Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 9

5-1-2007

Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics

Gibbs Y. Kanyongo
Duquesne University, kanyongog@duq.edu

Gordon P. Brook
Ohio University

Lydia Kyei-Blankson
Ohio University

Gulsah Gocmen
Ohio University

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 81-90

1538 – 9472/07/$95.00

Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics

Gibbs Y. Kanyongo
Duquesne University

Gordon P. Brook   Lydia Kyei-Blankson   Gulsah Gocmen
Ohio University

The relationship between reliability and statistical power is considered, and tables that account for reduced reliability are presented. A series of Monte Carlo experiments were conducted to determine the effect of changes in reliability on parametric and nonparametric statistical methods, including the paired samples dependent t test, pooled-variance independent t test, one-way analysis of variance with three levels, Wilcoxon signed-rank test for paired samples, and Mann-Whitney-Wilcoxon test for independent groups. Power tables were created that illustrate the reduction in statistical power from decreased reliability for given sample sizes. Sample size tables were created to provide the approximate sample sizes required to achieve given levels of statistical power based for several levels of reliability.

Key words: Pseudorandom generation, effect size, Monte Carlo simulations.

Introduction

Students of statistics usually become familiar with the factors that affect statistical power. For example, most students learn that sample size, level of significance, and estimated effect size all determine the a priori power of a statistical analysis. Some know that how effectively a particular design reduces error variance affects power, as does the directionality of the alternative hypothesis. However, many students do not realize that the reliability of measurements may also affect the statistical power (Hopkins & Hopkins, 1979). Light, Singer, and Willett (1990) provided tables to illustrate the point. Unfortunately, their tables provide only a very few situations and are therefore limited in their usefulness. It is not clear how the Light et al. tables were developed. The present study extends their tables and provides such information for additional statistical methods.

Using the information provided in these tables, researchers can account for different levels of reliability as they determine sample sizes for their studies. Perhaps the converse approach is even more useful; however, that is, researchers might be encouraged to improve the reliability of their instruments in order to need fewer participants in their studies. These tables can also be useful tools in teaching students the relationship between reliability of a survey instrument and statistical power.

Background

One of the chief concerns of research design is to ensure that a study has adequate statistical power to detect meaningful differences, if indeed they exist. There is a very good reason researchers should worry about power a priori: If researchers are going to invest a great amount of money and time in carrying out a study, then they would certainly want to have a reasonable chance, perhaps 70% or 80%, to find a difference between groups if it does exist. Thus, a priori power (the probability of rejecting a null hypothesis that is false) will inform researchers how many subjects per group

Gibbs Y. Kanyongo is an Assistant Professor of Education. He teaches educational statistics and research. Email: kanyongog@duq.edu. Gordon Brooks is an Associate Professor in the College of Education at Ohio University. He teaches educational statistics and research, and Monte Carlo simulations. He is also an author of several computer simulation programs. Lydia Kyei-Blankson and Gulsah Gocmen are graduate student in the College of Education.

will be needed for adequate power. Several factors affect statistical power. That is, once the statistical method and the alternative hypothesis have been set, the power of a statistical test is directly dependent on the sample size, level of significance, and effect size (Stevens, 2002). Often overlooked, however, is the relationship that variance has with power. Specifically, variance influences power through the effect size. For example, Cohen (1988) defined the effect for the t statistic as

$$\delta = (\mu_1 - \mu_0) / \sigma_X \qquad (1)$$

If variance can be reduced, effect size increases. Variance reduction techniques include using a more homogeneous population and improving the reliability of measurements (Aron & Aron, 1997; Zimmerman, Williams, & Zumbo, 1993). Similarly, because variance is reduced, analysis of covariance is more powerful than analysis of variance when a useful covariate is incorporated into the design.

Reliability and Effect Size

Cleary and Linn (1969) reported that "in the derivation and interpretation of statistical tests, the observations are generally considered to be free of error of measurement" (p. 50). From a classical test theory perspective, an individual's observed score (X) is the sum of true score (T) and error score (E); that is, $X = T + E$. Thus, if there is no error of measurement, then the observations are the true scores; implicitly, statistical hypotheses are proposed in terms of true scores. For a set of scores, however, measurements made without error occur only when the instruments provide perfectly reliable scores. Observed score variance, σ_X^2, is defined as the sum of true score variance, σ_T^2, and measurement error variance. Because reliability, $\rho_{xx'}$, is defined as the ratio of true score variance to observed score variance,

$$\rho_{XX'} = \sigma_T^2 / \sigma_X^2 = 1 - \sigma_E^2 / \sigma_X^2 \,, \qquad (2)$$

reliability can only be perfect (i.e.,) when there is no measurement error (Lord & Novick, 1968). Because σ_X can be written as

$$\sigma_T / \sqrt{\sigma_{XX'}}, \qquad (3)$$

the standardized effect size for the t test can be written as

$$\delta = (\mu_1 - \mu_0)(\sqrt{\sigma_{XX'}}\,)/\, \sigma_T \quad . (4)$$

(Levin & Subkoviak, 1977; Williams & Zimmerman, 1989). Consequently, reliability affects statistical power indirectly through effect sizes. Cohen (1988) reported that reduced reliability results in reduced effect sizes in observed data (ES), which therefore reduces power. That is, observed effect sizes,

$$ES = ESP * \sqrt{r_{XX'}}\,, \qquad (5)$$

where ESP is the population effect size. Therefore, when reliability is perfect, observed ES equals ESP; but when reliability is less than perfect, ES is a value smaller than the true ESP. Some introductory statistics textbooks discuss this problem in reference to attenuation in correlation due to unreliability of measures (e.g., Glass & Hopkins, 1996).

Reliability and Power

Controversy surrounds the relationship between power and reliability (Williams & Zimmerman, 1989). Good statistical power can exist with poor reliability and a change in variance unrelated to reliability can change power. However, there are persuasive reasons to consider reliability as an important factor in determining statistical power. For example, statistical power is a function of level of significance, sample size, and effect size only under the assumption of no measurement error, but measures in the social sciences are typically not measured perfectly (Cleary & Linn, 1969; Levin & Subkoviak, 1977). Indeed, the implicit assumption that our measures are perfectly reliable is not justified in practice and therefore measurement error should be considered a priori for sample size (Crocker & Algina, 1986; Subkoviak & Levin, 1977; Sutcliffe, 1958).

There is no controversy that statistical power depends on observed variance. Zimmerman and Williams (1986) noted that when speaking of statistical power it is irrelevant whether the observed variance is all true score variance or contains some amount of measurement error; that is, "the greater the observed variability of a dependent variable, whatever its source, the less is the power of a statistical test" (p. 123). However, because reliability is defined by observed variance in conjunction with either true or error variance, one cannot be certain which source of variance is changed when reliability improves. That is, if observed variance increases, one cannot be certain whether the increase is due to an increase in true score variance or a increase in error variance, or both. Or as Zimmerman et al. (1993) reported, power changes as reliability changes only if observed score variance changes simultaneously.

Knowing that improved reliability results in less measurement error, if it is assumed that true variance is a fixed value for the given population, it follows that a change in reliability will result in a change in observed score variance. Indeed, statistical power is a mathematical function of reliability only if either true score variance or error variance is a constant; otherwise power and reliability are simply related (Cohen, 1988; Williams & Zimmerman, 1989). But, improvement in reliability is usually interpreted as a reduction in the measurement error variance that occurs from a more precise measurement (Zimmerman & Williams, 1986). Therefore, a reduction in reliability that is accompanied by an increase in observed score variance will indeed reduce statistical power (Zimmerman et al., 1993). That is, if true score variance remains constant but lower reliability leads to increased error variance, then statistical power will be reduced because of the increased observed score variance (Humphreys, 1993).

Based on such an assumption, Light et al. (1990) advised that when measurements are less than perfectly reliable, improving the power of statistical tests involves a decision either to increase sample size or to increase reliability—the researcher must compare the costs associated with instrument improvement to the costs of adding study participants (see also Cleary & Linn, 1969; Feldt & Brennan, 1993). Researchers may encounter such a situation if an instrument does not perform as reliably in a given study as it has elsewhere, leading to increased variance in the current project. Assuming that the increased variance is not due to more heterogeneity in the population and that the true score variance of the population hasn't changed, the observed score variance will change as a consequence of the change in reliability.

Unfortunately, there are few easy ways to account for reliability when determining sample sizes. The tables found in Cohen (1988) do not provide the option to vary reliability. Computer programs such as Sample Power and PASS 2000 also assume perfect reliability. This article will report on the impact of reliability on power as well as provide tables to assist researchers in finding sample sizes necessary with fallible measures.

Methodology

Two Monte Carlo programs, MC2G (Brooks, 2002) and MC3G (Brooks, 2002) written in Borland Delphi Professional version 6.0, were used to create normally distributed but unreliable data and perform analyses for several statistical methods, namely: (a) paired samples dependent t test, (b) pooled-variance independent t test, (c) one-way analysis of variance with three levels, (d) Spearman rank correlation, (e) Wilcoxon signed-rank test for paired samples, and (f) Mann-Whitney-Wilcoxon test for independent groups. The program output was used to create power and sample size tables for these tests.

Reliability was varied from .70 to 1.0 in increments of 0.05. For power tables, power rates varied from .70 to .90 by .10. Population effect sizes were varied from small to large using Cohen's (1988) conventional standards. Specifically, for t tests and their nonparametric alternatives, a small standardized effect size was set at d = .20, medium was d = .50, and a large effect was set to be d = .80; for correlations, a small effect was set at r = .10, medium was r = .30, and a large effect was set to be r = .50; for ANOVA, a small standardized difference effect was set at f = .10, medium was f = .25, and a large effect was set to be f = .40. For the power tables, the sample sizes were obtained under the assumption of perfect reliability. That is, the sample sizes were fixed at the values needed to achieve power levels of .70, .80 and .90, respectively, when reliability was 1.0. The remaining values in the power tables were determined by systematically varying the reliability with that given sample size. For the sample sizes tables, power was fixed, reliability was varied, and sample sizes were tried repeatedly until the desired power was achieved.

Data Generation

The two Monte Carlo programs generate uniformly distributed pseudorandom numbers that are used as input to the procedure that converts them into normally distributed data. All data were generated to follow the standard normal distribution. For each sample, the appropriate statistical analysis was performed. The number of correct rejections of the null hypothesis was stored and reported by the program. These procedures were repeated as necessary for each sample condition created. The programs use the L'Ecuyer (1988) uniform random number generator. Specifically, the Fortran code of Press, Teukolsky, Vetterling, and Flannery (1992), was translated into Delphi Pascal. The L'Ecuyer generator was chosen because of its large period and because combined generators are recommended for use with the Box-Muller method for generating random normal deviates (Park & Miller, 1988).

The computer algorithm for the Box-Muller method used by the MC2G and MC3G programs was adapted for Delphi Pascal from the standard Pascal code provided by Press, Flannery, Teukolsky, and Vetterling (1989). The programs generate normally distributed data of varying reliability based on classical test theory. That is, reliability is not defined using a particular measure of reliability (e.g., splithalf or internal consistency); rather it is defined as the proportion of raw score variance explained by true score variance (Equation 2). Each raw score generated is taken to be a standardized total score.

In order to generate data with less-than-perfect reliability, scores were generated using the true-score standard deviations provided by the researchers; then for each score, the programs added a random error component. Consequently, as reliability decreased, the variation of the random error component increased, resulting in increased raw score variance. For correlation analyses, the same reliability was used for both measures; for independent sample analyses, the same reliability was used each for each group.

Monte Carlo

The number of iterations for the study is based on the procedures provided by Robey and Barcikowski (1992). Significance levels for both tests on which Robey and Barcikowski's method is based were set at .05 with a power level of .90; the magnitude of departure was chosen to be $\alpha \pm .2$, which falls between their intermediate and stringent criteria for accuracy. The magnitude of departure is justified by the fact that at $\pm .2\ \alpha$, the accuracy range for $\alpha = .05$ is $.04 \leq \alpha \leq .06$.

Based on the calculations for these parameters (this set of values was not tabled), 5422 iterations would be required to “confidently detect departures from robustness in Monte Carlo results” (Robey & Barcikowski, 1992, p. 283), but applies to power studies also (Brooks, Barcikowski, & Robey, 1999). However, to assure even greater stability in the results, a larger number of simulations was chosen for each type of analysis. Specifically, 10,000 samples were used for the power tables. The sample size algorithm used by the programs runs repeated analyses beginning with 100 samples per analysis, gradually increasing to 10,000 samples per analysis. Sometimes, however, the algorithm aborts before the 10,000 sample level is reached when the desired power level is approximated closely enough earlier in the process (at least 1000 samples were run in every case).

Results

Tables 1 through 5 show the relationship between statistical power and reliability for the dependent t test, independent t test, one-way ANOVA with three groups, Wilcoxon signed ranks, and Mann-Whitney-Wilcoxon tests, respectively. The tables clearly show that, as reliability is reduced while true score variance remains constant, statistical power is reduced. There is a relatively linear relationship between statistical power and reliability when sample size is fixed. For example, Table 1 shows that when statistical power is chosen to be .71 for the dependent t test, 12 cases are required when perfect reliability is assumed and a large effect size (d = .8) is expected. When reliability was changed to .90 with 12 cases, the actual statistical power was observed to be .63. Reliability set at .80 resulted in observed statistical power of .54. Finally, actual power for 12 cases was .46 when reliability was set at .70. Such depreciation of power occurs for all other tests examined in the study.

Table 1. Actual statistical power for paired-samples dependent t tests resulting from different reliability values for given sample sizes at two-tailed α = .05

	N	Reliability						
Effect Size	per group	1.0	.95	.90	.85	.80	.75	.70
Large	12	.71	.67	.63	.59	.54	.50	.46
(d=.8)	15	.82	.78	.74	.70	.65	.61	.56
	19	.91	.88	.85	.81	.77	.73	.68
Medium	27	.71	.66	.62	.58	.53	.49	.45
(d=.5)	34	.81	.77	.73	.68	.64	.59	.55
	44	.90	.87	.83	.80	.75	.71	.66
Small	157	.70	.66	.62	.57	.53	.49	.45
(d=.2)	199	.80	.76	.72	.67	.63	.59	.54
	264	.90	.87	.83	.80	.75	.71	.66

Table 2.Actual statistical power for pooled-variance independent t tests resulting from different reliability values for given sample sizes at two-tailed α = .05

	N	Reliability						
Effect Size	per group	1.0	.95	.90	.85	.80	.75	.70
Large	21	.72	.70	.67	.65	.62	.59	.56
(d=.8)	26	.81	.79	.77	.74	.72	.69	.66
	34	.90	.89	.87	.85	.83	.81	.78
Medium	51	.70	.68	.66	.64	.61	.58	.55
(d=.5)	64	.80	.78	.76	.74	.71	.68	.65
	86	.90	.89	.87	.85	.83	.81	.78
Small	309	.70	.68	.65	.63	.60	.58	.54
(d=.2)	393	.80	.78	.76	.73	.71	.68	.65
	526	.90	.89	.87	.85	.83	.80	.77

Table 3. Actual statistical power for one-way analysis of variance with three groups resulting from different reliability values for given sample sizes at two-tailed α = .05

	N	Reliability						
Effect Size	per group	1.0	.95	.90	.85	.80	.75	.70
Large	17	.70	.67	.65	.63	.60	.56	.53
(f=.40)	21	.80	.78	.75	.73	.71	.67	.64
	28	.91	.89	.87	.85	.83	.80	.77
Medium	41	.70	.67	.65	.62	.60	.57	.54
(f=.25)	51	.80	.78	.75	.73	.70	.67	.64
	66	.90	.88	.86	.84	.82	.79	.76
Small	269	.71	.68	.65	.62	.60	.57	.54
(f=.10)	333	.80	.78	.75	.73	.70	.67	.64
	441	.90	.89	.87	.85	.82	.80	.77

Table 4. Actual statistical power for Wilcoxon signed-rank tests resulting from different reliability values for given sample sizes at two-tailed α = .05

	N	Reliability						
Effect Size	per group	1.0	.95	.90	.85	.80	.75	.70
Large	12	.70	.66	.62	.58	.54	.50	.45
(d=.8)	15	.80	.76	.71	.67	.63	.58	.54
	19	.90	.87	.83	.79	.75	.71	.66
Medium	28	.70	.65	.61	.57	.53	.48	.44
(d=.5)	35	.80	.76	.72	.68	.63	.58	.54
	46	.90	.87	.84	.80	.75	.71	.66
Small	164	.70	.66	.62	.57	.53	.49	.45
(d=.2)	208	.80	.76	.72	.68	.63	.59	.54
	276	.90	.87	.83	.80	.75	.71	.66

Table 5. Actual statistical power for Mann-Whitney-Wilcoxon tests resulting from different reliability values for given sample sizes at two-tailed α = .05

	N	Reliability						
Effect Size	per group	1.0	.95	.90	.85	.80	.75	.70
Large	21	.69	.67	.64	.62	.59	.57	.53
(d=.8)	27	.80	.78	.76	.74	.71	.68	.65
	35	.90	.88	.86	.84	.82	.80	.77
Medium	53	.70	.68	.66	.63	.61	.60	.55
(d=.5)	67	.80	.78	.76	.73	.71	.68	.65
	90	.90	.89	.87	.85	.83	.81	.78
Small	323	.70	.67	.65	.62	.59	.57	.54
(d=.2)	411	.80	.78	.76	.73	.71	.68	.65
	550	.89	.88	.87	.85	.82	.80	.77

Tables 6 through 10 show the sample sizes required to maintain a given power level when reliability is less than perfect. Again, there are relatively linear relationships for all tests at all power levels. For example, Table 6 shows that when the desired statistical power level is set at .80 and a large effect size (d = .8) is expected, the use of 15 cases results in power of .80 when reliability is 1.0; but when reliability is reduced to .90, 17 cases are required. If reliability is .80, then the study needs 21 participants. Finally, 25 cases must be used to achieve power of .80 when reliability is .70.

Table 6. Sample sizes required for paired-samples dependent t tests in order to achieve the given statistical power values under different reliability conditions at two-tailed $\alpha = .05$

		Reliability						
Effect Size	Power	1.0	.95	.90	.85	.80	.75	.70
Large	.70	12	13	14	15	17	18	20
(d=.8)	.80	15	16	17	19	21	23	25
	.90	19	20	22	24	27	29	33
Medium	.70	27	29	32	35	39	43	48
(d=.5)	.80	34	37	40	44	49	54	60
	.90	44	49	53	59	65	72	80
Small	.70	157	172	192	214	234	258	287
(d=.2)	.80	199	220	243	266	289	329	369
	.90	264	286	328	364	400	440	492

Table 7. Sample sizes required for pooled-variance independent t tests in order to achieve the given statistical power values under different reliability conditions at two-tailed $\alpha = .05$

		Reliability						
Effect Size	Power	1.0	.95	.90	.85	.80	.75	.70
Large	.70	21	22	23	24	25	27	29
(d=.8)	.80	26	27	28	30	32	34	37
	.90	34	36	38	40	42	45	48
Medium	.70	51	53	56	59	63	67	72
(d=.5)	.80	64	67	71	75	79	85	91
	.90	86	89	95	102	107	114	121
Small	.70	309	327	345	365	387	415	443
(d=.2)	.80	393	415	438	466	492	527	566
	.90	526	557	583	618	658	702	755

Table 8. Sample sizes required for one-way analysis of variance with three groups in order to achieve the given statistical power values under different reliability conditions at two-tailed α = .05

		Reliability						
Effect Size	Power	1.0	.95	.90	.85	.80	.75	.70
Large	.70	17	18	19	20	21	22	24
(f=.40)	.80	21	22	23	25	26	28	30
	.90	28	29	30	32	34	36	39
Medium	.70	41	44	45	48	50	54	58
(f=.25)	.80	51	54	56	61	65	68	73
	.90	66	70	75	78	83	88	95
Small	.70	269	288	300	314	332	356	382
(f=.10)	.80	333	353	374	395	419	451	482
	.90	441	464	488	516	551	583	619

Table 9. Sample sizes required for Wilcoxon signed-ranks tests in order to achieve the given statistical power values under different reliability conditions at two-tailed α = .05

		Reliability						
Effect Size	Power	1.0	.95	.90	.85	.80	.75	.70
Large	.70	12	13	14	16	17	19	20
(d=.8)	.80	15	17	18	20	21	24	26
	.90	19	21	23	25	28	31	34
Medium	.70	28	31	34	37	40	45	50
(d=.5)	.80	35	39	42	46	51	57	63
	.90	46	51	56	62	68	75	85
Small	.70	164	181	201	222	246	273	304
(d=.2)	.80	208	225	253	282	314	346	387
	.90	276	307	338	376	417	462	511

Table 10. Sample sizes required for Mann-Whitney-Wilcoxon tests in order to achieve the given statistical power values under different reliability conditions at two-tailed α = .05

		Reliability						
Effect Size	Power	1.0	.95	.90	.85	.80	.75	.70
Large	.70	21	22	23	25	27	28	30
(f=.40)	.80	27	28	30	32	34	36	39
	.90	35	37	40	42	44	47	51
Medium	.70	53	56	58	62	67	69	75
(f=.25)	.80	67	70	74	79	84	89	96
	.90	90	93	97	105	113	117	127
Small	.70	323	339	358	386	405	437	463
(f=.10)	.80	411	430	458	484	517	551	593
	.90	550	575	611	653	692	733	796

Conclusion

In social sciences, few things are measured perfectly (Subkoviak & Levin, 1977). Researchers should therefore make an effort to minimize the effects of measurement error. Although some authors suggest that lower reliability is acceptable for group studies of attitudes or personality variables (e.g., Fink & Kosecoff, 1998; McMillan & Schumacher, 2001), it becomes obvious based on the tables provided here that improving reliability will increase power and therefore fewer members of these groups will be needed to participate in the study.

For example, in a two group study using a dependent measure that produces scores with a reliability of .70, 91 participants are required for a medium effect size at a power of .80; if reliability is improved to .85, the number of participants can be reduced to 75 (see Table 7). Perhaps for some studies, the additional effort required to improve the instrument is not justifiable; but for research with high per-subject costs, investment to improve the instrument may be very worthwhile. As well, the effect of measurement fallibility is even more dramatic for small effect sizes. In the same example as above, but for a small effect size, an improvement from reliability of .70 to .85 will result in a sample size reduction of around 100 (see Table 7).

Perhaps the most advantageous way for researchers to use the sample size information provided here is to make informed decisions about the trade-off between sample size and reliability. That is, researchers can make informed decisions about the costs and benefits of spending time and effort to improve an instrument. The issue really isn't how many more people do we need because our instrument is not perfectly reliable? Researchers would already have an estimate of variance based on that level of unreliability from pilot studies or previous research—after all, the effect size would be based on that observed variance—not true score variance. Rather, the implication intended from this work is more emphasis on the development of reliable and valid instruments. As instruments and reliability improve, because the true score variance of participants would (presumably) remain the same, observed score variance will decrease and would provide additional power. There are several strategies that have been developed for minimizing the effects of measurement error and increasing reliability. These include revising items, increasing the number of items, lengthening item scales, administering the instrument systematically, timing of data collection and use of multiple raters or scores (Light et al., 1990).

Before choosing a final sample size, the possibility of measurement error should be considered. To determine sample sizes "without simultaneously considering errors of measurement is to live in a 'fool's paradise'" (Levin & Subkoviak, 1977, p. 337). If one suspects that measurement error exists and there is no viable means to reduce it, sample size should be increased accordingly. Researchers can identify potential problems with measurement error through pilot studies or previous research. Where reliability information is lacking, the researcher should use cautious estimates, with a preference toward more conservative values, when deciding sample sizes (Levin & Subkoviak, 1977).

References

Aron, A., & Aron, E. N. (1997). *Statistics for the behavioral and social sciences: A brief course*. Upper Saddle River, NJ: Prentice Hall.

Brooks, G. P. (2002). *MC2G: Monte Carlo Analyses for 1 or 2 Groups* (Version 3.0.7) [Computer software]. Retrieved from http://oak.cats.ohiou.edu/~brooksg/mc2g.htm

Brooks, G. P. (2002). *MC3G: Monte Carlo Analyses for 3 Groups.* (Version 1.1.1) [Computer software]. Retrieved from http://oak.cats.ohiou.edu/~brooksg/mc3g.htm

Brooks, G. P., Barcikowski, R. S., & Robey, R. R. (1999, April). *Monte Carlo simulation for perusal and practice.* Paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology, 22*, 49-55.

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

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart, & Winston.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.). *Educational measurement* (pp. 105-146). Phoenix, AZ: Oryx.

Fink, A., & Kosecoff, J. (1998). *How to conduct surveys: A step-by-step guide* (2nd. ed.). Thousand Oaks, CA: Sage.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Hopkins, K. D., & Hopkins, B. R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education, 13*, 463-466.

Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement, 17*, 11-14.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM, 31*, 742-749, 774.

Levin, J. R., & Subkoviak, M. J. (1977). Planning an experiment in the company of measurement error. *Applied Psychological Measurement, 1*, 331-338.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McMillan, J. H., & Schumacher, S. (2001). *Research in education: A conceptual introduction* (5th ed.). New York: Longman.

Nicewander, W. A. & Price J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin, 94*, 524-533.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. *Communications of the ACM, 31*, 1192-1201.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal: The art of scientific computing*. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). New York: Cambridge University.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a statistical test. *Journal of Educational Measurement, 14*, 47-52.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 23*, 9-17.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology, 116*, 359-369.

Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the poser of significance tests. *Psychological Bulletin, 100*, 123-124.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17*, 1-9.