in *IJL*: descriptive and inferential.

[6]Calculating the median (ME = 160) when leaving the outliers in the distribution shows that the two measures of central tendency are approximately equal.

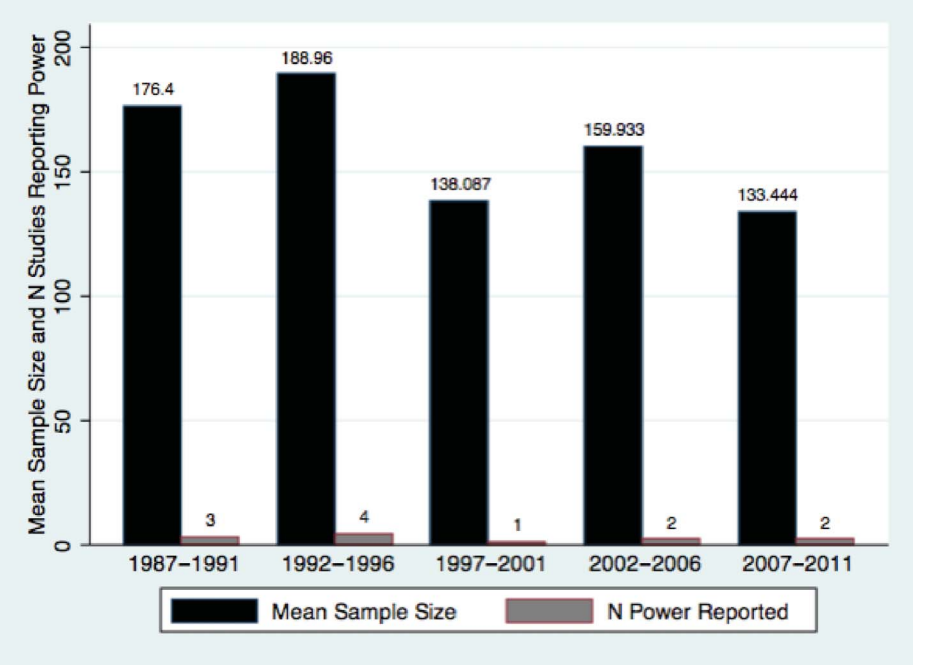

FIGURE 4 Sample size and power reporting across five-year intervals (color figure available online).

Descriptive statistics

Descriptive statistics, as differentiated from inferential, present summaries about the data sample involved in the analysis. Traditionally, descriptive statistics are not used to generalize from the sample gathered to a population from which data are derivative, and descriptive statistics are not founded in probability theory (The International Statistical Institute, 2003). However, descriptive statistics are often adequate for some examinations, and they tell an interesting story about data not available from inferential statistics.

Univariate descriptives include information about central tendency (mean, median, mode), range, variability (variance, standard deviation), and shape (skewness, kurtosis). As a method of developing principled arguments through descriptive statistics, central tendency (in the form of means) was used 47.8% of the time ($N = 23$). Variability was reported in the form of standard deviation three times (13.0%). No information about the shape of the distribution was reported for those using descriptive statistics to summarize data.

Bivariate descriptives comprise information in the form of cross-tabs, scatterplots, measures of dependence (Pearson's r when both variables are continuous, Spearman's ρ if one variable is discrete), covariance (which reflects measurement scales for the variables), and slope (a one unit change in the dependent variable for a one unit change in the independent variable). Only two studies used bivariate statistics (Pearson's r) in their descriptive analyses.

Inferential statistics

Statistical inference results from evaluating hypotheses to draw generalizable conclusions from gathered data to a larger population. Where descriptive statistics are simply presenting facts, statistical inference requires making assumptions that are based upon probability theory, allowing researchers to make predictions. When researchers can assert that their samples are approximately normally distributed, fully parametric statistical tests may be implemented. However, if a sample is nonparametric (not approaching normal or using only nominal or ordinal data), there are corresponding statistical tests that may be substituted. Using parametric statistics on non-normal samples can present a variety of issues. APA requires that testing for *assumptions of normality* be included when using inferential statistics and any *violations* of those assumptions. Next, procedures concerning the frequency or percentage of *missing data* should also be discussed, and it is important to know when to use the correct procedure to answer a research question or assess a hypothesis. Finally, null hypothesis significance testing (NHST) is only a beginning point and other reporting elements such as *effect sizes* and *confidence intervals* should also be incorporated (Levine, Weber, Park, & Hullett, 2008).

To make statistical inferences from the general linear model (GLM) which assumes normality, the following tests were reported (for an in depth review of the techniques, see Tabachnick & Fidell, 2007): For evaluating bivariate relationships amongst continuous variables, correlation ($n = 40$) and regression ($n = 11$) were common statistical tools; at the bivariate level these two statistics are equivalent. When adding continuous independent variables (IVs), simple multivariate (i.e., multiple regression) statistics fulfilled the need. However, when all of the IVs were discrete, simple multivariate analysis of variance (ANOVA; $n = 39$) was implemented. Main effects ($n = 31$, 79.5%) and interactions ($n = 20$, 51.3%) were intermittently reported. When estimating models where some IVs were continuous and others discrete, analysis of covariance (ANCOVA; $n = 2$) was estimated. For studies including a dichotomous dependent variable (DV) and continuous IVs, a discriminant analysis ($n = 3$) was implemented (which is, incidentally, the opposite of ANOVA), and if the DV was discrete and the IVs continuous and/or discrete, then logistic regression models could be the best choice (but none reported in this sample).

For full multivariate forms (multiple DVs) where all DVs and IVs were continuous, canonical correlation ($n = 5$) was applicable. If all of the DVs were continuous and all IVs were discrete, it was suitable to estimate multivariate analysis of variance (MANOVA; $n = 18$; single, $n = 11$; repeated measures, $n = 7$), multivariate analysis of covariance (MANCOVA; none reported), or for a mixture of discrete and continuous IVs, multivariate regression (none reported). When DVs and IVs were continuous, but the DVs were observed and the IVs latent, measurement models ($n = 28$, eight unidentified) such as factor analysis (exploratory, $n = 4$; confirmatory, $n = 3$) and principle components analysis (PCA, $n = 13$) were potential techniques. Relationships between latent variables require structural models; consequently when DVs and IVs were latent, then structural equation modeling (SEM) would have been desirable (but none were reported). Likewise, hierarchical models are appropriate when error terms are likely not independent, though none were reported.

When estimating differences in central tendency between variables, the following procedures were reported: When normality could be assumed, *t*-tests, ($n = 29$) were used. For nonparametric

samples, χ^2 ($n = 9$), Mann-Whitney ($n = 4$), and Kruskal-Wallis ($n = 1$) were utilized. Post-hoc tests for ANOVA and Kruskal-Wallis were reported 13 times (32.5%).

Finally, effect sizes were reported for 32 (29.4%) studies, and only three (3.4%) included details about confidence intervals. Eight studies (7.34%) reported procedures concerning missing data. For all studies, means ($n = 71$) and standard deviations ($n = 44$) were reported for comparisons across groups. However, out of all of the studies surveyed, only six reported tests for assumptions and only four reported that they satisfied those assumptions. These statistics mean that in many—if not most—of these studies, it is difficult to support statistical inferences drawn from the uses of these methods.

DISCUSSION

The purpose of this article is to assess the statistical and methodological acuity of social scientific research reported in the *International Journal of Listening*, particularly those studies that utilized numerical data to make principled arguments. Our sample included articles published in the first 32 issues, representing 25 years of scholarship, 1987–2011. While this Silver Jubilee is certainly a time to celebrate, it is also a time to critically reflect on what we actually know about listening. As Charles Peirce once said in a July 4 address, "it is usual enough to indulge . . . in self-glorification at our successes and it is equally useful to submit ourselves to a little self-humiliation at our shortcomings" (W 4.152). We agree. Below we first briefly review our findings then provide recommendations for improving listening scholarship.

Summary of Results

Interestingly, our findings suggest that the use of numerical data to make principled arguments about and advance knowledge of listening is not overutilized. Indeed, quantitative social scientific research represented 45.8% of all published material in the first 25 years of the *IJL*. Thus, we encourage future editors to continue this balance and to seek out various ways to understand how and why people listen in the ways that they do, whether through scientific, rhetorical, critical/cultural, or qualitative means.

In terms of presenting numerical data, it does not appear that reporting practices for numerical data within the *IJL* are consistent with recommendations of its guiding style manual. In particular, our data suggest that reporting basic sample characteristics like age, ethnicity, and biological sex distribution is not the norm. We also found that, contrary to APA recommendations, there is an overreliance on null hypothesis significance testing (NHST) with much less reporting of effect sizes and confidence intervals. Other reporting practices inconsistent with APA recommendations include the infrequent reporting of basic descriptive statistics (e.g., measures of central tendency and variability), a complete lack of focus on the shape of sample distributions, a tendency not to report tests relevant to statistical assumptions (e.g., normality), and a lack of clarity with regard to missing data. Lastly, there were some noteworthy misappropriations of statistical techniques that should be discussed at length. Given these concerns, we outline several recommendations and the rationale behind these recommendations.

Recommendation 1: Look at Your Data

There were no studies published in the first 25 years of the *IJL* that visually displayed data for inspection. This finding conflicts with recommendations provided by the Task Force on Statistical Inference, from which APA gleans much of the information for its style guide. The Task Force suggests the following:

> As soon as you have collected your data, before you compute any statistics, look at your data. Data screening is not data snooping. It is not an opportunity to discard data or change values to favor your hypotheses. However, if you assess hypotheses without examining your data, you risk publishing nonsense. . . .We are warned against fishing expeditions for understandable reasons, but blind application of models without screening our data is a far graver error. (Wilkinson, 1999, pp. 599–600)

Visual inspection of data is important for several reasons, and we will discuss two of the more essential reasons here.

First, descriptive statistics allow the researcher to see trends and patterns in data (Levine, 2011). For example, one would be hard pressed to find results from a criminologist merely testing whether a certain number of murders occurring in a particular area were different from zero (which is what the NHST methods used in the *IJL* ultimately do). Instead, most criminology data is arrayed using frequency distributions and histograms to show trends over time or to illustrate that aggregate crime statistics can be misleading and that there are, for instance, certain crime-prone areas of a city or state. Listening scholars can learn a great deal from criminologists and other social scientists who utilize descriptive statistics to make principled arguments. We believe the recommendation is true both for studies that present only descriptive data as well as for studies that also utilize inferential statistics. In both cases, descriptive data can be informative and can provide information beyond the story that inferential statistics can tell. There are excellent examples of Communication scholars who utilize descriptives to tell interesting stories and who regularly publish in the top journals (see Levine, 2011).

Second, visual inspection and related descriptive statistics can help researchers spot violations of one or more assumptions underlying an inferential procedure. Choosing the best test for respective research questions or hypotheses benefits the researcher and aids in replicability. The use of most of these statistical methods begins by assuming that their samples are normally distributed. When sample distributions are significantly skewed or suffer from positive or negative kurtosis, assumptions derived from statistics assuming normal distributions can be invalid. Many of these issues can be alleviated with a large enough sample size. For instance, in a large enough sample, skewness does not digress enough from normality to create noteworthy differences in analyses. Furthermore, positive ($n > 100$) and negative ($n > 200$) underestimates of variance associated with kurtosis wane with large enough samples (Tabachnick & Fidell, 2007).

Out of the 107 studies sensitive to normality assumptions, 39 (36.4%) had samples of fewer than 100, and 71 (66.4%) had fewer than 200. Out of the 39 studies using samples of less than 100, zero reported testing for normality. If the analysis is expanded to those with less than 200, a total of three studies reported assessing their samples (all three met the requirements). However, those with samples with fewer than 100 observations often used statistical tests that assume normality, such as Pearson's r ($n = 10$), regression ($n = 4$), ANOVA ($n = 15$), ANCOVA ($n = 1$), discriminant analysis ($n = 1$), MANOVA ($n = 4$), measurement models (EFA, $n = 1$; CFA, $n = 1$; PCA, $n = 3$), and t-tests ($n = 10$). None of those studies reported assumption testing. For

samples greater than 100 but fewer than 200, the statistical analyses included Pearson's r ($n = 13$), regression ($n = 4$), ANOVA ($n = 9$), discriminant analysis ($n = 1$), canonical correlation ($n = 3$), MANOVA ($n = 2$), and the use of measurement models ($n = 11$; PCA, $n = 5$; for all samples under 200, there were a total of eight unidentified measurement models), and t-tests ($n = 8$). Out of these, one tested for assumptions that used Pearson's r, one using ANOVA, and one using t-tests. The rest were not reported.

While many maintain that parametric statistical models may still be used providing the deviations from normality are not acute (Hubbard, 1978), especially given that many nonparametric tests lack versatility in multivariate situations (Nunnally, 1978), serious consequences still can result should the sample exhibit a distribution that is not close to normal. Using parametric statistics based on t, F, or χ^2 to generalize findings from sample distributions not approaching normal can, among other outcomes, compromise the estimation of coefficients and confidence intervals. Therefore, this study recommends testing for normality (both graphically and with descriptive statistics) and using these tests only after a principled case has been made.

Recommendation 2: Report Effect Sizes and Confidence Intervals

Few reputable scholars deny that the social and behavioral sciences are marked by an overreliance on NHST (Cohen, 1994). When engaged in NHST, the researcher is ultimately making a dichotomous judgment. Merely proclaiming statistical significance does not provide a complete picture of the results of a study. Real science is concerned with finding the magnitude of an effect, not with a dichotomous decision rule regarding whether the null (and usually nill) hypothesis is a valid assumption (Ziliak & McCloskey, 2009). Moreover, given a large enough sample, the conclusion drawn from inspecting a p-value is meaningless (Meehl, 1990). Indeed, merely reporting a p-value and claiming a result is "statistically significant" gives readers no indication as to the clinical or practical significance of the results, and failure to discuss the latter limits future attempts to replicate (or refute) results or to conduct meta-analyses.

A common misperception exists that probability values tell us something about the weight of an effect. They do not. Probability values only convey the likelihood of a Type I error (incorrectly rejecting the null). In other words, when a significant effect is detected in a sample, we can, with a particular degree of certainty, claim that this sample is not derivative from a population where this effect is statistically improbable. In this case it is no longer meaningful to retain the null, so it is rejected, allowing the researcher to reasonably claim that the sample in question can be generalized as representative of a population where an effect does in fact occur. Therefore, because probability values do not give us information concerning the size of an effect, reporting effect sizes and confidence intervals becomes essential. Without this additional information, we do not have a credible science of listening.

Effect sizes and confidence intervals also give consumers of scientific inquiry a basis for deciding if a study is practically significant rather than only statistically significant. Statistical significance does not necessarily imply that findings are of consequence, and nonsignificant results are not necessarily unimportant. Effect sizes and confidence intervals help deflate the overvalued importance of statistical significance and allow for nonsignificant findings that may have practical significance to see the light of day in journal space. Reporting effect sizes along with other descriptive statistics like measures of central tendency (e.g., means) and variability (e.g., standard deviation)—as well as the inferential probability that a population mean, for instance,

lies between two values (i.e., a confidence interval)—also allow researchers to double check the results of studies or to do meta-analyses of many studies. For results of one study to be compared to those of another and to ultimately build a cumulative body of knowledge, scholars need to know the practical and theoretical importance of findings and how these findings can be interpreted given what we already know. For practitioners to derive any set of best practices from scientific research, they need to appropriately know effect sizes and confidence intervals; that is, they need to be able to discern not the statistical but the practical significance of study results.

Recommendation 3: Psychometrically Validate Scores Derived from the Use of Instruments

For the total sample of studies, 28 reported some sort of measurement model or factor analysis. Given the large number of studies reporting data from the use of multi-item scales (e.g., LSP-16) or assessment tests (e.g., WBLT), this number is drastically low. For the most part, the non-use of factor analysis seems to stem from the assumption that existing scales, especially those with a rich history, have been "previously validated" and should thus be treated differently than newer or less established scales. As stated by Levine, Hullett, Turner, and Lapinski (2006), "this [belief] is unfortunate, and the view that a once-validated scale can or should be treated as an always-valid scale is neither reasonable nor consistent with good scientific practice" (p. 310). Not only is validity an ongoing process, suggesting that scales are not "valid or invalid" but can be said to have more or less robust validity portfolios, but scales often exhibit different properties when utilized with different populations (Little, 1997). As such, our recommendation is for authors to report the psychometric properties of data derived from the use of instruments, irrespective of the status of the instrument.

Our third recommendation is particularly important for instruments that are assumed to have vast validity portfolios. Two recent examples in the listening literature will illustrate this point. In a recent study, Bodie, Worthington, and Fitch-Hauser (2011) reported data inconsistent with the measurement model of the Watson-Barker Listening Test (WBLT), Form C. In particular, the reported data showed that items on the WBLT-C were largely uncorrelated with each other (r_{ave} = .03) and that no pattern of association among items could explain the small amount of shared variance that did exist. Ultimately, the WBLT-C consists of 40 unrelated multiple-choice items.[7] A similar project is underway to assess the LSP-16 with results suggesting major modifications of the scale are needed (Bodie & Worthington, 2010) and when made result in a much more potentially valid scale (Bodie, Worthington, & Gearhart, 2013). Ultimately, listening researchers are warned not to be like the "innovators" described by the Task Force on Statistical Inference:

> Innovators, in the excitement of their discovery, sometimes give insufficient attention to the quality of their instruments. Once a defective measure enters the literature, subsequent researchers are reluctant to change it. In these cases, editors and reviewers should pay special attention to the psychometric properties of the instruments used, and they might want to encourage revisions (even if not by the scale's author) to prevent the accumulation of results based on relatively invalid or unreliable measures. (Wilkinson, 1999, p. 598)

[7]The authors are fully aware that newer versions, Forms D and E, are now commercially available. We warn against their use in studies, however, without proper inspection of psychometric properties of the data prior to reporting results. The results presented with respect to Form C are not uncommon, as similar results were reported for earlier versions (e.g., Villaume & Weaver, 1997; Fitch-Hauser & Hughes, 1987).

Another noteworthy problem concerns the use of scales exhibiting reliability estimates that do not meet recommended criteria. The mean α of all of the samples was .71, which is problematic because this is a commonly regarded "minimum value" of "acceptable" internal consistency. Customary evaluative criteria are often in the range of $0.7 \leq \alpha < 0.8$ for acceptable values with $0.6 \leq \alpha < 0.7$ deemed questionable (George & Mallery, 2003; Kline, 1999). Regardless of the source, however, higher values of internal consistency are universally recognized as more desirable than lower values. Not only do low levels of internal consistency attenuate relationships between variables and differences between groups, but they are vitally important when making practical recommendations from studies (Nunnally, 1978, is still the best source for interested readers).

Most listening scholars ultimately want their research to be useful, to help themselves or others improve the lives of the everyday people about whom we theorize and for whom our work should be targeted. Listening scholars who create or utilize instruments which produce low estimates of internal consistency are like the doctor who declares a patient has an incurable disease based on the results of a test that produces more false positives than true scores. We would certainly be unconvinced of any medical advice stemming from tests not backed by empirical data, and we recommend that listening educators and practitioners not prescribe listening treatments if the only evidence for their effectiveness comes from data using instruments that have not been adequately vetted. Whatever the chosen evaluative method, it is clear that listening researchers need to take greater care in operationalizing listening constructs.

Recommendation 4: Correctly Utilize and Report Factor Analytic Techniques

Continuing our discussion of measurement models from Recommendation 3, of the 28 reports of factor analysis only three utilized confirmatory factor analysis (CFA). The two more commonly utilized procedures were Exploratory Factor Analysis (EFA) and Principle Components Analysis (PCA). EFA attempts to discern an underlying structure of latent variables by grouping observed variables that are correlated. EFA utilizes only shared variance, while PCA uses all of the variance in the data. Consequently, the resultant factors from EFA are thought to be explanatory mechanisms. Components gleaned from PCA, however, are only descriptive—not inferential—groupings of associated items and are not thought to be explanatory or causal. Thus, researchers must know their specific goals (research questions and hypotheses) to choose the correct procedure because using PCA to deduce factors is an inefficient and incorrect use of the method. It is more appropriate to use PCA to reduce the number of items in exploratory scale development (see Park, Dailey, & Remus, 2002).

Of the 13 studies reporting PCA, 11 (84.6%) reported a desire to deduce an explanatory factor structure. In one instance, the use of PCA was deemed appropriate, and in another the outcome variable was dichotomous. Therefore, 92.3% of the occasions where results were reported using PCA were inappropriate or problematic. Of the four studies using EFA, two used it appropriately, while two others also used dichotomous outcome variables (more aptly estimated with another technique, such as logistic regression). The three instances of CFA were suitable for the designs in which they were implemented.

Last, and perhaps most challenging, eight studies did not report the type of "factor analysis" used, so it was unclear the technique utilized. However, given that PCA is the default setting on most statistical software (e.g., SPSS), it may be likely that PCA was used in these scenarios. Given that it was not reported, replication of these studies would prove difficult.

Given the above, we recommend that when the goal is to develop underlying explanatory frameworks of latent, correlated variables that EFA and not PCA is a more fitting method, and reporting the specific procedure is essential in aiding the replicability of the study. When the goal is to report the psychometric properties of scores derived from established measures, then CFA is the preferred technique.

Recommendation 5: Match Sampling to Population of Interest

The next major point of interest concerns the homogeneity of the overall sample of participants used in studies reported in the *IJL*. A typical participant is a 23-year-old female, white college-student (usually a freshman or sophomore) from the United States. In an international journal, this observation poses a concern about the title of the journal at best, and at worst indicates that the bulk of what we know about listening is about how young Americans listen; we are not convinced our knowledge to date is generalizable to other cultural contexts. Although there has been some research in Europe, Asia, and the Middle East, 82.7% of the studies utilized participants residing in the US. The youthful age of participants has not varied over the course of time (Figure 2), and although sex remained somewhat consistent over the first 15 years of the journal, in the last decade the difference between the number of male and female participants has significantly diverged (Figure 3). Although white, middle-class, college-educated, young Americans are a viable population from which to learn about basic structures and functions of listening (see Shapiro, 2002), resting our entire knowledge base on this single population is certainly curious. To truly be an international journal, it is clear that something needs to be done to encourage more heterogeneous samples. Otherwise, it is the Journal for the Study of Listening as Defined by College Students.

Perhaps more important is the lack of precision with regard to the populations of principal interest. The people, stimuli, and events that a study seeks to illustrate or draw inferences about affects nearly every conclusion of a given investigation; thus the lack of information regarding the individuals included in a given inquiry is unfavorable to making claims about what we know about listening. One strategy from which we could determine the extent of homogeneity in samples and aid in future endeavors to replicate research is to report ranges, central tendencies, and variability of samples. Reporting means and standard deviations, for instance, helps readers determine exactly what type of sample is being discussed and generalized, and what types of samples need to be tested in future research.

Recommendation 6: Be Clear as to the Implications of Study Results

Related to the issue of samples and sampling discussed in Recommendation 5 is the issue of our ability to make causal claims about listening—the antecedents and consequences of listening in particular ways, whether individual differences in listening reliably produce differences in processing and/or vice versa, whether active listening behaviors influence processing or whether the reverse casual direction is more plausible, and the list goes on. Based on our analysis, there is an overreliance on cross-sectional research and a striking lack of experimental and longitudinal studies. Because the only claims scholars can make about listening when data are gathered simultaneously is that certain variables are related, it seems that our knowledge of listening is

largely of a bidirectional nature rather than of a causal one. We suspect that the overreliance of cross-sectional research is one reason that listening research is heavily atheoretical (Bodie, 2009, 2010, 2011; Wolvin, Halone, & Coakley, 1999), and we suspect that concerted efforts to include experimental and longitudinal research in the pages of future *IJL* issues will change this predicament. At minimum, scholars who pursue cross-sectional research should include a discussion of the limitations of such research and speculate about the theoretical structure among the variables of interest.

Recommendation 7: Limit the Reliance on Self-Report Measures of Listening

In a similar vein, there is an overreliance on self-report measures of listening. Although self-reporting listening is certainly not universally inappropriate—for instance, the Listening Concepts Inventory (LCI; Imhof & Janusik, 2006) assesses individual conceptualizations of listening akin to the work by O'Keefe (1988) on implicit theories of communication (i.e., message design logics)—most scales are aimed at assessing the general enactment of specific behaviors. The Self-Perceived Listening Competence scale (SPLC; Ford et al., 2000; Mickelson & Welch, 2012) includes items such as "I can interpret correctly persons' facial expressions." While attempts to assess the validity of self-reporting listening behaviors are available, they are rare (Bodie, Jones, & Vickery, 2012). Indeed, most studies utilizing self-reports of listening behaviors do not attempt to empirically dismiss other plausible explanations for found associations among measures of listening and important antecedents and consequences, such as common method variance (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Other research assumes that different perspectives (e.g., direct supervisors versus peers) are driving variability in scores without submitting such speculations to full tests (Cooper & Husband, 1993). As has been pointed out by others, listening is a socially desirable behavior, perhaps even more suspect to social desirability effects than other communication actions (Lawson & Winkelman, 2003). Moreover, there are readily available statistical (e.g., structural equation modeling) and methodological techniques (e.g., round robin designs, multitrait-multimethod studies) that listening scholars can utilize to address these issues.

Irrespective of statistical and other operationally-relevant concerns, what scholars and practitioners of listening are most interested in is what listeners do when interacting with others and whether the enactment of specific behaviors impacts important outcomes. If so, relying too heavily on self-report measurement for the advancement of knowledge about listening seems counterproductive. However, of all the measurement choices available to those interested in listening, the least employed is the assessment of actual behaviors. Those behaviors include not only linguistic responses indicating understanding or seeking clarity (e.g., asking questions) but also those nonlinguistic acts such as smiling and eye contact generally discussed in the literature as nonverbal immediacy (Bodie, St. Cyr, Pence, Rold, & Honeycutt, 2012). Perhaps one reason for the relative lack of behavioral listening research is its costs. It is far less time and labor intensive to collect a battery of self-report scales than it is to videotape conversations or group discussions (the LSU Listening Lab, or instance, has spent more than two years collecting a single data corpus, and we are not even finished). Indeed, behavioral listening research raises extensive logistical issues (Bodie, 2013). Likewise, while self-report data are easily analyzed using readily available statistical packages, behavioral data have to be coded, transformed, or otherwise handled in line with specific theoretical and practical purposes. Decisions relevant to this latter issue are not

easy to make, especially when research interests go beyond readily available coding rubrics or established rating scales.

Even so, behavioral data are rich and can offer insights not afforded by self-reports. As an example, Bavelas and colleagues have spent several decades exploring the listener as addressee, or "the person the speaker is addressing directly and who can respond to and interact with the speaker in a dialogue" (Bavelas & Gerwing, 2011, p. 180). Perhaps most important is that the addressee is a "full partner in creating the dialogue" (p. 180). Attending to how the listener contributes to dialogue shifts the notion of listener as a passive recipient and retainer of information to active constructer of meaning. Such a shift is largely impossible if we blindly stick to self-report measurement out of shear convenience.

Recommendation 8: Avoid the Use of Intact Classes or Groups

In our results, we reported that 28 studies (25.5% of all studies) utilized intact classes during data collection efforts. This statistic is astounding; the practice has potentially serious ramifications concerning the internal validity of a study (for more information, see Babbie, 1992; Barnett, 1991; Kerlinger, 1986; Moser & Kalton, 1972; Slonim, 1960; Smith, 1988; Thompson, 1992). Selecting participants from intact groups can result in selection bias, and this malady can affect the ability of a study to detect a true relationship between the independent and dependent variables. Because the participants are from an intact group, the effect of X on Y may in fact be due to another variable that every participant is exposed to equally. In experimental work using intact classes, the independent variable can no longer truly be said to be *independent* because the researcher is not determining the level of the variable that each subject will experience.

If participants were assigned to the intact group using preexisting information about them, then the sample is not random. If comparing two intact groups, the groups may consistently differ because of the non-random assignment and not because of the relationship between X and Y. Participants assigned to groups because of ability are also guilty of this problem; however, intact groups may be considered to be experimental if the participants were randomly assigned to the intact groups prior to the experiment. For instance, an introductory general education communication course containing 100 students from all over the university is closer to representing all college students than an upper level research methods course of 25 communication majors. Even in the former case, however, there are issues. For instance, if a researcher is interested in a particular training protocol on abilities to retain information, using existing classes and providing one class with the training while treating the other as the control group ultimately conflates training with characteristics of the teacher, time of the class, and other potential nuance variables. The better strategy is to randomly assign participants to group and to control or measure any extraneous variables not of primary interest.

CONCLUSION

The recent marking of the *International Journal of Listening's* Silver Jubilee supports the staying power of the journal and its affiliate, the International Listening Association (ILA). Seeking to "be the international leader of listening practices, teaching and research," it is important for ILA

to continually strive for excellence. Excellence in social scientific research is one way to meet this vision, and our manuscript should be a welcomed overview.

The results of this investigation suggest that the journal and its editorial staff strive to be inclusive with regard to methodological or philosophical preferences of listening scholars. We reported that fewer than half (45.8%) of the manuscripts between 1987 and 2011 utilized numerical data to make principled arguments. This trend should continue, and we are confident that even more ways to investigate the importance of listening to our daily lives will be represented in the future.

At the same time, however, our results also suggest in studies that do utilize numbers, more variation in demographics is necessary, especially ethnicity, culture, age, and the use of college students. More balance in the use of male and female participants is also needed. If the ILA is to truly be the international leader of listening education and research, its primarily publication outlet should be reflective of that goal.

In addition, psychometric properties should be reported more often when using self-report, other-report, evaluation or assessment, including descriptions of measurement models and reliability estimates. Reliability estimates are also a concern as the results imply that they should be higher across the board. Samples should be larger and more randomized, and when using descriptive statistics, measures of central tendency and variability should be reported with greater frequency. Samples should also be tested for normality, and these statistics should be reported, along with information about missing data. Effect sizes and confidence intervals are a concern, as they are rarely included in results. Lastly, more attention should be given to the appropriate application of statistical methods.

To aid in alleviating these observations, these authors made eight recommendations. These recommendations are far from novel, but they do represent fundamentally important decisions that scholars should make. We believe that following these recommendations will only benefit the next 25 years of *IJL* scholarship. In this light, it bears mentioning that the authors of this article are not exempt from these recommendations. On many occasions the second author has appeared in this journal. On some of these he has failed to report, among other crucial data, information on gender, race, age, and class. He has exclusively relied on data from U.S. participants, and in one case he neglected to report effect sizes. Thus, we can certainly stand to improve reporting methods and join the overall effort at making the *International Journal of Listening* more competitive in the intellectual marketplace.

Finally, these authors would also like to point out that we are not advocating that methods should drive research interests. Methodological rigor is beneficial in many ways, but ultimately theory and interpretation guide research. As such, we leave you with the following thoughts retold by Wilkinson (1999, p. 608), written more than 50 years ago by Hotelling, Bartky, Deming, Friedman, and Hoel (1948) but that still hold true today:

> "Unfortunately, too many people like to do their statistical work as they say their prayers—merely substitute in a formula found in a highly respected book written a long time ago" (p. 103). Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. If editors keep in mind Fisher's (1935) words . . . then there is less danger of methodology substituting for thought. Statistical methods should guide and discipline our thinking but should not determine it.

ACKNOWLEDGEMENT

The authors would like to thank Michelle Pence for assistance with data coding.

REFERENCES

American Psychological Association. (2010). *The publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Babbie, E. R. (1992). *The practice of social research* (6th ed.). Belmont, CA: Wadsworth.

Barnett, V. (1991). *Sample survey principles and methods*. New York, NY: Oxford University Press.

Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*, *25*, 178–198. doi: 10.1080/10904018.2010.508675

Bodie, G. D. (2009). Evaluating listening theory: Development and illustration of five criteria. *International Journal of Listening*, *23*, 81–103. doi:10.1080/10904010903014434

Bodie, G. D. (2010). Treating listening ethically. *International Journal of Listening*, *24*, 185–188. doi: 10.1080/10904018.2010.513666

Bodie, G. D. (2011). The understudied nature of listening in interpersonal communication: Introduction to a special issue. *International Journal of Listening*, *25*, 1–9. doi: 10.1080/10904018.2011.536462

Bodie, G. D. (2013). Issues in the measurement of listening. *Communication Research Reports*, *30*, 76–84. doi: 10.1080/08824096.2012.733981

Bodie, G. D., & Fitch-Hauser, M. (2010). Quantitative research in listening: Explication and overview. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 46–93). Malden, MA: Blackwell.

Bodie, G. D., Jones, S. M., & Vickery, A. J. (2012, May). *Measuring supportive listening: A multitrait-multimethod validity assessment*. Paper presented at the annual conventional of the International Communication Association, Phoenix, AZ.

Bodie, G. D., St. Cyr, K., Pence, M., Rold, M., & Honeycutt, J. M. (2012). Listening competence in initial interactions I: Distinguishing between what listening is and what listeners do. *International Journal of Listening*, *26*, 1–28. doi: 10.1080/10904018.2012.639645

Bodie, G. D., & Worthington, D. L. (2010). Revisiting the Listening Styles Profile (LSP-16): A confirmatory factor analytic approach to scale validation and reliability estimation. *International Journal of Listening*, *24*, 69–88. doi: 10.1080/10904011003744516

Bodie, G. D., Worthington, D. L., & Fitch-Hauser, M. (2011). A comparison of four measurement models for the Watson-Barker Listening Test (WBLT)-Form C. *Communication Research Reports*, *28*, 32–42. doi:10.1080/08824096.2011.540547

Bodie, G. D., Worthington, D. L., & Gearhart, C. C. (2013). The Revised Listening Styles Profile (LSP-R): Development and validation. *Communication Quarterly*, *61*, 72–90. doi: 10.1080/01463373.2012.720343

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cooper, L., & Husband, R. L. (1993). Developing a model of organizational listening competency. *Journal of the International Listening Association*, *7*, 6–34.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). New York, NY: Hafner Press. (Original work published 1935)

Fitch-Hauser, M., & Hughes, A. (1987). A factor analytic study of four listening tests. *The Journal of the International Listening Association*, *1*, 129–147.

Ford, W. S., Wolvin, A. D., & Chung, S. (2000). Students' self-perceived listening competencies in the basic speech communication course. *International Journal of Listening*, *14*, 1–13.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update.* Boston, MA: Allyn & Bacon.

Hotelling, H., Bartky, W., Deming, W. E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *Annals of Mathematical Statistics*, *19*, 95–115.

Hubbard, R. (1978). The probable consequences of violating the normality assumption in parametric statistical analysis. *Area*, *10*, 393–398.

Imhof, M., & Janusik, L. A. (2006). Development and validation of the Imhof-Janusik Listening Concepts Inventory to measure listening conceptualization differences between cultures. *Journal of Intercultural Communication Research*, *35*, 79–98. doi: 10.1080/17475750600909246

The International Statistical Institute. (2003). *The Oxford dictionary of statistical terms* (Y. Dodge, Ed.). Oxford, England: Oxford University Press.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart, & Winston.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London, England: Routledge.

Lawson, M., & Winkelman, C. (2003). The social desirability factor in the measurement of listening skills: A brief report. *Counseling Psychology Quarterly*, *16*, 43–45. doi: 10.1080/0951507021000050212

Levine, T. R. (2011). Quantitative social science methods of inquiry. In M. Knapp & J. Daley (Eds.), *Handbook of interpersonal communication* (pp. 25–58). Thousand Oaks, CA: Sage.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communicating research. *Human Communication Research*, *28*, 612–625.

Levine, T. R., Hullett, C. R., Turner, M. M., & Lapinski, M. K. (2006). The deirability of using confirmatory factor analysis on published scales. *Communication Research Reports*, *23*, 309–314. doi: 10.1080/08824090600962698

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188–209.

Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: Practical and theoretiacl issues. *Multivariate Behavioral Research*, *32*, 53–76. doi: 10.1207/s15327906mbr3201_3

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66* (Monograph Supplement 1-Vol. 66), 195–244.

Meline, T., & Wang, B. (2004). Effect-size reporting practices in AJSLP and other ASHA journals, 1999–2003. *American Journal of Speech-Language Pathology*, *13*, 202–207.

Mickelson, W. T., & Welch, S. A. (2012). Factor analytic validation of the Ford, Wolvin, and Chung Listening Competence Sccale. *International Journal of Listening*, *26*, 29–39. doi:10.1080/10904018.2012.639646

Moser, C. A., & Kalton, G. (1972). *Survey methods in social investigation* (2nd ed.). New York, NY: Basic Books.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

O'Keefe, B. J. (1988). The logic of message design: Individual differences in reasoning about communication. *Communication Monographs*, *55*, 80–103. doi: 10.1080/03637758809376159

Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research*, *27*, 562–577. doi: 10.1111/j.1468-2958.2002.tb00824.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903. doi: 10.1037/0021-9010.88.5.879

Shapiro, M. A. (2002). Generalizability in communication research. *Human Communication Research*, *28*, 491–500. doi: 10.1111/j.1468-2958.2002.tb00819.x

Slonim, M. J. (1960). *Sampling*. New York, NY: Simon & Schuster.

Smith, M. J. (1988). *Contemporary communication research methods*. Belmont, CA: Wadsworth.

Sun, S., & Fan, X. (2010). Effect size reporting practices in communication research. *Communication Methods and Measures*, *4*, 331–340. doi: 10.1080/19312458.2010.527875

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

Thompson, S. (1992). *Sampling*. New York, NY: Wiley.

Villaume, W. A., & Weaver, J. B., III. (1996). A factorial approach to establishing reliable listening measures from the WBLT and the KCLT: Full information factor analysis of dichotomous data. *International Journal of Listening*, *10*, 1–20.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi: 10.1037/0003-066X.54.8.594

Wolvin, A. D., Halone, K. K., & Coakley, C. G. (1999). An assessment of the "intellectual discussion" on listening theory and research. *International Journal of Listening*, *13*, 111–129.

Ziliak, S., & McCloskey, D. (2009). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.