DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2

Article 2

11-1-2013

Preliminary Testing for Normality: Is This a Good Practice?

H. J. Keselman
University of Manitoba, Winnipeg, Manitoba, kesel@ms.umanitoba.ca

Abdul R. Othman
Universiti Sains Malaysia, Georgetown, Penang, arothman60@yahoo.com

Rand R. Wilcox
University of Southern California, Los Angeles, rwilcox@usc.edu

Recommended Citation

Keselman, H. J.; Othman, Abdul R.; and Wilcox, Rand R. (2013) "Preliminary Testing for Normality: Is This a Good Practice?," *Journal of Modern Applied Statistical Methods*: Vol. 12: Iss. 2, Article 2.
Available at: http://digitalcommons.wayne.edu/jmasm/vol12/iss2/2

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 2-19.

ISSN 1538 − 9472

Invited Article: Preliminary Testing for Normality: Is This a Good Practice?

H. J. Keselman
University of Manitoba
Winnipeg, Manitoba

Abdul R. Othman
Universiti Sains Malaysia
Georgetown, Penang

Rand R. Wilcox
University of S. California
Los Angeles, CA

Normality is a distributional requirement of classical test statistics. In order for the test statistic to provide valid results leading to sound and reliable conclusions this requirement must be satisfied. In the not too distant past, it was claimed that violations of normality would not likely jeopardize scientific findings (See Hsu & Feldt, 1969; Lunney, 1970). Recent revelations suggest otherwise (See e.g., Micceri, 1989; Keselman, Huberty, Lix et al., 1998; Erceg-Hurn, Wilcox, & Keselman, 2013; Wilcox and Keselman, 2003; Wilcox, 2012a, b). Unfortunately the data obtained in psychological investigations rarely, if ever, meet the requirement of normally distributed data (Micceri, 1989; Wilcox, 2012a, b). Consequently, it could be the case that the results from many of the investigations conducted in psychology provide invalid results. Accordingly, authors recommend that researchers attempt to assess the validity of assuming data are normal in form prior to conducting a test of significance (Erceg-Hurn, et al., 2013; Keselman, et al., 1998). Present evidence suggests that a popular fit-statistic, the Kolmogorov-Smirnov test does a poor job of evaluating whether data are normal. Our investigation based on this statistic and other fit-statistics provides a more favorable picture of preliminary testing for normality.

Keywords: Assessing normality, fit statistics, g-and-h non-normal skewed and kurtotic data, contaminated mixed-normal distributions; outlying value(s), Likert scales

Introduction

Psychological researchers are often reminded that the validity of their statistical tests and the conclusions derived from these tests depends to a great extent on whether the derivational assumptions of the test procedures have been satisfied (e.g., See Keselman, Huberty, Lix et al., 1998; Wilcox, 2012a, b; Wilcox &

H. J. Keselman is a Professor of Psychology. His research interests are in applied statistics. Email him at: kesel@ms.umanitoba.ca. Abdul R. Othman is a Professor of Statistics. Email him at: arothman60@yahoo.com. Rand R. Wilcox is a Professor of Psychology. Email him at: rwilcox@usc.edu.

Keselman, 2003). Consequently, though not a common practice, researchers are still reminded about assessing derivational assumptions (See Erceg-Hurn, Wilcox, & Keselman, 2013; Kirk, 2013; Schoder, Himmelmann & Wilhelm, 2006; Wilcox & Keselman, 2003). Almost all inferential methods require that in the population(s) the data is (are) normally distributed (as well as other requirements not relevant to this paper). Violation of the normality assumption can have a deleterious effect on the Type I error rate of test statistics (See Wilcox, 2012a, b; Wilcox & Keselman, 2003). Although the Type I error rate is widely viewed as being relatively unaffected by non-normality, Bradley (1980) has pointed out conditions in which this is not true. This finding is also evident in the findings of recent studies and published texts (e. g., See Hempel, Ronchetti, & Rousseeuw, 1986; Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006; Micceri, 1989; Schoder, et al., 2006; Staudte & Sheather, 1990; Wilcox, 2012a, b; Wilcox & Keselman, 2003).

Applied researchers can examine plots of their data and/or perform tests to assess the assumption, i. e., normality. Evaluating graphs (e.g., box-plots, stem-and-leaf, box and whisker, QQ plots) of ones data to assess whether data are normally distributed can be problematic since the determination relies on a subjective assessment (Wilk & Gnanadesikan, 1968). Thus, this practice is oftentimes not typically used when assessing the shape of the distribution of data (See Schoder, et al., 2006). Researchers tend to prefer exact methods based on formal tests for normality such as the Kolmogorov-Smirnov (K-S) goodness-of-fit statistic (See Muller & Fetterman, 2002, Chapter 7). Furthermore, researchers commonly use the result from a goodness-of-fit test to determine whether the normality of classical test procedures is satisfied thus providing legitimacy to the use of a classical test statistic. Consequently, preliminary testing for normality or any distributional shape is quite important in the whole inferential process and has been discussed in various contexts (See e.g., Cardoso de Oliveira & Ferreira, 2010; Doornik & Hansen, 2008; Sürücü, 2006). However, if the assumption of normality does not appear to be satisfied, researchers use this information to select alternative procedures such as nonparametric methods. Thus, it is important to know how well a preliminary test for normality, e. g., the K-S test, works in detecting non-normal data.

Unfortunately, according to Schoder, Himmelman, and Wilhelm (2006) “The Kolmogorov-Smirnov test performs badly on data with single outliers, 10% outliers, and skewed data at sample sizes <100.” (p. 757) These authors investigated the performance of the K-S test for four types of non-normal data (e.g., normal distribution with a single outlier, normal distribution with 10%

outliers, skewed lognormal distribution with varying skewness, and an ordinal 5-point Likert scale with varying multinomial probabilities) and varying sample size in a pretest-posttest design. The assessment for normality was conducted at a 5% significance level. Unfortunately, the results tabled by Schoder et al. do not support the use of the K-S test as a preliminary test to assess normality of the data.

Because it is strongly believed that validity assumptions such as normality should be verified before adopting a classical test of significance that assumes the data in the population is normal in shape, it important to replicate the findings reported by Schoder, Himmelmann, and Wilhelm (2006) and extend their study in important ways. (For a contrary view previously noted in this journal, see Sawilowsky, 2002, p. 466-467). Other goodness-of-fit statistics are available (see, e.g., Muller & Fetterman, 2002, Chapter 7). Accordingly, a simulation study was conducted investigating three goodness-of-fit statistics, varying the degree of non-normality with other distributional shapes not investigated by Schoder, Himmelmann, and Wilhelm (2006), using sample sizes more likely to be encountered in psychological and educational research.

Method

Specifically, in this study the following are manipulated: (1) the procedure used to assess shape of distribution [K-S, Cramer-von Mises (CvM), Anderson-Darling (A-D)] fit-statistics (available through the SAS system), (2) the shapes of distributions (26 cases—14 g-and-h distributions, 8 contaminated normal mixture models, and 4 multinomial models), (3) the sample sizes (20, 40, and 80), depending on distribution, and (4) the level of significance for the fit-statistics (i.e., $\alpha = .05, .10, .15 \text{ and } .20$).

Most statistical packages (e.g., the SAS system) provide numerous fit statistics. Accordingly, it is possible that other tests other than the K-S can adequately assess whether normality, or other distributions, exists in the data. The SAS system was used to implement the K-S, CvM, and A-D fit-statistics. The choices of non-normal distributions are modifications from Schoder, Himmelmann and Wilhelm (2006) and Zimmerman (1998). Schoder, et al. (2006) investigated a normal distribution with a single outlier, a normal distribution with 10% outliers, skewed lognormal distributions with varying skewness, and an ordinal 5-point Likert scale with varying multinomial probabilities (common they state in dermatological investigations). Many non-normal distributions were investigated via g-and-h distributions (See Headrick, Kowalchuk, & Sheng, 2008; Hoaglin, 1983; 1985; Kowalchuk & Headrick, 2010; Tukey, 1960). These distributions with their values for skewness and kurtosis are enumerated in Table 1. A substantial number of values of g and h were chosen to cover as broad a spectrum of non-normal distributions that could occur in psychological and behavioral science experiments (e.g., See Keselman, Huberty, Lix et al., 1998; Micceri, 1989; Wilcox, 2012a, b).

Table 1. g-and h-distributions examined in the simulation study with their corresponding measures of skewness and kurtosis

Distribution	Skewness	Kurtosis	Distribution	Skewness	Kurtosis
g=0,h=.05	0.00	0.82	g=.4,h=0	1.32	3.26
g=0,h=.075	0.00	1.49	g=.6,h=0	2.26	10.27
g=0,h=.1	0.00	2.51	g=1,h=0	6.19	110.94
g=0,h=.125	0.00	4.16	g=.2, h=.1	1.08	5.50
g=0,h=.15	0.00	7.17	g=.4,h=.1	2.45	20.30
g=0,h=.2	0.00	33.22	g=.6,h=.1	4.69	89.80
g=.2, h=0	0.61	0.68	g=.8,h=.1	9.27	603.61

Table 2. Contaminated mixed-normal distributions used in the power studies of the three goodness-of-fit-tests for normality

		Outliers	
n	Distribution	Distance (in standard Deviations)	Number
20	(.95)N(0,1) + (.05)N(0,25)	5	1
20	(.90)N(0,1) + (.10)N(0,25)	5	2
20	(.95)N(0,1) + (.05)N(0,100)	10	1
20	(.90)N(0,1) + (.10)N(0,100)	10	2
40	(.975)N(0,1) + (.025)N(0,25)	5	1
40	(.95)N(0,1) + (.05)N(0,25)	5	2
40	(.90)N(0,1) + (.10)N(0,25)	5	4
40	(.975)N(0,1) + (.025)N(0,100)	10	1
40	(.95)N(0,1) + (.05)N(0,100)	10	2
40	(.90)N(0,1) + (.10)N(0,100)	10	4
80	(.9875)N(0,1) + .(0125)N(0,25)	5	1
80	(.975)N(0,1) + (.025)N(0,25)	5	2
80	(.95)N(0,1) + (.05)N(0,25)	5	4
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1
80	(.975)N(0,1) + (.025)N(0,100)	10	2
80	(.95)N(0,1) + (.05)N(0,100)	10	4

The SAS system was used on a Sun Fire X4600 M2 x64 server: 8 x AMD Opteron Model 8220 processor (2.8GHz-dual-core) to generate g- and-h data, by modifying standard normal variates $Z \sim N(0,1)$ to non-normal variates by specifying values of g and h in the following quantile functions:

$$q(Z) = q_{g,h}(Z) = \frac{\exp(gZ)-1}{g}\exp\left(\frac{hZ^2}{2}\right), \tag{1}$$

$$q(Z) = q_{g,0}(Z) = \frac{\exp(gZ)-1}{g}, \tag{2}$$

$$q(Z) = q_{0,h(Z)} = Z \;\; \exp\left(\frac{hZ^2}{2}\right) \tag{3}$$

Equations (2) and (3) generate lognormal and symmetric h distributions, respectively. As Kowalchuk and Headrick (2010) noted "The parameter $\pm g$ controls the skew of a distribution in terms of both direction and magnitude. The parameter h controls the tail weight or elongation of a distribution and is positively related with kurtosis." (p. 63). As well, Type I error rates were investigated when data were obtained from a normal distribution [$g = h = 0$, the standard normal distribution (skewness and kurtosis = 0)].

A number of different contaminated mixed-normal distributions were examined, such as those reported in Zimmerman (1998). Contaminated mixed-normal distributions have one or more outlying values that deviate from the central mean of the distribution by some amount measured in standard deviation units. For example, Zimmerman examined a mixed normal distribution consisting of samples from $N(0,1)$ with probability .95 and from $N(0,400)$ with probability .05. Tukey (1960) suggested that outliers are a common occurrence in distributions and others have indicated that skewed distributions frequently depict psychological data (e.g., reaction time data). Accordingly, eight contaminated mixed normal distributions were examined that had one, two, or four outlying values which were five or ten standard deviations from the mean value. These distributions are enumerated in Table 2.

Finally, like Schoder, Himmelmann and Wilhelm (2006), a 5-point Likert scale was simulated; such data is frequently gathered in psychological (e.g., from clinical, personality, and social psychological investigations) and other behavioral science investigations. The same conditions investigated by Schoder et al. (2006) were investigated. Specifically,

1) even distribution (*p*=.02 for each category 0-4);
2) symmetric distribution
($p_0 = 0.1,\ p_1 = 0.2,\ \mathrm{p}_2 = 0.4,\ p_3 = 0.2,\ p_4 = 0.1$) ;
3) moderately skewed distribution
($p_0 = 0.5,\ p_1 = 0.3,\ p_2 = 0.15,\ p_3 = 0.04,\ p_4 = 0.01$) ; and
4) heavily skewed distribution
($p_0 = 0.7,\ p_1 = 0.2,\ \mathrm{p}_2 = 0.06,\ p_3 = 0.03,\ p_4 = 0.01$) .

Thus, for the 5-point Likert scale data there were 4 multinomial distributions that were simulated (See Table 3).

Table 3. Multinomial distributions based upon Schoder, Himmelmann, and Wilhelm's (2006) probabilities simulated as a five-point Likert Scale

	Even	Symmetric	Moderately Skewed	Heavily Skewed
n	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)
20	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)
40	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)
80	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)

The same number of sample size conditions as Schoder, Himmelmann, and Wilhelm (2006) were not investigated, but a reasonable range of values (i.e., n = 20,40,80) were incldled, depending on the condition investigated. Specifically,

(i) for the 14 g- and h- distributions, and 2 contaminated normal distributions, .95*N*(0,1) + .05*N*(0, *k*), *k*=25, 100, sample sizes of 20, 40 and 80 were chosen.
(ii) For 2 contaminated normal distributions, .9*N*(0,1) + .1*N*(0,*k*), *k*=25, 100, sample sizes of 20 and 40 were chosen.
(iii) For 2 contaminated normal distributions, .975*N*(01) + .025*N*(0,*k*), *k*=25, 100, sample sizes of 40 and 80 were chosen.
(iv) For 2 contaminated normal distributions, .9875*N*(0,1) + .0125*N*(0,k), *k*=25, 100, sample size of 80 was chosen.

Lastly, because in preliminary testing it would be quite important to guard against a Type II error (falsely accepting the null hypothesis that the data are normal in form), we selected significance levels of .10 , .15, and .20, in addition to the standard .05. Each condition in the investigation was replicated 5,000 times.

Table 4. Power rates for the goodness-of-fit test on normality (*n* = 20).

	Distribution	**Skewness**	**Kurtosis**	**α = .05**	**α = .10**	**α = .15**
Kolmogorov-Smirnov	***Normal****	***0.00***	***0.00***	***0.0524***	***0.1082***	***0.1530***
	g=0,h=.05	0.00	0.82	0.0726	0.1304	0.1834
	g=0,h=.075	0.00	1.49	0.0870	0.1540	0.2094
	g=0,h=.1	0.00	2.51	0.1066	0.1838	0.2392
	g=0,h=.125	0.00	4.16	0.1320	0.2156	0.2726
	g=0,h=.15	0.00	7.17	0.1626	0.2502	0.3100
	g=0,h=.2	0.00	33.22	0.2296	0.3194	0.3834
	g=.2, h=0	0.61	0.68	0.1030	0.1678	0.2286
	g=.4,h=0	1.32	3.26	0.2436	0.3506	0.4262
	g=.6,h=0	2.26	10.27	0.4450	0.5662	0.6416
	g=1,h=0	6.19	110.94	0.7852	0.8648	0.9008
	g=.2, h=.1	1.08	5.50	0.1662	0.2530	0.3100
	g=.4,h=.1	2.45	20.30	0.3218	0.4204	0.4842
	g=.6,h=.1	4.69	89.80	0.5018	0.6026	0.6642
	g=.8,h=.1	9.27	603.61	0.6698	0.7602	0.8096
Cramer-von Mises	***Normal****	***0.00***	***0.00***	***0.0494***	***0.1036***	***0.1490***
	g=0,h=.05	0.00	0.82	0.0752	0.1368	0.1952
	g=0,h=.075	0.00	1.49	0.0970	0.1658	0.2260
	g=0,h=.1	0.00	2.51	0.1286	0.1996	0.2632
	g=0,h=.125	0.00	4.16	0.1608	0.2426	0.3038
	g=0,h=.15	0.00	7.17	0.1990	0.2842	0.3400
	g=0,h=.2	0.00	33.22	0.2756	0.3580	0.4232
	g=.2, h=0	0.61	0.68	0.1100	0.1814	0.2444
	g=.4,h=0	1.32	3.26	0.3082	0.4064	0.4872
	g=.6,h=0	2.26	10.27	0.5570	0.6590	0.7204
	g=1,h=0	6.19	110.94	0.8822	0.9268	0.9484
	g=.2, h=.1	1.08	5.50	0.1990	0.2826	0.3454
	g=.4,h=.1	2.45	20.30	0.3808	0.4730	0.5370
	g=.6,h=.1	4.69	89.80	0.5922	0.6728	0.7216
	g=.8,h=.1	9.27	603.61	0.7594	0.8552	0.8626
Anderson-Darling	***Normal****	***0.00***	***0.00***	***0.0494***	***0.1036***	***0.1490***
	g=0,h=.05	0.00	0.82	0.0810	0.1456	0.2040
	g=0,h=.075	0.00	1.49	0.1090	0.1816	0.2378
	g=0,h=.1	0.00	2.51	0.1444	0.2162	0.2766
	g=0,h=.125	0.00	4.16	0.1784	0.2582	0.3200
	g=0,h=.15	0.00	7.17	0.2182	0.2992	0.3590
	g=0,h=.2	0.00	33.22	0.2924	0.3798	0.4386
	g=.2, h=0	0.61	0.68	0.1222	0.1966	0.2584
	g=.4,h=0	1.32	3.26	0.3388	0.4456	0.5258
	g=.6,h=0	2.26	10.27	0.6012	0.6988	0.7528
	g=1,h=0	6.19	110.94	0.9086	0.9448	0.9602
	g=.2, h=.1	1.08	5.50	0.2190	0.2984	0.3610
	g=.4,h=.1	2.45	20.30	0.4084	0.4968	0.5590
	g=.6,h=.1	4.69	89.80	0.6168	0.6972	0.7444
	g=.8,h=.1	9.27	603.61	0.7876	0.8474	0.8766

*Type 1 error rates

Table 5. Power rates for the goodness-of-fit test on normality (*n* = 40).

	Distribution	Skewness	Kurtosis	*α* = .05	*α* = .10	*α* = .15
Kolmogorov-Smirnov	Normal*	0.00	0.00	0.0524	0.1082	0.1530
	g=0,h=.05	0.00	0.82	0.0726	0.1304	0.1834
	g=0,h=.075	0.00	1.49	0.0870	0.1540	0.2094
	g=0,h=.1	0.00	2.51	0.1066	0.1838	0.2392
	g=0,h=.125	0.00	4.16	0.1320	0.2156	0.2726
	g=0,h=.15	0.00	7.17	0.1626	0.2502	0.3100
	g=0,h=.2	0.00	33.22	0.2296	0.3194	0.3834
	g=.2, h=0	0.61	0.68	0.1030	0.1678	0.2286
	g=.4,h=0	1.32	3.26	0.2436	0.3506	0.4262
	g=.6,h=0	2.26	10.27	0.4450	0.5662	0.6416
	g=1,h=0	6.19	110.94	0.7852	0.8648	0.9008
	g=.2, h=.1	1.08	5.50	0.1662	0.2530	0.3100
	g=.4,h=.1	2.45	20.30	0.3218	0.4204	0.4842
	g=.6,h=.1	4.69	89.80	0.5018	0.6026	0.6642
	g=.8,h=.1	9.27	603.61	0.6698	0.7602	0.8096
Cramer-von Mises	Normal*	0.00	0.00	0.0564	0.1068	0.1542
	g=0,h=.05	0.00	0.82	0.0950	0.1622	0.2212
	g=0,h=.075	0.00	1.49	0.1332	0.2094	0.2746
	g=0,h=.1	0.00	2.51	0.1860	0.2692	0.3328
	g=0,h=.125	0.00	4.16	0.2448	0.3314	0.3996
	g=0,h=.15	0.00	7.17	0.3132	0.4012	0.4722
	g=0,h=.2	0.00	33.22	0.4490	0.5294	0.5908
	g=.2, h=0	0.61	0.68	0.1936	0.2786	0.3474
	g=.4,h=0	1.32	3.26	0.5528	0.6618	0.7352
	g=.6,h=0	2.26	10.27	0.8628	0.9116	0.9360
	g=1,h=0	6.19	110.94	0.9948	0.9980	0.9990
	g=.2, h=.1	1.08	5.50	0.3286	0.4220	0.4894
	g=.4,h=.1	2.45	20.30	0.6394	0.7218	0.7738
	g=.6,h=.1	4.69	89.80	0.8728	0.9120	0.9308
	g=.8,h=.1	9.27	603.61	0.9664	0.9798	0.9866
Anderson-Darling	Normal*	0.00	0.00	0.0564	0.1024	0.1556
	g=0,h=.05	0.00	0.82	0.1036	0.1724	0.2326
	g=0,h=.075	0.00	1.49	0.1504	0.2278	0.2968
	g=0,h=.1	0.00	2.51	0.2082	0.2978	0.3612
	g=0,h=.125	0.00	4.16	0.2766	0.3678	0.4288
	g=0,h=.15	0.00	7.17	0.3460	0.4326	0.4960
	g=0,h=.2	0.00	33.22	0.4740	0.5620	0.6216
	g=.2, h=0	0.61	0.68	0.2130	0.3046	0.3750
	g=.4,h=0	1.32	3.26	0.6130	0.7160	0.7776
	g=.6,h=0	2.26	10.27	0.8946	0.9398	0.9586
	g=1,h=0	6.19	110.94	0.9974	0.9988	0.9998
	g=.2, h=.1	1.08	5.50	0.3556	0.4522	0.5194
	g=.4,h=.1	2.45	20.30	0.6698	0.7510	0.7956
	g=.6,h=.1	4.69	89.80	0.8958	0.9252	0.9444
	g=.8,h=.1	9.27	603.61	0.9738	0.9858	0.9890

*Type 1 error rates

Table 6. Power rates for the goodness-of-fit test on normality (*n* = 80).

	Distribution	Skewness	Kurtosis	*α* = .05	*α* = .10	*α* = .15
Kolmogorov-Smirnov	**_Normal*_**	**_0.00_**	**_0.00_**	**_0.0534_**	**_0.1082_**	**_0.1580_**
	g=0,h=.025	0.00	0.35	0.0696	0.1314	0.1828
	g=0,h=.05	0.00	0.82	0.0968	0.1742	0.2252
	g=0,h=.075	0.00	1.49	0.1446	0.2318	0.2980
	g=0,h=.1	0.00	2.51	0.2114	0.3172	0.3928
	g=0,h=.125	0.00	4.16	0.3012	0.4194	0.4934
	g=0,h=.15	0.00	7.17	0.3950	0.5154	0.5958
	g=0,h=.2	0.00	33.22	0.5904	0.6938	0.7526
	g=0,h=.225	0.00	154.84	0.6736	0.7624	0.8098
	g=.2, h=0	0.61	0.68	0.2530	0.3758	0.4494
	g=.4,h=0	1.32	3.26	0.7334	0.8334	0.8792
	g=.6,h=0	2.26	10.27	0.9692	0.9872	0.9936
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.4448	0.5604	0.6296
	g=.4,h=.1	2.45	20.30	0.8196	0.8870	0.9132
	g=.6,h=.1	4.69	89.80	0.9762	0.9882	0.9932
	g=.8,h=.1	9.27	603.61	0.9982	1.0000	1.0000
Cramer-von Mises	**_Normal*_**	**_0.00_**	**_0.00_**	**_0.0558_**	**_0.1030_**	**_0.1512_**
	g=0,h=.05	0.00	0.82	0.1194	0.1872	0.2480
	g=0,h=.075	0.00	1.49	0.1896	0.2740	0.3442
	g=0,h=.1	0.00	2.51	0.2792	0.3834	0.4538
	g=0,h=.125	0.00	4.16	0.3912	0.4936	0.5570
	g=0,h=.15	0.00	7.17	0.5004	0.5990	0.6626
	g=0,h=.2	0.00	33.22	0.6914	0.7654	0.8128
	g=.2, h=0	0.61	0.68	0.3172	0.4314	0.5108
	g=.4,h=0	1.32	3.26	0.8526	0.9082	0.9372
	g=.6,h=0	2.26	10.27	0.9950	0.9980	0.9990
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.5402	0.6346	0.6942
	g=.4,h=.1	2.45	20.30	0.8926	0.9302	0.9498
	g=.6,h=.1	4.69	89.80	0.9928	0.9968	0.9982
	g=.8,h=.1	9.27	603.61	1.0000	1.0000	1.0000
Anderson-Darling	**_Normal*_**	**_0.00_**	**_0.00_**	**_0.0548_**	**_0.1046_**	**_0.1526_**
	g=0,h=.05	0.00	0.82	0.1316	0.2112	0.2694
	g=0,h=.075	0.00	1.49	0.2158	0.3046	0.3804
	g=0,h=.1	0.00	2.51	0.3196	0.4220	0.4946
	g=0,h=.125	0.00	4.16	0.4328	0.5290	0.5996
	g=0,h=.15	0.00	7.17	0.5420	0.6396	0.7004
	g=0,h=.2	0.00	33.22	0.7270	0.7960	0.8358
	g=.2, h=0	0.61	0.68	0.3606	0.4802	0.5604
	g=.4,h=0	1.32	3.26	0.8982	0.9430	0.9608
	g=.6,h=0	2.26	10.27	0.9976	0.9996	0.9998
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.5816	0.6692	0.7234
	g=.4,h=.1	2.45	20.30	0.9104	0.9424	0.9608
	g=.6,h=.1	4.69	89.80	0.9942	0.9976	0.9986
	g=.8,h=.1	9.27	603.61	1.0000	1.0000	1.0000

*Type 1 error rates

Table 7. Number of times the g- and h- non-normal power values are equal to or greater than .80 for the fit-statistics (K-S, CvM, and A-D)

	n	K-S	CvM	A-D	
***α* = .05**	20	0	1	1	
	40	3	4	4	
	80	5	6	6	
	Total	**8**	**11**	**11**	**30**
***α* = .10**	20	1	2	2	
	40	4	4	4	
	80	6	7	8	
	Total	**11**	**13**	**14**	**38**
***α* = .15**	20	2	2	2	
	40	4	4	5	
	80	7	8	8	
	Total	**13**	**14**	**15**	**42**
***α* = .20**	20	**---**	2	2	
	40	**---**	5	6	
	80	**---**	8	8	
	Total	**---**	**15**	**16**	**31***
Grand Total			**53**	**56**	

Note: --- and *: PROC UNIVARIATE in SAS does not provide exact p-values for K-S at α = .20

Results

g- and h- Non-normal Distributions

Table 4 presents Type I error and power rates for the K-S, CvM, and A-D fit-statistics when sample size was 20. A number of conclusions can be drawn from this table. First, Type I error was controlled for each level of significance. Second, for the non-normal alternatives investigated, the K-S was typically the least powerful procedure, followed by CvM, and the A-D is typically most powerful. Also evident from the data is that for kurtotic data, none of the procedures displayed reasonable power (i.e., >.80). Although for skewed and kurtotic data the fit-statistics were only reasonably powerful for extreme departures from normality. As expected, power to detect non-normal distributions increased with more liberal levels of significance; we excluded the $\alpha = .20$ values from the tables since the values are naturally larger than those reported for the other significance levels examined.

For moderate sample size case (See Table 5) the same pattern of results held; however, the fit-statistics had more power to detect non-normal data when sample size was 40. Finally, the same pattern of results occurred for our largest investigated sample size of 80 (See Table 6). And as expected the power to detect non-normal data increased with the increase in sample size.

To summarize the findings for the g-and-h non-normal distributions examined in this study we provide in Table 7 a count of the number of times the power values were equal to or greater than .80 across the simulated conditions. Over the significance levels that can be used with the K-S test (i.e., $\alpha = .05, .10,$ and $.15$) the A-D procedure was most powerful to detect non-normal distributions, followed closely by CvM and then by K-S. Clearly the A-D is most sensitive of the three. Also most evident is that the power to detect non-normal distributions is affected by the level of significance as would be expected. Also evident is that contrary to the warning given by Schoder, Himmelmann, and Wilhelm (2006) researchers can detect non-normal distributions with sample sizes less than 100 (80 in our case).

Contaminated Mixed-Normal Distributions

The power rates for the contaminated normal distributions for the three fit-statistics, K-S, CvM, and A-D are contained in Tables 8, 9, and 10, respectively. As we found for the g- and- h non-normal data, the A-D fit-statistic was most powerful for detecting normal data with outlying values than both the CvM and K-S fit-statistics. And, as expected, power increased with sample size and level of significance. Indeed, to a large extent the reported power values are in reasonably close proximity to .80 for most of the contaminated normal distributions examined. Furthermore, again, as expected the power values were largest when the level of significance was > .05.

Likert Non-normal data

The final type of non-normal data that we investigated was data that is obtained when five-point Likert scales are used in measuring the dependent variable. Subjects in the investigations indicate their preference, liking, attitude, etc. on five point type scales (e.g., very unfavorable, unfavorable, neutral, pleasant, very pleasant). Such responses obviously cannot be normally distributed.

Table 8. Power of the Kolmogorov-Smirnov goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.3238	0.4040	0.4568
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.4950	0.5748	0.6270
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6022	0.6526	0.6866
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8164	0.8566	0.8782
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.2898	0.3670	0.4240
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.4630	0.5424	0.5988
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.7050	0.7748	0.8144
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.5838	0.6462	0.6864
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8160	0.8520	0.8732
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9660	0.9768	0.9818
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.2472	0.3210	0.3804
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.4144	0.5006	0.5572
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.6754	0.7482	0.7852
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.5436	0.6052	0.6464
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.7874	0.8288	0.8546
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9606	0.9714	0.9778

Table 9. Power of the Cramer-von Mises goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.3692	0.4362	0.4844
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.5582	0.6220	0.6700
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6386	0.6844	0.7164
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8534	0.8802	0.8962
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.3374	0.4000	0.4590
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.5346	0.6018	0.6428
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.7776	0.8264	0.8540
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.6250	0.6698	0.7028

Table 9, continued. Power of the Cramer-von Mises goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Numb er**	***α*** **=.05**	***α*** **=.10**	***α*** **=.15**
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8478	0.8716	0.8864
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9784	0.9836	0.9868
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.2928	0.3596	0.4150
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.4884	0.5548	0.6086
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.7534	0.8002	0.8298
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.5924	0.6366	0.6714
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.8258	0.8580	0.8768
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9736	0.9800	0.9836

Table 10. Power of the Anderson-Darling goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α*** **=.05**	***α*** **=.10**	***α*** **=.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.4024	0.4650	0.5136
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.5974	0.6596	0.6994
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6688	0.7100	0.7368
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8704	0.8922	0.9076
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.3802	0.4466	0.4944
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.5860	0.6432	0.6854
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.8174	0.8558	0.8762
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.6572	0.7000	0.7276
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8664	0.8920	0.9056
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9824	0.9866	0.9896
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.3356	0.4050	0.4576
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.5460	0.6138	0.6574
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.8036	0.8440	0.8694
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.6284	0.6726	0.7042
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.8568	0.8830	0.8990
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9792	0.9850	0.9886

Table 11. Power of the goodness–of-fit test on normality of data for multinomial data representing five-point Likert scale scores

Kolmogoroc-Smirnov*

Even						*Symmetric*					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.2, .2, .2, .2, .2)	0.2742	0.4490	0.5630		10	(.1, .2, .4, .2, .1)	0.4568	0.6164	0.7212	
20	(.2, .2, .2, .2, .2)	0.6368	0.8248	0.8918		20	(.1, .2, .4, .2, .1)	0.8132	0.9390	0.9606	
40	(.2, .2, .2, .2, .2)	0.9978	1.0000	1.0000		40	(.1, .2, .4, .2, .1)	0.9998	1.0000	1.0000	

Moderately Skewed						*Heavily Skewed*					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.5, .3, .15, .04, .01)	0.7629	0.9176	0.9530		10	(.7, .2, .06, .03, .01)	0.9747	0.9905	0.9971	
20	(.5, .3, .15, .04, .01)	0.9970	0.9992	1.0000		20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	
40	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000		40	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	

Cramer-von Mises

Even						Symmetric					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.2, .2, .2, .2, .2)	0.2610	0.4000	0.5348	0.6620	10	(.1, .2, .4, .2, .1)	0.4520	0.5596	0.7008	0.7714
20	(.2, .2, .2, .2, .2)	0.6710	0.9060	0.9946	1.0000	20	(.1, .2, .4, .2, .1)	0.8494	0.9664	0.9986	1.0000
40	(.2, .2, .2, .2, .2)	0.9978	1.0000	1.0000	1.0000	40	(.1, .2, .4, .2, .1)	1.0000	1.0000	1.0000	1.0000

Moderately Skewed						Heavily Skewed					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.5, .3, .15, .04, .01)	0.8501	0.9134	0.9734	0.9814	10	(.7, .2, .06, .03, .01)	0.9825	0.9916	0.9981	0.9988
20	(.5, .3, .15, .04, .01)	0.9998	1.0000	1.0000	1.0000	20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000
40	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000	1.0000	40	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000

Anderson-Darling[b]

Even						Symmetric					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.2, .2, .2, .2, .2)	0.3202	0.5086	0.6220	0.7250	10	(.1, .2, .4, .2, .1)	0.4248	0.5820	0.6628	0.7898
20	(.2, .2, .2, .2, .2)	0.8420	0.9888	1.0000	1.0000	20	(.1, .2, .4, .2, .1)	0.8668	0.9916	1.0000	1.0000

Moderately Skewed						Heavily Skewed					
n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20	n	(p_0,p_1,p_2,p_3,p_4)	*α* = .05	*α* = .10	*α* = .15	*α* = .20
10	(.5, .3, .15, .04, .01)	0.8996	0.9526	0.9738	0.9928	10	(.7, .2, .06, .03, .01)	0.9897	0.9969	0.9988	0.9996
20	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000	1.0000	20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000

*PROC UNIVARIATE does not allow α = .20 for the Kolmogorov-Smirnov test.

[b] The power values for non-tabled n = 40 values are all 1.000

Table 11 provides power rates for the three fit-statistics for detecting non-normality arising from using a Likert scale for assessing the dependent variable. Preliminary findings indicated that power values were 100% for sample sizes greater than 20 in the vast majority of cases. Thus, it was decided to include a smaller sample size case (i.e., $n = 10$) to examine power values for a relatively modest number of subjects. The findings are quite positive; that is, in just about every case examined, the power to detect non-normality is > .80. Indeed, out of the 106 tabled values 83 are greater in value than .80. Once again, the A-D statistic provides the best power values, followed by CvM, and then by K-S.

Discussion

Applied researchers use statistical tests to assess whether or not the effect of an experimental manipulation is significant. Unfortunately, the results of many of these investigations are suspect as they often involve the use of statistical procedures with questionable validity. In these cases, the reported effects may be misleading or, in many cases, wrong. Clearly, such erroneous decisions can have serious negative consequences for both the advancement of knowledge in a given field as well as the effective translation of research results into practice. The intent of this paper was to examine whether one can effectively test whether one's data confirms to the validity assumption of normality—a requirement for most classical test statistics. Prior research suggested that one could not use the Kolmogorov-Smirnov goodness-of-fit test to effectively test whether data were normally distributed or not (See e.g., Schoder, Himmelmann, and Wilhelm, 2006).

We looked into this negative finding by also investigating other fit statistics, the Cramer von Mises and Anderson-Darling tests (See Muller & Fetterman, 2002 Chapter 7), varying the skewness and kurtosis values of numerous g-and h-distributions, examining a number of contaminated mixed-normal distributions and examining results when the dependent variable was obtained from non-normal five-point Likert data. We also manipulated sample sizes (n = 20,40,80 and the level of significance for the test of normality $\alpha = .05, .10, .15$ and $.20$).

Of the three fit-statistics we found that the Anderson-Darling procedure was most effective in detecting non-normality being superior to both the Kolmogorov-Smirnov and Cramer-von Mises tests. We also determined that one could reasonably detect non-normality with reasonable sample sizes (n = 10,20,40), unlike what was reported by Schoder, Himmelmann, and Wilhelm (2006). Lastly, and importantly, since in this context one would want to increase the power to detect effects and concomitantly reduce the probability of falsely accepting the null hypothesis that data are normally distributed, we suggest that preliminary testing be performed with significance levels larger than .05, say $\alpha = .15$ or $\alpha = .20$.

We conclude by reminding researchers that if normality is not present in the data current analytic practices allow researchers to test hypotheses say about mean equality in multiple group designs with software that does not require that data be normally distributed (See e. g., SAS's Glimmix procedure). Or, researchers can choose to replace classical test statistics and their least squares estimators for the mean and variance with robust test statistics with robust estimators (i.e., trimmed means and Winsorized variances (See e.g., Wilcox, 2012a, b; Wilcox & Keselman, 2003), procedures that have been found to be robust to non-normality [e.g., Erceg-Hurn, Wilcox, & Keselman (2013); Keselman, Algina, Lix, Wilcox, & Deering (2008a, b)].

References

Bradley, J. V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society, 16*, 333-336.

Cardoso de Oliveira, I. R., & Ferreira, D. F. (2010). Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation*, *80*, 513-526.

Doornik, J. A., & Hansen, H. (2008). Practioners' corner: An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70 supplement: 927-939.

Erceg-Hurn, D. M., Wilcox, R. R., & Keselman, H. J. (2013). Robust statistical estimation. In T. Little (Ed.), *The Oxford handbook of quantitative methods, Vol. 1*, 388-406. New York: Oxford University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986).*Robust statistics*. New York: Wiley.

Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2008). Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. *Applied Mathematical Sciences, 2(9)*, 449-462.

Hoaglin, D. C. (1983). G-and-h distributions. In Kotz, S., & Johnson, N. L. Eds.), *Encyclopedia of statistical sciences*, Vol. 3, pp. 298-301. New York: Wiley.

Hoaglin, D. C. (1985). Summarizing shape numerically; The g-and h-distributions. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.), *Exploring data, tables, trends, and shapes*, pp. 461-511. New York: Wiley

Hsu, T., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal*, *6*, 515-527.

Huber, P. J. & Ronchetti, E. (2009). *Robust statistics*, (2nd Ed.) New York: Wiley.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008a). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*, 110-129.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008b). Supplemental materials to 129a. A SAS program to implement a general approximate degrees of freedom solution for inference and estimation. http://dx.doi.org/10.1037/1082-989X.13.2.110.supp

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kirk, R. E. (2013). Experimental design: Procedures for the Behavioral Sciences (4th ed). Los Angeles: Sage.

Kowalchuk, R. K., & Headrick, T. C. (2010). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*, *63*, 63-74.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*, 263-269.

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). *Robust statistics*: *Theory and methods*. New York: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Muller, K. E., & Fetterman, B. A. (2002). Regression and ANOVA: An integrated approach using SAS software. Cary, NC: SAS Institute, Inc.

SAS Institute. (2012). Statistics: ANOVA and regression. Cary, NC: SAS Institute, Inc.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, *1*(2), 461-472.

Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: some statistical aspects of a common concept. *Clinical Dermatology*, *31*,757-761.

Staudte, R. G. & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Sürücü, B. (2006). Goodness-of-fit tests for multivariate distributions. *Communications in Statistics—Theory and Methods*, *35*, 1319-1331.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Wilcox, R. R. (2012a). *Introduction to robust estimation and hypothesis testing*, (3rd ed.) San Diego, CA: Academic Press.

Wilcox, R. R. (2012b). *Modern statistics for the social and behavioral sciences*: *A practical introduction*. New York: Chapman & Hall/CRC Press.

Wilcox, R. R. & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254-274.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1-17.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, *67*, 55-68.