Supplementary Materials for

Bishara, A. J., Li, J., Conley, C. (in press). Informal versus formal judgment of statistical models: The case of normality assumptions. *Psychonomic Bulletin & Review*.

Supplement 1: Identification of 20 "Popular" Statistics Textbooks

Ten of these textbooks were identified by the Amazon bestseller list in the Statistics category, which captures primarily undergraduate texts because large undergraduate courses produce the demand that makes textbooks best-sellers. The other 10 textbooks were identified by a WorldCat database search sorted by the number of libraries holding the book. Libraries often avoid purchasing undergraduate textbooks, and library holdings are biased instead to favor more advanced and specialized textbooks that are requested by faculty members. Thus, the two sources (Amazon and WorldCat) are each biased, but toward different ends of the "introductory-advanced" continuum.

The Amazon list came from https://www.amazon.com/Best-Sellers-Books-Statistics/zgbs/books/491548/. Amazon's Best Seller lists are updated hourly. This list was obtained on August 31, 2016 at 3:35pm.

WorldCat holdings were examined with the search phrase: (ti:statistic* OR (ti:data AND ti:analysis)) AND (de:textbook* OR su:"Textbooks.") This phrase was decided upon after considering a list of 5 textbooks familiar to the authors, and then identifying terms that would include those books in the results while simultaneously minimizing the number of off-topic results. The search was conducted on August 31, 2016.

We excluded textbooks below the undergraduate level, as well as all non-textbooks (e.g., general audience books about statistics) and books that were not about statistics. Regardless of the specific edition identified in searches, we opted to review the most recent and longest edition available for each textbook (i.e., we replaced "essentials" or "fundamentals" versions with longer versions).

Supplement 2: Graph and Hypothesis Test Technical Details

For both graph and formal test methods, we assume that the observed random sample x_1, x_2, ..., x_n has been standardized to have a mean of 0 and standard deviation of 1.

Graphs

Q-Q plots. The theoretical Q-Q plots used here compare the sorted observed quantiles (y_i) against the quantiles theoretically expected under the standard normal distribution (m_i). If the sample is approximately normal, then increases in y_i should correspond to increases in m_i. More specifically, let $y_1 < y_2 < ... < y_n$ be the order statistics of the observed random sample x_1, x_2, ..., x_n, with no two observations being equal. The empirical distribution function is defined as:

$$F_n(x) = \begin{cases} 0, & x < y_1 \\ k/n, & y_k \le x < y_{k+1} \\ 1, & x \ge y_n \end{cases} \tag{1}$$

That is, the empirical distribution function is a "step" function that jumps $1/n$ in height at each observation x_k.

The actual expected order statistics of the standard normal distribution (m_i) are often difficult to compute. Intuitively, one could pair y_i with the quantiles for the standard normal distribution using cumulative probability, i/n:

$$\Phi^{-1}(i/n) \tag{1}$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function. However, this approach would pair y_n , the largest observation, with infinity. Instead, the expected order statistics are often approximated by one of several rank-based inverse normal transformations:

$$\hat{m}_i = \Phi^{-1}\left(\frac{i-a}{n-2a+1}\right) \tag{2}$$

where a is some constant (for reviews, see Harter, 1984; Thode, 2002). Here, a was set to .5, the default for Q-Q plots in the popular software package, R (R Core Team, 2016). Other common choices of a lead to similar results, as they are all approximately linear transformations of one another.

More consequential are the default settings for reference lines. By default, R's reference lines are drawn to connect the 1st and 3rd quartiles of the observed scores (e.g., Figure 1, row 2). This causes the slope of the line to vary from sample to sample, even when data are standardized.

P-P plots. The theoretical P-P plots used here compare the ordered observed percentiles plotted against the theoretical percentiles expected if the scores were normally distributed. As with a Stable Q-Q plot, if the sample is approximately normal, the pattern should be approximately linear along the diagonal.

Specifically, in the P-P Plot condition, the plotting position of score y_{i} was $\{\mathrm{i}/(n\text{-}1), \Phi(y_i)\}$, where i is the rank of the observation (i=1 for the lowest score, etc.), Φ is the cumulative normal distribution function, and y_i is the observed value for the i rank (i.e., y_i is the order statistic). If the data are from the standard normal distribution, the scatterplot should be close to the reference line $y = x$.

Formal Hypothesis Tests of Normality

Across all formal tests described below, the null hypothesis is that the data are drawn from a standard normal distribution, and the alternative hypothesis is that they are not.

Pearson Chi-Squared test statistic.

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}, \tag{3}$$

where k is the number of equal-probable subintervals that the whole real number line is divided into, O_i is the observed frequency and E_i is the expected frequency from the standard normal distribution in subinterval i (Pearson, 1900). The observed frequency should be close to the expected frequency under the null hypothesis, so the null hypothesis is rejected if χ^2 is large. χ^2 asymptotically approaches to chi-square distribution with $k-3$ degree of freedom under the null hypothesis.

One limitation of the Pearson chi-squared test statistic is that it depends on the choice of k. Here, k was set to be the ceiling of $2n^{2/5}$, which is a common setting (Moore, 1986, p. 70). A second limitation of this test is that it tends to have less power than other tests (e.g., Gan & Koehler, 1990; Shapiro & Wilk, 1965). The Pearson chi-squared test was included in the present investigation to provide at least one mechanical decision rule that relies on binning, much as do human decisions based on histograms.

Kolmogorov-Smirnov (Lilliefors) test statistic. Let $\Phi(x)$ be the cumulative distribution function of the standard normal distribution at $x \in$ R. The test statistic D_n is based on the least upper bound of the absolute difference between the empirical distribution function and the cumulative distribution function that would be expected if the population were standard normal:

$$D_n = \sup_x \left|F_n(x) - \Phi(x)\right|. \tag{4}$$

If there is, at any point z, a large difference between the empirical distribution $F_n(x)$ and the hypothesized $\Phi(x)$, it would suggest that the empirical distribution $F_n(x)$ does not equal the hypothesized $\Phi(x)$. Therefore, if D_n is too large, the null hypothesis is rejected. Let y_0 be such that $F_n(y_0)=0$, then the Kolmogorov-Smirnov (Lilliefors) test statistic is computed as the maximum of the differences:

$$\left|F_n(y_k) - \Phi(y_k)\right| \text{ and } \left|F_n(y_{k-1}) - \Phi(y_k)\right| \tag{5}$$

for $k=1,2,\ldots,n$.

The original Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948) required a simple null hypothesis with parameters fully specified. Lilliefors (1967) presented results for the more common situation where the population mean and standard deviation are not known, and this Lilliefors version is used here (see Conover, 1999, pp. 443-447, for critical values).

This test is almost universally recommended against as a test for normality because of its low power, especially in small samples (Thode, 2002, pp. 145, 150). It is included in the present investigation because it is well-known and is commonly described (though rarely endorsed) in statistics textbooks.

Anderson-Darling test statistic. A more powerful test based on the empirical distribution function is one developed by Anderson and Darling (1952, 1954):

$$A_n^2 = n\int_{-\infty}^{+\infty} \frac{\left[F_n(x) - \Phi(x)\right]^2}{\Phi(x)\left[1-\Phi(x)\right]} d\Phi(x). \tag{6}$$

A large A_n^2 would suggest that the empirical distribution $F_n(x)$ does not equal the hypothesized $\Phi(x)$. Therefore we reject the null hypothesis if A_n^2 is too large. One can compute the test statistic by

$$A_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\log u_{(j)} + \log(1-u_{(n-j+1)})\right], \tag{7}$$

where $u_{(j)} = \Phi(y_j)$ (see Stephens, 1976, for critical values). The Anderson-Darling test tends to show competitive power for many situations, and even higher power than that of the Shapiro-Wilk test for symmetric distributions with high kurtosis (Stephens, 1974).

Shapiro–Francia test statistic. The test statistic, W', is the squared Pearson correlation coefficient between the ordered observations and the expectations of the ordered observations under normality. In other words, it is a measure of the linearity of the Q-Q plot of the data. Specifically,

$$W' = \frac{\left(\sum_{i=1}^{n} y_i m_i\right)^2}{(n-1)\sum_{i=1}^{n} m_i^{\,2}}, \tag{8}$$

where m_{i} is the expectation (mean) of the i_{th} order statistic in a standard normal distribution (Shapiro & Francia, 1972; specifically, m was estimated here with Eq. 3, a=3/8). W' should be close to 1 if the null hypothesis of normality is true, and so a small W' leads to rejection of the null hypothesis (see Royston, 1983, for critical values).

Note that the Shapiro-Francia test is closely related to numerous other tests that depend on the correlation between the coordinates on some normal probability plot. For example, the Ryan-Joiner (1976) test relies on the unsquared rather than squared correlation; the test described by Filliben (1975) relies on the median instead of the mean of the order statistics (also see Gan & Koehler, 1990; Looney & Gulledge 1985, 1985b). The Shapiro-Francia test was developed as a simplified version of the Shapiro-Wilk test, described next.

Shapiro-Wilk test statistic.

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{n-1}, \tag{9}$$

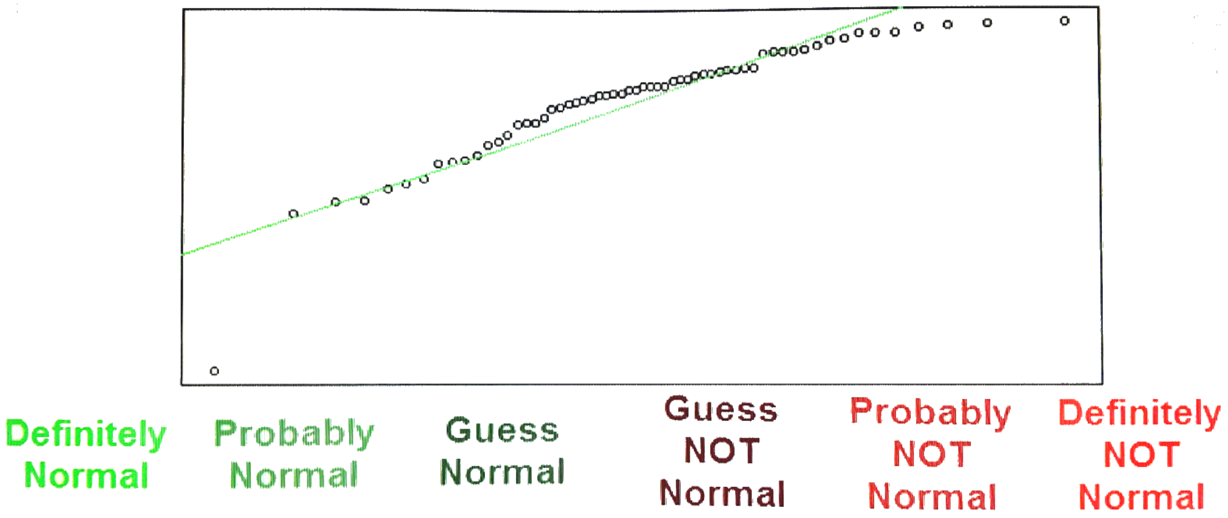
The a_i weights are defined as

$$\left(a_1,\mathrm{K}\ ,a_n\right) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}, \tag{10}$$

Where $m = (m_1, \ldots, m_n)^T$ are the expectations of the order statistics of a sample of size n from the standard normal distribution, and V is the covariance matrix of the order statistics (Shapiro & Wilk, 1965). The mean of W for non-normal distributions tends to shift to the left of that for the null case, at least based on heuristic considerations and extensive empirical sampling results (Shapiro & Wilk 1964). Therefore, the null hypothesis is rejected when W is small (see Shapiro

& Wilk, 1965, for critical values). Among the normality tests considered here, the Shapiro-Wilk test has shown the highest power for skewed and short-tailed symmetric alternatives and respectable power for long-tailed symmetric alternatives (Gan & Koehler, 1990; Stephens, 1974). Various other correlation tests have shown nearly equivalent power, with sometimes a slight advantage for the Shapiro-Wilk test (Looney & Gulledge, 1984, 1985).

Supplement 3: Procedure Screenshot



Note. Taken from Experiment 1 Q-Q Plot trial.

Supplement 4: Subjective Reports of Accuracy

	Subjective Accuracy		Objective Accuracy		Correlation (*r*) between Subjective & Objective Accuracy	
Experiment, Condition	Overall	Post-Train	Overall	Post-Train	Overall	Post-Train
Exp. 1						
Histogram	.59	.54	.71	.73	.39	.32
Q-Q Plot	.64	.68	.77	.80	.44	.26
Exp. 2						
Q-Q Plot	.67	.64	.80	.81	.52	.48
Stable Q-Q Plot	.71	.69	.84	.83	.38	.38
P-P Plot	.64	.64	.81	.81	.39	.18
Exp. 3						
Stable Q-Q Plot	.72	.63	.76	.77	.32	.02
Exp. 4						
Stable Q-Q Plot	.77	.72	.81	.80	-	-

Note. Subjective Accuracy shows participants' mean answers, converted to proportions, to the questions asked at the end of the experiment: "From 0 to 100, what percentage do you think you had correct overall? (counting any green button as correct for NORMAL and any red button as correct for NOT NORMAL)" and "In the very last series of graphs where you weren't told what the correct answer was, what percentage do you think you had correct?". In Exp. 3 and 4, the Subjective Accuracy questions did not distinguish between small and large sample sizes, and so objective accuracy is collapsed across those conditions. Because Exp. 4 had only 3 participants, correlations were not calculated.

Two general patterns are apparent. First, participants tended to underestimate their own accuracy, both Overall and in the Post-Training block. Second, there were modest but positive correlations between subjective and objective accuracy.

Supplement 5: Optimal Alphas for Formal Hypothesis Tests

For Experiments 3 and 4, we estimated optimal alphas for formal normality tests, that is, alphas that would maximize the long-run power of the selected tests of equal means (*t* vs. Mann-Whitney-Wilcoxon; MWW). For the 240 stimuli in a particular condition (e.g., Small Sample), let $n_1, n_2, \ldots, n_{240}$ be the normality test p-values for a particular test (e.g., Shapiro-Wilk) with those stimuli. For the corresponding stimuli, let $t_1, t_2, \ldots, t_{240}$ be the t-test powers to detect a medium size effect of $d = .5$, and $m_1, m_2, \ldots, m_{240}$ be the MWW test powers. Let α be some particular criterion value, not necessarily the customary .05, for the normality test. For stimulus *j*, if $n_j \geq \alpha$, the distribution is decided to be normal enough that a parametric t-test is chosen, with power t_j. If instead $n_j < \alpha$, the distribution is decided to be non-normal enough that a nonparametric MWW test is chosen, with power m_j. So, for a particular α of the normality test, the long-run power of testing equal means can then be estimated as

$$1 - \beta = \frac{1}{240} \Sigma_{j=1}^{240} (\mathbf{1}_{n_j \geq \alpha} \cdot t_j + \mathbf{1}_{n_j < \alpha} \cdot m_j) \, , \qquad (12)$$

where $\mathbf{1}_{expression}$ is an indicator function that has value 1 when the expression is true and 0 when it is false. To estimate the optimal criterion, that is, the alpha that maximizes power, we examined all $\alpha \in \{n_1, n_2, \ldots, n_{240}, 1\}$, saving the maximum value of $1\text{-}\beta$ and the α that achieved it.

These optimal alphas were established in a way that makes them favorable to this stimulus set, but not realistic enough for general use. Stimuli here were sampled from realistic marginal base-rates (Cain et al., 2017), but unrealistically, the sampling was stratified so that half of the stimuli were normal enough for a t-test to be more powerful, and half were non-normal enough for a MWW test to be more powerful. Additionally, Cain et al. report marginal skewness and kurtosis. It is unclear how similar marginal base-rates are to residual base-rates. Finally, because our approach was limited to sets of 240 stimuli, it allowed for only 241 possible values for the optimal alpha, so the estimates were coarse. Identifying optimal criteria for assumption tests is a challenging problem, and one worthy of more extensive investigation than the simple approach described here.

Supplement 6: Supplement References not Found in Article

Conover, W.J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: Wiley.

Harter, H. L. (1984). Another look at plotting positions. *Communications in Statistics-Theory and Methods, 13*, 1613-1633.

Looney, S. W., & Gulledge, T. R. (1984). Regression tests of fit and probability plotting positions. *Journal of Statistical Computation and Simulation, 20*, 115-127.

Looney, S. W., & Gulledge, T. R. (1985b). Probability plotting positions and goodness of fit for the normal distribution. *The Statistician, 34,* 297-303.

Pearson, A. V., & Hartley, H. O. (1972). *Biometrica tables for statisticians, Vol. 2*, Cambridge, England: Cambridge University Press.

Royston, J.P. (1983): A simple method for evaluating the Shapiro-Francia W' test for non-normality. *The Statistician, 32*, 297-300

Ryan, T. A., & Joiner, B. L. (1976). Normal probability plots and tests for normality. *Minitab Statistical Software: Technical Reports*. The Pennsylvania State University, State College, PA.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association, 69*, 730-737.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, *4*, 357-369.