Perceptual and Motor Skills, 2011, 112, 2, 331-348.

SAMPLE SIZE IN PSYCHOLOGICAL RESEARCH OVER THE PAST 30 YEARS[1, 2]

JACOB M. MARSZALEK, CAROLYN BARBER, JULIE KOHLHART

University of Missouri–Kansas City

COOPER B. HOLMES

Emporia State University

Summary.—The American Psychological Association (APA) Task Force on Statistical Inference was formed in 1996 in response to a growing body of research demonstrating methodological issues that threatened the credibility of psychological research, and made recommendations to address them. One issue was the small, even dramatically inadequate, size of samples used in studies published by leading journals. The present study assessed the progress made since the Task Force's final report in 1999. Sample sizes reported in four leading APA journals in 1955, 1977, 1995, and 2006 were compared using nonparametric statistics, while data from the last two waves were fit to a hierarchical generalized linear growth model for more in-depth analysis. Overall, results indicate that the recommendations for increasing sample sizes have not been integrated in core psychological research, although results slightly vary by field. This and other implications are discussed in the context of current methodological critique and practice.

One of the characteristics that sets science apart from other epistemologies is the use of empirical, public observations, which give it the ability to self-correct (Elmes, Kantowitz, & Roediger, 1992). However, the extent to which modern science does correct itself has recently been called into question in both the academic (e.g., Ioannidis, 2005; Merenda, 2007) and popular presses (e.g., Hotz, 2007). The field of psychology has shown susceptibility to systemic breakdowns in the conduct of science, as demonstrated by the formation of the Task Force on Statistical Inference by the Board of Scientific Affairs of the American Psychological Association (APA) in 1996 (Wilkinson & the Task Force on Statistical Inference, 1999). The Task Force was formed in response to a growing body of research demonstrating methodological issues in the field that threatened the credibility of psychological research, and its recommendations have influenced journal article reporting standards published in the past three editions of the *Publication Manual* (APA Publication Board, 2008). One is-

[1]Address correspondence to Dr. Jacob M. Marszalek, Division of Counseling and Educational Psychology, University of Missouri–Kansas City, 215 Education, 5100 Rockhill Road, Kansas City, MO 64110 or e-mail (marszalekj@umkc.edu).

[2]The authors would like to thank Courtney Moody and Rachel Kirkpatrick for their assistance in the 2006 data collection.

DOI 10.2466/03.11.PMS.112.2.331-348 ISSN 0031-5125

sue addressed by the Task Force was the size of samples used in studies published by leading journals (e.g., Holmes, 1979, 1983; Holmes, Holmes, & Fanning, 1981), which affects generalizability, estimation of effect sizes, and accurate planning of future studies.

The issue of sample size in psychological research is closely tied to power and effect size, both of which have been surveyed periodically over the past 50 years, and the current study seeks to add to that body of knowledge. Cohen (1962) surveyed power in articles published in the 1960 volume of the *Journal of Abnormal and Social Psychology,* and depended on his own estimates of effect sizes obtained from the reported statistics. Cohen found that, on average, power was woefully inadequate for studies investigating what he termed "small" and "medium" effects; power was adequate on average for "large" effect sizes. Similar to the present study, Sedlmeier and Gigerenzer (1989) assessed progress made in the field since Cohen's study by surveying power in the 1984 volume of the *Journal of Abnormal Psychology* (the *Journal of Abnormal and Social Psychology* split into the *Journal of Abnormal Psychology* and the *Journal of Social Psychology* in 1965). After correcting for the use of adjusted alpha levels, which were not in widespread use in 1960, Sedlmeier and Gigerenzer found that despite Cohen's recommendations, power had actually declined over 24 years! With a similar purpose, Rossi (1990) conducted a wider survey of articles in the 1982 volumes of the *Journal of Abnormal Psychology, Journal of Social Psychology,* and *Journal of Consulting and Clinical Psychology,* and found that power had not improved in the 20 years since Cohen's study.

Other surveys of observed statistical power have found similar trends (e.g., Maddock & Rossi, 2001; Maxwell, 2004). Maddock and Rossi (2001) surveyed the 1997 volumes of three health psychology journals, and found power for studies reporting small effect sizes to be .36—much less than Cohen's suggested minimum (1988) of .80—whereas studies reporting medium and large effect sizes averaged power levels of .77 and .92, respectively. However, Maddock and Rossi (2001) seemed to be an exception unique to health psychology, because Maxwell (2004) reported four other studies published during the same period (1993–2001) that, like Cohen (1962), found low power for studies reporting medium effect sizes.

One possible reason that power has failed to improve is that sample sizes have remained small. Cohen (1962) concluded, "Increased sample size is likely to prove the most effective general prescription for improving power" (p. 153), but there is little evidence that the field has taken note. After reviewing the literature, Holmes (1979) reported finding only two studies that examined sample sizes directly. One study reported the number of articles published about single-subject samples (Dukes, 1965), and the other examined sample sizes reported in two British journals, finding that every reported study had $N \leq 25$ (Cochrane & Duffy, 1974).

Holmes (1979, 1983) himself examined sample sizes in four APA journals in 1955 and 1977, and reported median sample sizes for the total study and each of the comparison groups. His general conclusions were that sample size had not changed significantly between 1955 and 1977, and that the typical sample size in psychology did not seem large (Holmes, 1979). Holmes also criticized the difficulty he had in determining the sample sizes in many articles, finding the reporting to be unnecessarily obtuse or convoluted. Our review of the literature since 1979 revealed just one additional direct examination of sample sizes used in psychological research, Holmes, *et al*. (1981). One possible weakness of the Holmes (1979, 1983) study was that its targeted sample of four highly selective journals may not have adequately represented the field. To account for this, Holmes, *et al*. (1981) replicated his methodology in examining sample sizes reported in 10 non-APA journals from 1977, and found similar results. To the best of our knowledge, no examination of sample sizes had been conducted since Wilkinson and the Task Force on Statistical Inference (1999) published its recommendations 10 years ago, and in light of the age of previous results, it seemed time to conduct a new examination.

The sample size findings of Holmes (1979, 1983) and Holmes, *et al*. (1981) parallel the surveys of observed power reported in the literature (Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Rossi, 1990; Maddock & Rossi, 2001; Maxwell, 2004), and even overlap in specific journals (i.e., *Journal of Abnormal Psychology*). Thus, sample size inadequacy may indeed explain much of the observed deficiencies in power. However, Holmes' research stops in the 1977 literature, whereas more recent studies on power extend to the 2001 literature, although that is also aging. Therefore, the purpose of the present study was to examine sample sizes reported in the same four journals examined by Holmes (1979, 1983), but in more recent volumes. Two additional data collections were undertaken, one in 1995 (about the time the Task Force was formed), and the other in 2006, so that both descriptive and trend data would be available for several periods spanning Cohen's study (1962), to the more recent surveys of observed power that were undertaken, to the present. It was hypothesized that: (a) the sample sizes reported for 1995 and 2006 would be no different than those reported in 1955 and 1977; (b) the sample sizes reported in the four psychology subfields represented by the chosen flagship journals would be no different; and (c) the sample sizes had no increase after the Task Force report in 1999. We tested each of these hypotheses for both total sample size in a given article study, and for the size of study groups.

Method

Published research articles in four APA journals selected by Holmes (1979, 1983) were used. Holmes examined the years 1955 and 1977, be-

cause 1977 was the most recent available year at the time and 1955 allowed a sufficient amount of time to assess any appreciable change in average sample size. APA journals were selected, "because their rejection rate … suggests that only the best articles were published" (Holmes, 1979, p. 284). In order to obtain a broad representation of psychology, Holmes selected the following journals: *Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance,* and *Developmental Psychology. Developmental Psychology* was not included in the 1955 data collection, because it did not exist prior to 1969. Another difference in the 1955 data collection was that the *Journal of Experimental Psychology* was used, because it had not yet separated into different journals. Additional data from 1995 and 2006 were collected on the same four journals.

Procedures

Holmes (1979, 1983) used the following procedure for determining sample sizes reported in 1955 and 1977 journal articles, which was also used in the 1995 and 2006 data collections. Each article was read to determine two different sample sizes, a total sample size and an individual sample size. The total sample size represented the total number of subjects used in the study regardless of design groupings. The individual sample sizes represented the sizes of each of the design groupings in the study. For example, a study comparing treatment and control groups with 100 participants randomly assigned would have a total sample size of 100 and two individual sample sizes of 50. The reason for examining individual sample sizes was that "whatever conclusions are reached about a treatment factor, they are based on the number of subjects under that condition, not on the total sample size" (Holmes, 1979, p. 284). When an article reported two or more studies, each was recorded as a separate article with its own total and individual sample sizes.

For the 2006 data collection, our research team found it necessary to formulate some additional rules for deciding when an article should be recorded as having more than one study. One reason for the need for additional rules was to facilitate communication among investigators, ensuring a standardized data collection. Another reason was an increase in the complexity and comprehensiveness of research results; multiple manipulations of the same set of data in an article seemed more commonplace in 2006 than in the previous years, perhaps as part of the shift from null hypothesis significance tests to statistical modeling identified by Rodgers (2010). Through discussion and trial and error, we eventually settled on the following rules: (a) the total sample was whatever general group was used for analysis; (b) each study was defined as an attempt to answer a research question about a unique sample or subsample with its own analysis; (c)

in longitudinal studies, the total sample was defined as whatever subjects completed all phases of the study; (d) regarding demographic variables (e.g., race/ethnicity), the groups were counted as individual sample sizes if the variable was an *a priori* part of the original study design, but not if it was an afterthought (i.e., part of a *post hoc* analysis); (e) statistical simulations were excluded; and (f) studies involving only forms of artificial intelligence (AI) were also excluded, but studies involving human-AI interaction were included. We found (d) to be necessary because of the sheer number of studies that would have been reported if *post hoc* analyses were included (*post hoc* in the sense of comparisons made merely in support of the main findings, not in the sense of statistical comparisons that were obviously planned but carried out with *post hoc* methods, such as Tukey tests). We found (e) and (f) to be necessary because of the inherently biasing effect that studies without humans have on sample sizes. For example, statistical simulations are limited only by the researcher's imagination and time. No animal studies were found in any of the four journals in 2006. It also bears mention that no articles in the sixth and final issue (a special issue) of the *Journal of Applied Psychology* reported studies using human subjects, but instead reported literature reviews and discussions of theory.

Interrater Agreement

After the 2006 data were collected, each article was assigned a unique number, and 10% ($n = 76$) were randomly selected using the Select Cases procedure in SPSS Version 16.0. One of the students re-recorded the total and individual sample sizes from each of the selected articles, and the results were compared to the initial data collection. Agreement was measured with Krippendorff's alpha, α_K, using the KALPHA SPSS (now PASW) macro (Hayes & Krippendorff, 2007), treating sample size as a ratio level measurement and performing 1,000 bootstrapping samples. Agreement was ascertained separately for total sample sizes and individual sample sizes, because many studies did not use comparison groups (43 to 46 out of 68 cases, or 63 to 68%). In three of 25 cases (12%), there was disagreement on whether there were comparison groups at all.

For the total sample sizes, 69 cases were evaluated for agreement (six had no sample size recorded because of the type of article or study), and $\alpha_K = .98$ (95%$CI = .95, 1.00$). We counted this as good evidence of the reliability of our measurement of total sample sizes. There was disagreement on six studies, which were reviewed by the authors. For five of the studies, the authors were able to determine that the original rater was correct, but the sixth study was sufficiently unclear in its sample description that a definitive judgment was impossible, and it was dropped from further analysis.

For the individual sample sizes, two measures of rater agreement were

taken. The first measure was of whether the raters agreed that comparison groups did exist in the study, and the second measure was of whether the raters agreed on the actual sample sizes themselves. Sixty-eight cases were evaluated for the first measure, and $\alpha_K = .90$ (95%*CI* = .77, 1.00); the raters disagreed on three of the 74 studies. We took this as evidence of acceptable reliability. The raters agreed that 22 studies actually had comparison groups, but disagreed on whether the individual sample sizes could actually be determined on two of the studies. In one study, no one in the research team could determine individual sample sizes consistent with the total sample size, so it was dropped from the present analysis of individual sample sizes. In the other case, the authors were able to determine that the original rater was correct by reviewing the original published article. The final 21 studies contained 51 individual sample sizes (in other words, comparison groups), and α_K was greater than .99 (95%*CI* = .99, 1.00); the raters disagreed on just one sample size and differed by a single participant. In summary, there was good evidence that total and individual sample sizes were reliably interpreted.

Analysis

Nonparametric analysis including older data waves.—Data for 1955 and 1977 were taken from Holmes (1979, 1983). More specifically, descriptive summary statistics from across all articles sampled were available for 1955, while summary data broken down by journal were available for 1977. In neither case were data available for individual articles. Therefore, our statistical analyses to address Hypothesis 1 relied on the comparison of summary statistics across journals and years, which required the use of nonparametric tests. More specifically, a series of three one-sample Wilcoxon signed-rank tests was employed to compare the median sample size from 1955 to the median of the within-journal sample sizes in 1977, 1996, and 2006, respectively. To compare the median sample sizes across the last three waves of data collection (when statistics were available by journal), we used an independent-samples median test.

To address Hypothesis 2 (regarding differences of sample sizes across journals), we conducted two different nonparametric tests. First, we considered an independent-samples test in order to test for overall differences between median sample sizes across journals using an independent-samples median test. Second, we considered the yearly median sample sizes within journals as repeated measures of journal sample sizes. Thus, we employed Friedman's related-samples test to determine whether the ranked order of journal sample sizes changed across the three data waves. All nonparametric tests were conducted using PASW Version 18.0.0.

Multilevel modeling of two data waves.—The use of nonparametric statistics was necessary given the lack of data available on individual articles

from 1955 and 1977. However, such article-level data were available from 1996 and 2006. In a separate analysis looking only at these last two data waves, we employed multilevel modeling techniques in order to incorporate all available data on individual articles, which were nested within years (or journal volumes) and further nested within journals. Such analysis gives additional insight into the testing of Hypotheses 1 and 2 by providing an in-depth examination of the differences in sample size between 1996 and 2006 across the four sampled journals. Moreover, these analyses enabled us to test Hypothesis 3 in a sophisticated way by focusing on differences in the data waves collected most immediately before and after the report by Wilkinson and the Task Force on Statistical Inference (1999).

Holmes (1979, 1983) found that sample distributions were positively skewed (as is typically found with count data), and employed the nonparametric median for a valid comparison of such nonnormal data (Corder & Foreman, 2009). We similarly anticipated that data from 1996 and 2006 would be positively skewed, rendering hierarchical linear modeling techniques inadequate. Therefore, we chose to analyze these data using hierarchical *generalized* linear modeling techniques (HGLM; Raudenbush, Bryk, Cheong, & Congdon, 2004). More specifically, a Poisson model with constant exposure was selected, as Poisson models are especially appropriate for count data (Raudenbush, *et al*., 2004). To analyze a model employing an outcome with a Poisson distribution, HLM 6 statistical software employs a log link function as follows:

$$E(N \mid \pi) = \lambda \quad [1]$$

$$Log[\lambda] = \eta \ . \quad [2]$$

In these equations, λ is the sample size of the article, while η is the log transformation of the article sample size. It is this transformation that was modeled, using the following equations:

$$\text{Level 1: } \eta = \pi_0 \quad [3]$$

$$\text{Level 2: } \pi_0 = \beta_{00} + \beta_{01}(YEAR) + r_0 \quad [4]$$

$$\text{Level 3: } \beta_{00} = \gamma_{000} + u_{00} \quad [5]$$

$$\beta_{01} = \gamma_{001} + u_{01} \ . \quad [6]$$

In these equations, π_0 is the average log sample size in a volume year, β_{00} is the average log sample size in a given journal across volume years, and γ_{000} is the overall average log sample size. β_{01} represents the effect of year on average log sample size in a volume year. Random effects are represented by r_0, u_{00}, and u_{01} for volume year, journal, and the effect of year within journal, respectively. Unit-specific models were computed using a full Penalized Quasi-likelihood estimation method using HLM 6 software

(Raudenbush, *et al*., 2004).

It should also be noted that the Poisson distribution, while generally thought to be the most appropriate for count data, has several assumptions that must be met. Most importantly, the Poisson distribution assumes that the standard deviation is equal to the mean, allowing for the dispersion parameter to be fixed to 1. Therefore, prior to the estimation of the hierarchical generalized linear model, we examined the distribution of article sample sizes to determine whether this assumption was met. The results of this process are summarized in our section on descriptive and exploratory analyses.

Results

Descriptive and Exploratory Analyses

Overall descriptive statistics are summarized in Table 1, and those for each journal from each year are summarized in Table 2 for total sample sizes, and Table 3 for individual sample sizes. The median sample sizes for 1955 reported in Table 1 served as our comparison figure for the Wilcoxon signed-rank tests conducted to test Hypothesis 1, while the median sample sizes reported in Tables 2 and 3 provided the data for additional tests of Hypothesis 1 as well as tests of Hypothesis 2.

We also calculated additional descriptive statistics to verify our assertion that the article-level data were positively skewed, and that hierarchical generalized linear modeling was warranted. This was confirmed for both the total sample size (skewness = 12.76, $SE = 0.07$) and the individual sample size (skewness = 12.75, $SE = 0.11$). Moreover, the standard deviations for the yearly sample sizes reported in Table 1 were larger than the reported means. These standard deviations represented variances in the tens of thousands, indicating the presence of extreme outliers when considered alongside interquartile ranges measureable by tens or hundreds. This supports the decision to relax the Poisson model to allow for overdispersion of the distribution (i.e., greater variance than expected), essentially requiring a negative binomial distribution where a dispersion parameter is estimated (Raudenbush, *et al*., 2004).

A final exploratory analysis in support of the negative binomial distribution was conducted using the Hierarchical Generalized Linear Mod-

TABLE 1

Aggregated Summary Statistics for Total Sample Sizes in All Four Journals by Year

Year	*N*	Min.	Q_1	Median	Q_3	Max.	Modes	*M*	*SD*
1955	448	1	25.50	59.95	131.30	16,584	60, 40, 30, 24	180.49	193.86
1977	507	1	18.13	48.40	94.00	45,144	8, 12, 6, 60	217.96	2,026.53
1995	527	1	14.00	32.00	87.50	16,930	8	211.03	983.38
2006	690	1	18.00	40.00	136.00	13,059	16	195.78	680.02

TABLE 2

TOTAL SAMPLE SIZES IN STUDIES REPORTED IN FOUR APA JOURNALS IN 1977, 1996, AND 2006

Journal	*N*	Min.	Max.	Range	Q_1	Median	Q_3	Modes	*M*	*SD*
Abnormal										
1977[a]	89	[b]	[b]	2,272	29.92	51.50	79.75	48	126.65	309.80
1995	74	1	7,691	7,690	42.00	70.00	183.00	32, 36, 40, 48	354.00	993.00
2006	107	22	5,847	5,825	45.00	107.00	269.00	34, 35	340.00	718.00
Applied										
1977	117	[b]	[b]	45,140	60.50	114.50	271.25	72	682.53	4,168.29
1995	69	25	16,930	16,905	86.00	146.00	385.00	25, 60, 84, 85, 86	740.00	2,219.00
2006	112	15	13,059	13,044	96.00	148.00	300.00	139	404.00	1,373.00
Developmental										
1977	140	[b]	[b]	1,472	33.50	60.16	87.50	32, 64, 80	90.88	147.73
1995	110	3	6,526	6,523	30.00	60.00	114.00	24	257.00	828.00
2006	145	16	4,511	4,495	50.00	114.00	313.00	50	295.00	543.00
Experimental										
1977	161	[b]	[b]	2,056	7.51	12.20	31.68	6, 8	41.33	166.40
1995	274	1	288	287	8.00	15.00	30.00	8	23.00	34.00
2006	334	1	533	532	10.00	18.00	32.00	10	26.00	35.00

[a]The 1977 data were taken from Holmes (1983). [b]Data unavailable.

TABLE 3

INDIVIDUAL (GROUP) SAMPLE SIZES IN STUDIES REPORTED IN FOUR APA JOURNALS IN 1977, 1996, AND 2006

Journal	*N*	Min.	Max.	Range	Q_1	Median	Q_3	Modes	*M*	*SD*
Abnormal										
1977[a]	324	[b]	[b]	1,299	9.62	12.37	26.83	10	34.79	91.52
1995	1,776	9	893	884	15.00	20.00	42.00	16.00	82.00	165.00
2006	2,568	2	1,462	1460	19.00	26.00	56.00	16.00	103.00	244.00
Applied										
1977	384	[b]	[b]	4,267	14.98	32.50	114.50	20, 12	207.96	573.52
1995	1,656	10	14,905	14,895	14.00	22.00	78.00	12.00	191.00	1,190.00
2006	2,688	5	1,468	1,463	21.00	21.00	83.00	21	89.00	218.00
Developmental										
1977	591	[b]	[b]	297	9.05	11.95	19.55	12	21.53	46.04
1995	2,640	7	2,526	2,519	15.00	24.00	44.00	10.00	83.00	258.00
2006	3,480	2	1,876	1,874	12.00	25.00	54.00	12.00	73.00	163.00
Experimental										
1977	241	[b]	[b]	2,057	6.30	10.27	17.08	8, 6, 12	27.61	132.04
1995	6,576	1	72	71	9.00	10.00	15.00	10.00	15.00	14.00
2006	8,016	1	82	81	10.00	12.00	19.00	12.00	17.00	13.00

[a]The 1977 data were taken from Holmes (1983). [b]Data unavailable.

els module in PASW Statistics 18. Unconditional negative binomial models with log link functions were calculated using both the total sample size and individual sample size as outcomes, and in each case a dispersion parameter was estimated. In each case, the estimated dispersion parameter was larger than 1 (total sample parameter = 2.30, $SE = 0.08$, 95%$CI = 2.15$, 2.46; individual sample parameter = 1.49, $SE = 0.11$, 95%$CI = 1.30$, 1.71). This provides additional support to the assumption that data are overdispersed, necessitating a negative binomial distribution and the estimation of a dispersion parameter greater than 1.

Analysis Across Three Data Waves

None of the analyses conducted to test differences in total sample sizes as outlined in Hypothesis 1 yielded statistically significant results. The Wilcoxon signed-rank tests revealed no statistically significant difference between the 1955 median sample size of 59.95 and the median sample sizes observed in 1977, 1996, or 2006. Similarly, the independent-samples median test conducted revealed no statistically significant differences among the median sample sizes examined in the last three waves of data collection.

In regards to Hypothesis 2 as it pertains to total sample sizes, the independent-samples median test revealed no statistically significant differences in the overall sample sizes of the four journals. However, Friedman's related-samples test revealed statistically significant differences in the ranked order of median within-journal sample sizes across the three years ($p = .039$). As reported in Table 2, this difference appears to be due to an increase in median sample size in the *Journal of Abnormal Psychology* from 1977 to 1996, resulting in a median sample size that is slightly larger than that reported for *Developmental Psychology.* By 2006, however, *Developmental Psychology* had increased sample sizes at a faster rate, resulting in a slightly larger median sample size in this journal as compared to the *Journal of Abnormal Psychology.* To contrast, sample sizes in the *Journal of Applied Psychology* and the *Journal of Experimental Psychology* stayed relatively constant. It should be noted, however, that the observed changes in rank order are small. In fact, follow-up pair-wise Wilcoxon signed-ranks tests to identify specific differences yielded only marginally significant results ($p = .068$ when comparing 1977 to 2006 and when comparing 1996 to 2006).

The same tests were conducted to test Hypotheses 1 and 2 as they pertained to individual sample sizes. No statistically significant results were found, indicating that individual group sample sizes stayed constant across years and journals.

Table 4 summarizes the hierarchical generalized linear model created to examine variations in sample sizes by journal in year. In regards to Hypothesis 1, there is no statistically significant effect of year on either to-

tal sample size or individual sample size. This result mirrors the results from the nonparametric analysis. In regards to Hypothesis 2, there is a statistically significant random effect of journal, indicating that the average sample size varies across the four journals examined. This significant random effect appears both when examining total sample size [$\tau_{(\beta)} = 1.53$, χ^2 (3) = 74.44] and individual sample size [$\tau_{(\beta)} = 0.44$, χ^2 (3) = 14.20]. This differs from the nonparametric analysis, which found no overall significant effect of journal, but may be reflective of the increased power of the hierarchical generalized linear model as compared to the median test (Corder & Foreman, 2009), as well as the availability of article-level data in this analysis. However, the random effect of year within journal is not statistically significant, indicating that there are no between-journal variations in changes in total or individual sample sizes between 1996 and 2006. Together, these findings provide support for Hypothesis 3, as neither total nor individual sample size appeared to increase significantly in any of the journals after the 1999 Task Force report.

TABLE 4
Hierarchical Generalized Linear Model for Total and Individual Sample Sizes Across Two Data Waves

	Total Sample Size				Individual Sample Size			
	β	*SE*	Rate	95%*CI*	β	*SE*	Rate	95%*CI*
Fixed Effect								
N, γ_{000}	5.34	0.63	208.88	51.71, 843.87	4.37	0.40	79.42	33.52, 188.18
Year, β_{01}	−0.01	0.02	0.99	0.95, 1.02	−0.01	.02	0.99	0.95, 1.02
	Variance	χ^2	*df*		Variance	χ^2	*df*	
Random Effect								
N between journals, μ_{00}	1.53	74.44*	3		0.44	14.20*	3	
N between volumes, r_0	4.29×10^{-4}				0.00			
Error, *e*	1,416.41				2,274.07			
Yr. between journals, μ_{01}	1.60×10^{-4}	4.11	3		2.10×10^{-4}	1.54	3	

*$p < .01$.

DISCUSSION

In examining the articles from these four psychological journals, we tested whether the sample sizes increased from 1995 to 2006. Holmes (1979) said in his initial study that "the size of the typical sample in psychological research does not appear to be large" (p. 288), and the same might be said of 1995 and 2006 if only looking at the overall statistics reported in Table 1, with medians ranging from 32 to 60. However, in addressing the second hypothesis—the trends in reported sample size in the four psychology subfields would be no different—we saw that such a con-

clusion would be superficial. Our modeling showed that sample size depends on the field. Smaller samples are needed in experimental settings, presumably because sufficient control of extraneous variation is in place, and standard errors tend to be smaller. (Higher cost per participant may also be a factor, due to sophisticated measurement equipment or laboratory controls.) However some fields, such as applied and developmental psychology, depend much more on quasi-experimental research because of their greater emphasis on comparisons of naturally occurring groups and ecological validity. Such research designs result in more variation in the data, and larger samples are necessary to gain feasible standard errors. (Lower cost per participant may also be a factor, because of the availability of institutional archival data.) Abnormal psychology may be an area in which Holmes' statement applies, but with the "deinstitutionalization" of the state hospital system between 1960 and 1995 (Dowdall, 1996), smaller than desired sample sizes prior to 2006 may be attributable to less accessibility to research populations caused by the broad shift to community-based mental health centers.

We also tested an extension of the second hypothesis, which stated that the sample sizes within journals would not differ over time. We found that overall, the relatively small sample sizes found by Holmes did not increase significantly over the next 29 years. However, there was significant variability in the change in sample size over time by field, with increases from 1977 to 2006 appearing in the *Journal of Abnormal Psychology* and *Developmental Psychology,* and no change in *Experimental Psychology* or *Applied Psychology* (which actually showed a slight decrease for individual sample size).

The third hypothesis was that sample sizes remained unchanged after the Task Force report in 1999. A change would have been reflected in a significant difference in sample size between 1995 and 2006, but none was found. This result is not surprising, given previous research on power (e.g., Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Rossi, 1990; Maddock & Rossi, 2001; Maxwell, 2004) and Holmes' own studies on sample size (Holmes, 1979, 1983; Holmes, *et al*., 1981). However, it is troubling, especially when one considers the increased use of sophisticated multivariate analyses and statistical modeling techniques during this time that would require the employment of larger sample sizes (Merenda, 2007; Rodgers, 2010).

The difficulty in discerning sample sizes remained the same in 1995 and 2006 as in 1955 and 1977, volume years about which Holmes (1979) complained of several problems. He reported being "surprised at the number of studies in which sample size was either not reported at all or was obtainable only by checking tables or making some mathematical cal-

culation based on the data that were available. This problem was especially acute in reference to individual sample size." (p. 287). Holmes also stated that many articles would report two different sample sizes, one in the abstract and the initial sample description, and the other in a later section after indicating the exclusion of some participants. Finally, Holmes reported difficulty in determining whether control groups had been used. We found similar difficulties in the 1995 data collection. Authors and editors seemed to be better about reporting sample sizes, but problems persisted. The comparison between the reporting of sample sizes in 1995 and in 1955 and 1977 was straightforward; most studies still entailed the reporting of group comparisons, such as ANOVA, ANCOVA, and multiple regression.

The 2006 data collection, however, indicated many studies with more sophisticated designs and analyses, such as MANOVA and SEM, which complicated the reporting of sample sizes and our ability to discern them. This change is in line with the shift from NHSTs to statistical modeling that occurred in the 1990s and 2000s (Rodgers, 2010). The reading and interpreting of articles from 2006 seemed much more arduous than that of articles from 1995, and partly explains our use of a research team. One of the doctoral students had this to say about her experience:

> In reviewing the articles for their sample size, it was often difficult to pinpoint how many subjects were used for the analysis. I often had to compare the number of subjects reported in the methods section with the *F* statistic degrees of freedom, because authors would sometimes fail to mention if cases were dropped. In looking at the sample sizes in the *Journal of Applied Psychology*, the authors were sometimes unclear about group assignments. In the *Journal of Experimental Psychology*, the authors were clearer about the sample sizes, which was important given the increased number of experiments per article.

More sophisticated treatments of data were more likely to be found in the *Journal of Applied Psychology* or *Developmental Psychology*, perhaps because those fields more often need them to statistically control extraneous variance. Even when methodological sophistication is the same, such as when statistical modeling is used, the models in those fields may be more complex because of the lack of direct investigator control over research conditions.

Limitations

There are several limitations associated with the analysis. We examined only one journal per field. Statistically significant random effects indicate some degree of variability among journals, and it appeared that this is due to different fields and methods of the disciplines. Including additional journals in each of these four disciplines would help. This would allow us to determine whether other applied or developmental journals, for example, similarly have larger sample sizes overall, or whether this is

a characteristic only of the major APA journals in these fields. Holmes, *et al*. (1981) provided some information about this already. In their examination of sample sizes of 10 non-APA journals of the volume year 1977, they reported finding results similar to those of Holmes' study (1979) of the four APA journals of 1955 and 1977. Among the 10 non-APA journals were *Child Development* and the *Journal of Applied Behavior Analysis*. Unfortunately, Holmes, *et al*. reported results only as an aggregate, and not by individual journal.

Statistical power could also be considered a limitation. Data were only available aggregated by journal in 1977, and were only available aggregated across all studies examined in 1955. Article-level data were only available for two time points (1995 and 2006). This meant that the complexity of the analyses that we could conduct using these older waves of data was limited, and that more intricate models of growth (e.g., modeling a quadratic function) could not be tested using HGLM. Ideally, we would like to consider more variation by year in order to better estimate the effect of time on article sample sizes. In making this critique, however, it should be noted that both sets of analyses conducted yielded statistically significant effects related to journal. Further, the nonsignificant fixed effect of year examined in the HGLM was very small in size, with total sample size decreasing at the rate of approximately two participants per year, which would not impugn the sensitivity of our tests.

Another limitation is the difference in procedures between 2006 and prior years, which was necessitated by the increased number of studies (Table 1 shows that there were 163 more studies in 2006 than 1995, a 31% increase) and the increased number of complex designs. We addressed this implementation threat through careful training and verification of procedural equality between Holmes' (1979, 1983) and our 1995 data collection, and again between the 1995 and 2006 data collections. However, stronger evidence of internal validity would be gained through successful replication of the 2006 methods with the same journals from a different year.

One explanation of the increase in the number of studies and study groups, and the complexity of designs, is more widespread use of multivariate methods, which should have been accompanied by an increase in sample sizes (Merenda, 2007). This relates to a final limitation of this study: the lack of differentiation between univariate and multivariate approaches. As related to our discussion of power, one could ask whether it is possible to identify *a priori* a target rate of growth that would result in an adequate average sample size in the last year of data collection. Power could then be assessed in terms of whether the study, as designed,

could detect this rate of growth as statistically significant. Taking this approach, however, relies on identifying an "adequate" target sample size, and that target depends heavily on the complexity of the models typically assessed. By examining univariate and multivariate approaches specifically, one could more specifically discuss the adequacy of sample-size practices in each area. This consideration will be especially important in examining sample sizes beyond 2006, as the use of multivariate methods continues to grow.

Recommendations

One easy recommendation based on the results of the present study would be that the APA, other research organizations, and journals go beyond merely providing guidelines for best practices regarding the reporting of sample sizes, and refuse to accept manuscripts for review if such guidelines are not followed. As shown by Alhija and Levy (2009), mere guidelines have not made much difference in increasing best practices in statistical reporting. Alhija and Levy (2009) reviewed the literature on reporting effect sizes, and noted that the fifth edition of the *Publication Manual of the American Psychological Association* (APA, 2001) made stronger recommendations for reporting effect sizes than previous editions. Filder (2002) concluded that practices of reporting effect sizes were only "modestly affected" (Alhija & Levy, 2009, p. 247). Alhija and Levy themselves compared the numbers of articles in which effect sizes were reported between five journals with explicit guidelines for effect-size reporting, and five journals without such guidelines. They found no significant difference. Therefore, we recommend that journal editors make clear reporting of sample sizes a criterion of acceptance of all manuscripts. Some journals already provide reviewers with rating scales for specific criteria, a practice we would like to see become more widespread and to include sample-size reporting, including the rationale for the sample size in each study described. Ideally, this would be based on power, detectable outcomes, and desired confidence interval width.

Maxwell (2004) also suggested that editors afford studies with nonsignificant results with as much chance of publication as studies with significant results, so that statistical significance is de-emphasized in favor of power and precision. Maxwell (2004) made two additional recommendations that we also endorse. One is more use of meta-analysis, which can help increase sample size, and thus power, and bypass typical obstacles, such as limited time and resources. The other is to increase the number of collaborative multisite studies, in which multiple scientists cooperate to conduct the same study at multiple sites and pool the data. We note with some unease that the continued lack of power identified in other studies, and use of smaller samples observed in the current study, correspond with

the decline in the number of psychometricians and quantitative methodologists observed by Merenda (2007). A final recommendation, then, is to redouble current efforts to educate psychology students about the importance of power and sample size.

REFERENCES

Alhija, F. N., & Levy, A. (2009) Effect size reporting practices in published articles. *Educational and Psychological Measurement,* 69, 245-265. DOI:10.1177/0013164408315266.

American Psychological Association. (2001) *Publication manual.* (5th ed.) Washington, DC: Author.

APA Publication Board Working Group on Journal Article Reporting Standards. (2008) Reporting standards for research in psychology: why do we need them? What might they be? *American Psychologist,* 63, 839-851. DOI: 10.1037/0003-066X.63.9.839.

Cochrane, R., & Duffy, J. (1974) Psychology and the scientific method. *Bulletin of the British Psychological Society,* 27, 117-121.

Cohen, J. (1962) The statistical power of abnormal–social psychological research: a review. *Journal of Abnormal and Social Psychology,* 65, 145-153.

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

Corder, G. W., & Foreman, D. I. (2009) *Nonparametric statistics for non-statisticians: a step-by-step approach.* New York: Wiley.

Dowdall, G. W. (1996) *The eclipse of the state mental hospital: policy, stigma, and organization.* New York: State Univer. of New York Press.

Dukes, W. (1965) *N* = 1. *Psychological Bulletin,* 64, 74-79.

Elmes, D. G., Kantowitz, B. H., & Roediger, H. L. (1992) *Research methods in psychology.* (4th ed.) St. Paul, MN: West.

Filder, F. (2002) The fifth edition of the APA publication manual: why its statistics recommendations are so controversial. *Educational and Psychological Measurement,* 64, 749-770.

Hayes, A. F., & Krippendorff, K. (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures,* 1, 77-89.

Holmes, C. B. (1979) Sample size in psychological research. *Perceptual and Motor Skills,* 49, 283-288.

Holmes, C. B. (1983) Sample size in four areas of psychological research. *Transactions of the Kansas Academy of Science,* 86(2-3), 76-80.

Holmes, C. B., Holmes, J. R., & Fanning, J. J. (1981) Sample size in non-APA journals. *The Journal of Psychology,* 108, 263-266.

Hotz, R. L. (2007) Most science studies appear to be tainted by sloppy analysis. *The Wall Street Journal,* September 14, p. B1. Retrieved on January 28, 2009, from http://online.wsj.com/article/SB118972683557627104.html

Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Medicine,* 2(8), e124. DOI:10.1371/journal.pmed.0020124.

Maddock, J. E., & Rossi, J. S. (2001) Statistical power of articles published in three health psychology-related journals. *Health Psychology,* 20, 76-78. DOI: 10.1037//0278-6133.20.1.76.

Maxwell, S. E. (2004) The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods,* 9(2), 147-163. DOI: 10.1037/1082-989X.9.2.147.

Merenda, P. F. (2007) Psychometrics and psychometricians in the 20th and 21st centuries: how it was in the 20th century and how it is now. *Perceptual and Motor Skills*, 104, 3-20.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2004) *HLM 6: hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Rodgers, J. L. (2010) The epistemology of mathematical and statistical modeling. *American Psychologist*, 65, 1-12.

Rossi, J. S. (1990) Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58, 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.

Wilkinson, L., & the Task Force on Statistical Inference. (1999) Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, 54, 594-604. DOI:10.1037/0003-066X.54.8.594.

Accepted February 11, 2011.