Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 21, Number 9, August 2016 ISSN 1531-7714

Measurement Error and Equating Error in Power Analysis

Gary W. Phillips & Tao Jiang
American Institutes for Research

Power analysis is a fundamental prerequisite for conducting scientific research. Without power analysis the researcher has no way of knowing whether the sample size is large enough to detect the effect he or she is looking for. This paper demonstrates how psychometric factors such as measurement error and equating error affect the power of statistical tests. The overall finding is that measurement error and equating error reduce power and inflate sample size requirements. It is recommended that researchers, where appropriate, incorporate these sources of error in conducting power analysis. If either of these two sources of error are present in the data but not accounted for in the power analysis, then power will be underestimated and sample size requirements will be underestimated.

Power is the probability of detecting whether the population effect you are looking for is in your sample. The mathematical statistics of power analysis have been well documented in the literature (Cohen, 1969, 1988; Mood, Graybill, & Boes, 1974). There are two types of power analysis: a priori power analysis and post hoc power analysis. A priori power analysis is used to determine the sample size needed to achieve a specified power in a planned research study for a given expected effect size. Post hoc power analysis is a retrospective power analysis after the research has been completed. This paper is really about how researchers need to conduct a more thorough a priori power analysis by taking into account psychometric factors such as measurement error and equating error.

The process of power analysis consists of deciding how big an effect you want to be able to detect and at what probability you want to detect it. One of the important uses of power analysis is to determine the sample size needed in a planned study. We want the sample size to be large enough to detect an effect that is substantively significant. Another way of saying this is that you want the sample size to be large enough to have the power you want to detect the effect you are looking for. Three decisions must be made before determining the sample size.

First, the probability of rejecting a true null hypothesis H_0 needs to be decided. This probability is denoted as α and is the probability of a Type I error. For illustrative purposes in this paper we use a two-tailed t-test with $\alpha / 2 = .025$.

Second, the probability of rejecting a false null hypothesis $1-\beta$ needs to be decided (i.e., statistical power). Power is defined as 1 minus β where β is the probability of a Type II error. A Type II error is failing to reject a false null hypothesis. The usual convention is that power should be at least equal to .80. This convention is adopted in this paper.

Third, the size of the effect you are trying to detect needs to be decided. As Cohen states, "neither the determination of power or necessary sample size can proceed without the investigator having some idea about the degree to which the H_0 is believed to be false" (Cohen, 1992, p. 156). For illustrative purposes we will

use Cohen's standardized effect size $\delta = \frac{\mu_2 - \mu_1}{\sigma}$ (Cohen, 1969, p. 19), which is defined as the population mean difference between group 2 and group 1, divided by the pooled population standard deviation. The population effect size δ is estimated in the sample by

$$d = \frac{\bar{y}_2 - \bar{y}_1}{\hat{\sigma}} = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}}$$

Cohen has recommended a convention to aid researchers in conducting power analysis. His convention is that researchers should consider small, medium, or large effect sizes, which he operationally defines as $\delta_S = .20, \delta_M = .50,$ and $\delta_L = .80$ (Cohen, 1969, p. 12). Cohen provided examples of these effect sizes (Cohen, 1988, p. 25–27, 79–80). For example, a small effect size is one that is meaningful but not visible to the naked eye such as the mean height between 15- and 16-year-old girls. A medium effect size is visible to the naked eye in the data such as the difference in mean height between 14 and 18 year-old girls. An example of a large effect size is the mean difference between 13- and 18-year-old girls. We will use Cohen's conventions to illustrate the effect of sample design effects, measurement error, and equating error on power analysis.

In large-scale scientific research, power analysis is far more complicated than indicated in Cohen's books. In many large-scale research studies the samples are complex samples and are not simple random samples (SRS) as assumed by Cohen. Essentially, any factor that perturbs the sampling distribution of the test statistic $\bar{y}_2 - \bar{y}_1$ will influence power. One factor that affects the sampling distribution is complex sampling (CS). The term *complex sampling* refers to sampling that may involve stratification, unequal probability of selection, and clustering. In general, stratification provides a small reduction in variance of the sampling distribution and sampling weights related to the unequal probability of selection causes a small increase. The biggest inflationary impact on the sampling distribution is usually the clustering. In complex sampling, data are often collected by sampling clusters first, then by sampling units within clusters. For example, in a randomized clinical trial of reading programs, schools may be randomly assigned to a treatment or control group. Then, within each school, a random sample of students will be selected for the study. Such research designs are called *cluster-randomized* designs. Another example would involve large-scale assessment field tests such as state testing programs. New items may be field-tested in samples where schools are first sampled, then students within schools are tested. As is shown below, complex sampling often increases the variance of the sampling distribution of the test statistic, which reduces statistical power and concomitantly increases the sample sizes needed to detect the target effect size.

A second factor that affects power analysis is measurement error. For example, measurements of blood pressure and glucose readings, survey instruments, behavioral observations, rating scales, and test scores all contain measurement error. Measurement error also reduces statistical power and increases sample size needed to detect the target effect size.

Third, many outcome variables contain equating error, which reduces statistical power. This is especially true in psychological and educational tests. Equating error occurs when the scale used in the analysis (such as the posttest) has been equated to previous versions of the test (such as the pretest). As is shown in the next section, equating error also reduces statistical power and increases the sample size needed to detect the target effect size.

Simple Random Samples

Let's assume that our sample is a simple random sample as assumed by Cohen (1969). Let $T_v(t, \gamma)$ be the Cumulative Distribution Function (CDF) of a non-central *t*-distribution at *t* ($t \in (-\infty, +\infty)$) with degrees of freedom *v* and non-centrality γ (Chow, Shao, & Wang, 2002). In this paper we expand the traditional non-central t-distribution to accommodate design effects, measurement error, and equating error. We use the example of two independent samples. The formula will work for both equal and unequal variances but for simplicity we assume equal variances. The probability β associated with the difference $\mu_2 - \mu_1$ being different from 0, based on two true scores with distributions from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, for given significance level α is shown in equation (1).

In equation (1), $\Delta = \mu_2 - \mu_1$ is the true mean difference we wish to detect, α is the specified probability of a Type I error, β is the probability of a Type II error, σ_1^2 and σ_2^2 are the variances, n_1 is the

$$\beta = T_v\left(t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\left(\sigma_1^2/n_1 + \sigma_2^2/pn_1\right)}}\right) - T_{(1+p)n-2}\left(-t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\left(\sigma_1^2/n_1 + \sigma_2^2/pn_1\right)}}\right) \quad (1)$$

sample size for sample 1, and pn_1 is the sample size for sample 2, where p is the ratio of the sample size in group 2 to group 1. When $\sigma_1^2 \neq \sigma_2^2$, the degrees of freedom associated with equation (1) is provided by Satterthwaite (1946) as

$$v = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{pn_1}\right)^2}{\frac{\left(\sigma_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(\sigma_2^2/pn_1\right)^2}{pn_1 - 1}} \quad (2)$$

and when $\sigma_1^2 = \sigma_2^2$ the degrees of freedom associated with equation (1) are

$$v = n_1 + pn_1 - 2 \quad (3)$$

The above non-central t-distribution can be used to replicate the sample size tables used by Cohen (1969) for various levels of $1-\beta$ and δ by specifying α. Note that $\delta = \Delta$ when $\sigma_1 = \sigma_2 = 1$. An example of a replication is provided in Table 1 where the cells are the values of n for different values of $1-\beta$ and δ given $\sigma_1^2 = \sigma_2^2 = 1.0$ and $p = 1.0$. If the reader compares Table 1 to Cohen's table the entries will either agree exactly or be off by no more than one.

From Table 1 we can see the sample size requirements in a simple random sample to detect Cohen's small, medium, and large effect sizes are 394, 64, and 26, respectively. In the rest of this paper we use Table 1 as a baseline and expand on equations (1)–(3) to explore the impact of design effects, measurement error, and equating error on power analysis and the estimation of required sample sizes. We note that a power equal to .25 in Cohen's Table 1 is not a value of power that researchers should aspire to use in practice. It represents a level of power associated with an inadequate sample size. Most researchers use power equal to .80 as a convention.

Complex Random Samples

Data in the social and behavioral sciences are often collected through complex sampling designs that

Table 1. Replicating Cohen's Sample Size Tables (Cohen, 1969, Table 2.4.1, Pages 52–53)

	δ										
1 – β	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00	1.20	1.40
0.25	331	84	38	22	15	11	8	7	5	4	3
0.50	770	194	87	49	32	23	17	14	9	7	6
0.60	981	246	110	63	41	29	21	17	11	8	7
0.67	1,143	287	128	73	47	33	25	19	13	9	7
0.70	1,236	310	139	79	51	36	27	21	14	10	8
0.75	1,390	348	156	88	57	40	30	23	15	11	9
0.80	1,571	394	176	100	64	45	34	26	17	12	10
0.85	1,797	450	201	114	73	51	38	30	19	14	11
0.90	2,103	527	235	133	86	60	44	34	23	16	12
0.95	2,600	651	290	164	105	74	55	42	27	20	15
0.99	3,676	920	410	231	148	104	76	59	38	27	20

include stratification, weighting, and clustering. With stratification the population of units is grouped into homogenous subgroups (strata), and then units are sampled within the subgroups. Often some strata are oversampled and sample stratum weights are used to approximate a representative sample. Stratification generally provides a moderate reduction in the design effect and weighting causes a moderate increase. However, the effects on the error variance due to stratification and weighting are often overshadowed by a larger inflation in the error variance due to cluster sampling. In many sample designs, clusters of units are sampled first, then individual units are sampled within the selected clusters. An example of clustering is when experimental studies and clinical trials use group randomized designs where treatment and control conditions are randomly assigned to clusters of units rather than the units themselves. For example, in pharmaceutical and health research the clusters may be hospitals, HIV prevention programs, or family planning centers. In educational research the clusters may be schools; then students in schools are sampled. Cluster sampling can have many hierarchical levels such as sampling school districts, then schools, then classrooms, then students.

A design effect (Deff) is the inflation in the error variance of your test statistic caused by the design of the sample. The design effect is the ratio of the error variance of a statistic with a complex sample to the error variance of the statistic with a simple random sample of the same size. The concept of the design effect originated with Cornfield (1951) when he compared the error variance of a simple random sample to a complex sample of the same sample size as a way to characterize the efficiency of sample design. Kish (1965) used the inverse of Cornfield's ratio and named it the "design effect." There are many factors that influence the design effect, including stratification, weighting and cluster sampling. Often stratification reduces the design effect and weighting increases the design effect. By far the largest increase in the design effect is caused by cluster sampling. Complex sampling and design effects are described in more detail by Cochran (1977), Levy (1999), and Lohr (1999).

Complex sampling influences power analysis through the sample's design effect. Power analysis and sample size estimation for clustered samples is well established in the literature (Raudenbush, 1997; Konstantopoulos, 2009; Hedges & Rhoads, 2009). If there is only one level of cluster sampling—such as when schools are sampled first, then students are sampled within schools—then the design effect is defined as $Deff = 1 + (\bar{n}_S - 1)\rho_S$.

In this formula $\bar{n}_S$ is the average cluster size and ρ_S is the intra-class correlation between students within clusters. The intra-class correlation is defined as $\rho_S = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$, where σ_B^2 is the variance between schools and σ_W^2 is the variance within schools. In this design effect formula, we see that the two factors that influence the magnitude of the design effect are $\bar{n}_S$ and ρ_S. Larger cluster sizes and larger intra-class correlations will result in larger design effects. If $\rho_S = 0.0$, which would occur in a simple random sample, then $Deff = 1.0$ regardless of the size of $\bar{n}_S$. If $\bar{n}_S = 1.0$, which would occur if only one observation is sampled per cluster, then $Deff = 1.0$ regardless of the size of ρ_S.

Design effects can be complicated depending on the number of levels of clustering. For example, for two levels of cluster sampling, where schools are sampled first, then classrooms within schools, for students within classrooms the design effect is $Deff = 1 + (\bar{n}_C - 1)\rho_C + \bar{n}_C(\bar{c}_S - 1)\rho_S$. In this two-level formula, $\bar{n}_C$ is the average number of students per classroom, $\bar{c}_S$ is the average number of classrooms per school, ρ_C is the intra-class correlation within classes, and ρ_S is the intra-class correlation within schools.

For three levels of cluster sampling such as sampling school districts first, then schools within districts, then classrooms within schools, then students within classrooms, the design effect is $Deff = 1 + (\bar{n}_C - 1)\rho_C + \bar{n}_C(\bar{c}_S - 1)\rho_S + \bar{n}_C\bar{c}_S(\bar{s}_D - 1)\rho_D$. In this three-level formula, $\bar{n}_C$ is the average number of students tested per class, $\bar{c}_S$ is the average number of classes sampled per school, and $\bar{s}_D$ is the average number of schools selected per district. The intra-class correlations are ρ_C, ρ_S, and ρ_D for classrooms, schools, and school districts, respectively.

In this paper we use the design effect to illustrate, primarily, the effects of clustering but in general the

design effect will also include the impact of stratification and sample weighting due to unequal probability of selection. Regardless of the complexity of the clustering, stratification and weighting in the sample design, the design effect can be used to conduct power analysis and sample size determination.

We can expand equations (1)–(3) to include the design effects from clustered sampling:

$$\beta = T_v\left(t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\left(\sigma_1^2 Deff_1 / n_1 + \sigma_2^2 Deff_2 / pn_1\right)}}\right) - T_{(1+p)n-2}\left(-t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\left(\sigma_1^2 Deff_1 / n_1 + \sigma_2^2 Deff_2 / pn_1\right)}}\right), \quad (4)$$

with degrees of freedom under $\sigma_1^2 \neq \sigma_2^2$

$$v = \frac{\left(\frac{\sigma_1^2 Deff_1}{n_1} + \frac{\sigma_2^2 Deff_2}{pn_1}\right)^2}{\frac{\left(\sigma_1^2 Deff_1 / n_1\right)^2}{n_1 / Deff_1 - 1} + \frac{\left(\sigma_2^2 Deff_2 / pn_1\right)^2}{pn_1 / Deff_2 - 1}} \quad (5)$$

and degrees of freedom when $\sigma_1^2 = \sigma_2^2$,

$$v = \frac{n_1}{Deff_1} + \frac{pn_1}{Deff_2} - 2 \quad (6)$$

Equations (4)–(6) can be used to estimate sample size requirements for complex samples. For our illustration we assume there is one level of school clustering, the test statistic is the true mean difference $\mu_2 - \mu_1$ the average cluster sizes in both groups are equal to 21 students per school, and the intra-class correlations in both groups are equal to .15. Therefore, the design effects are $Deff_1 = Deff_2 = 1 + .15(21 - 1) = 4.0$. A design effect of 4.0 means the error variance of the test statistic is four times larger in the clustered sample than it would be with an SRS of the same size. Alternatively, the root design effect, $\sqrt{Deff} = 2.0$ indicates that the standard error of the test statistic is twice as large in the clustered sample as in a simple random sample of the same size. The sample sizes required for a design effect of 4.0 are presented in Table 2.

One important observation in Table 2 is that the sample size requirements to detect $\delta_S = .20$, $\delta_M = .50$, and $\delta_L = .80$ for clustered samples, in this specific example, are approximately four times larger than the sample size requirements for a simple random sample. This is equal to the average of the design effects in group 1 and group 2. In general, the increase in the sample size caused by complex sampling can be approximated by

$$n_{CS} \approx n_{SRS}\left(\frac{\sigma_1^2 Deff_1 + \sigma_2^2 Deff_2 / p}{\sigma_1^2 + \sigma_2^2 / p}\right) \quad (7)$$

Table 2. Sample Size Table for Clustered Samples with $Deff_1 = Deff_2 = 4.0$

	δ										
1 – β	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00	1.20	1.40
0.25	1,322	334	151	87	57	41	32	25	18	14	12
0.50	3,077	773	346	196	127	90	67	53	35	26	21
0.60	3,923	984	440	249	161	113	84	66	44	32	25
0.67	4,570	1,146	512	290	187	131	98	76	50	36	28
0.70	4,942	1,239	553	313	202	142	105	82	54	39	30
0.75	5,557	1,392	621	351	226	159	118	91	60	43	33
0.80	6,283	1,574	702	397	256	179	133	103	67	48	37
0.85	7,187	1,800	802	453	292	204	151	117	76	55	41
0.90	8,410	2,106	938	530	341	238	176	136	89	63	48
0.95	10,400	2,603	1,159	654	420	293	217	167	108	77	58
0.99	14,702	3,679	1,637	923	592	413	304	234	151	107	80

Note: Given $\alpha = .05$, two-tailed t-test, $\sigma_1^2 = \sigma_2^2 = 1.0$, and $p = 1.0$

For example, the sample sizes required to detect $\delta_S = .20$, $\delta_M = .50$, and $\delta_L = .80$ with $1-\beta = .80$ for the clustered sample are now 1,574, 256, and 103, respectively. The increases in the sample size requirements from those required in a simple random sample in Table 1 are 1,574/394 = 3.99 for $\delta_S = .20$, 256/64 = 4.00 for $\delta_M = .50$, and 103/26 = 3.96 for $\delta_L = .80$. The sample size in each case is increased by a factor approximately equal to

$$\frac{\sigma_1^2 Deff_1 + \sigma_2^2 Deff_2 / p}{\sigma_1^2 + \sigma_2^2 / p} = \frac{4+4}{2} = 4.0$$

The sample sizes in Table 2 illustrate that the design of the sample can dramatically impact the sample sizes needed in the research study. This is why the sample designs should be as efficient as possible. For example, if the researcher has the choice of collecting data on 1,000 students from 10 schools versus 1,000 students from 20 schools, the second design is more efficient. The second design spreads the sample over more clusters, which reduces the cluster size and the design effect.

Measurement Error

In Section 3 we explored the impact of design effects when the outcome variable is measured without error. From a classical test theory true-score model point of view we treat the outcome measures as true scores. In this section we incorporate the reliability of the outcome measure into the power analysis.

Measurement error occurs in all scientific disciplines. Possibly the first formal model for measurement error was developed by Gauss in 1809 when he showed that the variation in astronomical measurements made by Galileo in the 17th century were due to imperfections in Galileo's telescopes. He showed that the distribution of such random errors of measurement followed a normal distribution. The classical test theory model used in this paper uses the same normal distribution that Gauss used to explain Galileo's errors in observations.

In this paper we use the classical test theory reliability coefficient R_1 for the outcome measure in group 1 and R_2 for the outcome measure in group 2 as our index of measurement error (Crocker & Algina, 1986; Lord, 1980; Lord & Novick, 1968). The reliability coefficient is the proportion of true score variance divided by the observed score variance in the outcome variable and can vary from 0.0 to 1.0. When the reliability coefficient equals 1.0, there is no measurement error in the outcome variable and we are using true measurements or true scores. If the reliability coefficient equals 0.0, then there is no measurement in the outcome variable and we are using random numbers. In practice, all outcome variables have some degree of unreliability. We show below that measurement error reduces power (Cleary & Linn, 1969; Williams & Zimmerman, 1989) and increases sample size requirements beyond the inflation caused by cluster sampling. We can further expand equations (1)–(3) to include reliability, *R*, as well as design effects:

$$\beta = T_v\left(t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\frac{(\sigma_1^2/R_1)(R_1 Deff_1 + 1 - R_1)}{n_1} + \frac{(\sigma_2^2/R_2)(R_2 Deff_2 + 1 - R_2)}{pn_1}}}\right) - T_{(1+p)n-2}\left(-t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\frac{(\sigma_1^2/R_1)(R_1 Deff_1 + 1 - R_1)}{n_1} + \frac{(\sigma_2^2/R_2)(R_2 Deff_2 + 1 - R_2)}{pn_1}}}\right), \quad (8)$$

with degrees of freedom when $\frac{\sigma_1^2}{R_1} \neq \frac{\sigma_2^2}{R_2}$,

$$v = \frac{\left(\frac{(\sigma_1^2/R_1)(R_1 Deff_1 + 1 - R_1)}{n_1} + \frac{(\sigma_2^2/R_2)(R_2 Deff_2 + 1 - R_2)}{pn_1}\right)^2}{\frac{\left((\sigma_1^2/R_1)(R_1 Deff_1 + 1 - R_1)/(n_1)\right)^2}{n_1/(R_1 Deff_1 + 1 - R_1) - 1} + \frac{\left((\sigma_2^2/R_2)(R_2 Deff_2 + 1 - R_2)/(pn_1)\right)^2}{pn_1/(R_2 Deff_2 + 1 - R_2) - 1}} \quad (9)$$

and degrees of freedom under $\frac{\sigma_1^2}{R_1} = \frac{\sigma_2^2}{R_2}$

$$v = \frac{n_1}{R_1 Deff_1 + 1 - R_1} + \frac{pn_1}{R_2 Deff_2 + 1 - R_2} - 2 \quad (10)$$

To see the effect of reliability on sample size requirements in power analysis, we assume the reliability of the outcome measures are $R_1 = R_2 = .75$ and the design effects in observed scores $Deff_{1Obs} = (R_1 Deff_1 + 1 - R_1) = 3.25$ an $Deff_{2Obs} = (R_2 Deff_2 + 1 - R_2) = 3.25$. Note that the true design effect in Table 3 and Table 4, below, is attenuated by measurement error in observed scores. Furthermore, when $\sigma_1 = \sigma_2 = 1.0$ and p=1, the

true effect size δ is also attenuated by measurement error $d = \delta * \sqrt{\frac{R_1 + R_2}{2}}$.

The sample sizes required for various levels of δ and are presented in Table 3.

The impact of unreliability on sample size estimation can be seen by comparing the clustered sample of observed scores in Table 3 to the clustered sample of true scores in Table 2. In general, the sample size requirements are approximately estimated by equation (11).

For example, in Table 3 the sample sizes required to detect $\delta_S = .20$, $\delta_M = .50$ and $\delta_L = .80$ with $1-\beta = .80$ for observed scores are now 1,704, 276, and 110. The increases in the sample size requirements beyond the requirements of a simple random sample of true scores are 1,704/394 = 4.32 for $\delta_S = .20$, 276/64 = 4.31 for $\delta_M = .50$, and 110/26 = 4.23 for $\delta_L = .80$. The sample size requirements due to the combined effect of complex sampling and unreliability are increased by approximately a factor of $Deff_{obs} / R = 4.33$.

The above results show that measurement error in the outcome variable reduces the power of the statistical test, which then requires a larger sample size to detect a given effect size. The takeaway in Section 3 is that sample designs should be as efficient as possible. Similarly, the takeaway in this section is that outcome variables should be as reliable as possible.

Equating Error

Equating error is ubiquitous in educational and psychological testing. The need to equate scores on one test to those on another test often comes about because of the need to create new versions of the test which are not perfectly parallel in difficulty. Equating and the associated equating error occurs in practically every test used in the social sciences. The role of complex sampling and measurement error has been addressed in the research literature but there is very little coverage of the role of equating error. We will therefore provide the necessary derivations to show how equating error influences power analysis. Since this is not a comprehensive paper on equating error we will only use one of the classical test theory models of equating error for illustrative purposes. More comprehensive

$$n_{CS,OBS} \approx n_{SRS,TRUE}\left(\frac{\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right) + \left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)/p}{\sigma_1^2 + \sigma_2^2/p}\right) \quad (11)$$

Table 3. Sample Size Table for Clustered Samples with $Deff_1 = Deff_2 = 4.0$ and $R_1 = R_2 = .75$

					d						
	0.09	0.17	0.26	0.35	0.43	0.52	0.61	0.69	0.87	1.04	1.21
					δ						
1 - β	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00	1.20	1.40
0.25	1,432	361	162	93	61	43	33	26	18	14	12
0.50	3,333	836	374	212	137	96	72	56	37	27	21
0.60	4,249	1,065	475	269	173	122	90	70	46	33	26
0.67	4,949	1,240	553	313	201	141	105	81	53	38	29
0.70	5,353	1,341	598	338	218	152	113	87	57	41	31
0.75	6,019	1,507	672	380	244	171	126	98	64	46	35
0.80	6,806	1,704	759	429	276	193	143	110	72	51	39
0.85	7,785	1,949	868	490	315	220	162	125	82	58	44
0.90	9,110	2,280	1,015	573	368	257	190	146	95	67	50
0.95	11,266	2,819	1,255	708	454	316	234	180	116	82	61
0.99	15,926	3,984	1,773	999	641	446	329	252	163	114	85

Note: Given $\alpha = .05$, .two-tailed t-test, $\sigma_1^2 / R_1 = \sigma_2^2 / R_2 = 1.33$, and $p = 1.0$

presentations of equating error (including item response theory equating error) can be found in Kolen and Brennan (2004).

Let's say that we plan to conduct an experiment to assess the efficacy of a new reading program for grade 3 students. At the end of year one, a sample of students are administered a reading test *X* in grade 3. In year two, half of the schools are randomly assigned to the new intensive reading program (treatment group) and half are assigned to the traditional reading program (control group). The study plans to administer a different test Y at the end of year two. Both test *X* and test *Y* measure the same reading content but because they consist of different items we expect they will vary in difficulty. In order to compare the results of test *X* to *Y* test we want both tests to be on the same scale, which means they need to be statistically equated. We create an equating sample of 400 students in year one in which a subset of 200 students are administered test *Y* and a randomly equivalent subset of 200 students are administered test *X*. The subsets of students will constitute the equating sample in an equating design referred to as the randomly equivalent groups design (Kolen & Brennan, 2004). We let the linear equating function be

$$y_x = \left(\overline{x} - \frac{s_x}{s_y}\overline{y}\right) + \left(\frac{s_x}{s_y}\right) y$$

$$\hat{A} = \overline{x} - \frac{s_x}{s_y}\overline{y} \qquad (12)$$

$$\hat{B} = \frac{s_x}{s_y}.$$

In equation (12) $\overline{x}$ and $\overline{y}$ are the sample means in the equating sample, and s_x and s_y are the sample standard deviations of *X* and *Y* in the equating sample, respectively. $\hat{A}$ and $\hat{B}$ are the intercept and slope of the linear relationship between *Y* and *X*. The equating process has now created a new set of scores for test *Y*. Instead of reporting *y* scores we will report y_x scores, which are observed scores on test *Y* converted to the scale of test X. The *y* scores contain measurement error, but the y_x scores contain both measurement error and equating error. The equating error in y_x is caused by the error in estimating $\hat{A}$ and $\hat{B}$ in the equating sample. The linear equivalent of the mean of *Y* is y_x, which is the mean of *Y* re-expressed on the *X* scale. Based on Taylor series linearization (Wolter, 1985) we can think of our test statistic y_x as a function *g* such that $\overline{y}_x = g\left(\overline{y}, \hat{A}, \hat{B}\right) = \hat{A} + \hat{B}\overline{y}$.

The error variance in $\overline{y}_x$ can be approximated by equation (13). Equation (13) shows that the error variance $\sigma^2_{\overline{y}_x}$ of our test statistic $\overline{y}_x$ in group 2 has two components. The first component is $\hat{B}^2\sigma^2_{\overline{y}}$, which is the square of the standard error of the mean, $\hat{\sigma}^2_{\overline{y}}$, rescaled by $\hat{B}$ to be on the *X* scale. The second component is equating error variance s^2_E, which results from the estimation of $\hat{A}$ and $\hat{B}$ in the equating sample. Equation (13) shows that equating error variance is an integral part of the variance of the sampling distribution of the test statistic $\overline{y}_x$ in group 2.

It should be noted that the size of standard error of equating is determined in the equating sample and is influenced by the reliability of the tests in the equating sample as well as the sample size of the equating sample. Less reliable tests will have larger standard errors of equating and smaller sample sizes and design effects in the equating sample will result in larger standard errors of equating. However, once the equating is completed, then the standard error of equating, derived from the equating sample, becomes a permanent margin of error in the scale of the equated test. In subsequent analyses of the equated test, the standard error of equating (obtained from the equating sample) is a part of the

$$\begin{aligned}\sigma^2_{\overline{y}_x} &= \begin{pmatrix} \frac{\partial g}{\partial \overline{y}} & \frac{\partial g}{\partial \hat{A}} & \frac{\partial g}{\partial \hat{B}} \end{pmatrix} \begin{pmatrix} \sigma^2_{\overline{y}} & 0 & 0 \\ 0 & \sigma^2_A & \sigma_{AB} \\ 0 & \sigma_{A,B} & \sigma^2_B \end{pmatrix} \begin{pmatrix} \frac{\partial g}{\partial \overline{y}} & \frac{\partial g}{\partial \hat{A}} & \frac{\partial g}{\partial \hat{B}} \end{pmatrix}^T \\ &= B^2\sigma^2_{\overline{y}} + \sigma^2_A + 2\overline{y}\sigma_{A,B} + \overline{y}^2\sigma^2_B \\ &= \hat{B}^2\sigma^2_{\overline{y}} + s^2_E. \end{aligned} \qquad (13)$$

error component of sampling distribution of the test statistic.

Equations (1)–(3) can again be expanded to include the impact of equating error in addition to sample design effects and test reliability: Equation (14) is an expansion of equation (1) with degrees of freedom shown in equation (15) when $\frac{\sigma_1^2}{R_1} \neq \frac{\sigma_2^2}{R_2}$. Under $\frac{\sigma_1^2}{R_1} = \frac{\sigma_2^2}{R_2}$, the degrees of freedom are:

$$v = \frac{n_1}{R_1 Deff_1 + 1 - R_1} + \frac{pn_1}{R_2 Deff_2 + 1 - R_2} - 2 \qquad (16)$$

With equations (14)–(16) we can see how equating error variance affects power calculations and sample size estimation. For illustrative purposes we assume the reliabilities of the outcome measures are $R_1 = R_2 = .75$ and $Deff_1 = Deff_2 = 4.0$. Furthermore, we assume there is no equating error in group 1 but there is equating error in group 2; therefore, $s^2_{E_1} = 0.0$, $s^2_{E_2} = 0.0025$. Note that the square root of the equating error variance $s^2_{E_2} = 0.0025$ is the standard error of equating $s_{E_2} = 0.05$, or 1/20th of a standard deviation unit. The sample sizes required for various levels of 1–β and δ are presented in Table 4.

The impact of equating error on sample size estimation can be seen by comparing the clustered sample of observed scores with equating error in Table 4 to the simple random sample of true scores in Table 1. The sample size requirements are estimated by equation (17).

For example, in Table 4 the sample sizes required to detect $\delta_S = .20, \delta_M = .50, \text{and } \delta_L = .80$ with 1–β=.80 for observed scores are now 3,345, 299, and 113. The increases in the sample size requirements caused by the combined effects of complex sampling, unreliability and equating error, beyond the requirements of a simple random sample of true scores, are 3,345/394 = 8.49 for $\delta_S = .20$, 299/64 = 4.67 for $\delta_S = .50$, and 113/26 = 4.35 for $\delta_L = .80$. As can be seen, the increase in the sample

$$\beta = T_v\left(t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\frac{\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right)}{n_1} + \frac{\left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)}{pn_1} + s^2_{E_1} + s^2_{E_2}}}\right) - T_{(1+p)n-2}\left(-t_{\frac{\alpha}{2},v}, \frac{\Delta}{\sqrt{\frac{\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right)}{n_1} + \frac{\left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)}{pn_1} + s^2_{E_1} + s^2_{E_2}}}\right), \qquad (14)$$

$$v = \frac{\left(\frac{\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right)}{n_1} + \frac{\left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)}{pn_1} + s^2_{E_1} + s^2_{E_2}\right)^2}{\frac{\left(\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right)/n_1 + s^2_{E_1}\right)^2}{n_1/\left(R_1 Deff_1 + 1 - R_1\right)_1 - 1} + \frac{\left(\left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)/pn_1 + s^2_{E_2}\right)^2}{pn_1/\left(R_2 Deff_2 + 1 - R_2\right) - 1}} \qquad (15)$$

$$n_{CS,OBS,E} \approx n_{SRS,TRUE} \frac{\left(\left(\sigma_1^2/R_1\right)\left(R_1 Deff_1 + 1 - R_1\right) + \left(\sigma_2^2/R_2\right)\left(R_2 Deff_2 + 1 - R_2\right)/p\right)}{\sigma_1^2 + \sigma_2^2/p - n_{SRS,TRUE}\left(s^2_{E_1} + s^2_{E_2}\right)} \qquad (17)$$

Table 4. Sample Size Table for Clustered Samples with $Deff_1 = Deff_2 = 4.0$, $R_1 = R_2 = .75$, and $s^2_{E_1} = 0.0, s^2_{E_2} = 0.0025$

						d					
	0.09	0.17	0.26	0.35	0.43	0.52	0.61	0.69	0.87	1.04	1.21
						δ					
1 - β	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00	1.20	1.40
0.25	2,434	402	170	95	62	44	33	26	18	14	12
0.50	83,835	1,100	418	225	142	99	73	57	37	27	21
0.60	*	1,535	550	291	182	126	93	71	47	34	26
0.67	*	1,927	657	343	214	147	108	83	54	39	29
0.70	*	2,183	722	374	232	159	117	89	58	41	31
0.75	*	2,662	832	426	262	179	131	100	65	46	35
0.80	*	3,345	971	489	299	204	148	113	73	52	39
0.85	*	4,440	1,157	570	346	234	170	130	83	59	44
0.90	*	6,641	1,434	685	411	277	200	152	97	68	51
0.95	*	15,007	1,964	888	522	348	250	189	120	84	62
0.99	*	*	3,620	1,401	785	511	363	272	171	118	87

Note: Given $\alpha = .05$, .two-tailed t-test, $\sigma_1^2 / R_1 = \sigma_2^2 / R_2 = 1.33$, and $p = 1.0$ "*" indicates the sample size requirement approaches infinity.

size requirements when equating error is introduced is a function of the sample size.

Equating error increases the sample size requirements disproportionately for small values of δ. For example, for $1-\beta = .80$ and $\delta_S = .20$, the sample size is 8.49 times larger than the sample required for a simple random sample. On the other hand, for $1-\beta = .80$ and $\delta_L = .80$, the sample size is 4.35 times larger than the sample required for a simple random sample. In fact, for very small values such as $\delta = .10$ it is not possible to have a sample size large enough to detect δ with power equal to .80.

Unlike sampling error and measurement error, equating error is not reduced by increases in sample size. Therefore, in larger sample sizes equating error has a proportionally larger impact on the variance of the sampling distribution. In practice this means that as the sample size grows, sampling and measurement contribute smaller and smaller components of error variance. Equating error, on the other hand, becomes the dominant component of error variance in the sampling distributions with large samples. This is because equating error variance is a constant component in the variance of the sampling distribution regardless of sample size. In fact, equating error variance is the lower limit of the variance of the sampling distribution. Even if the variance due to sampling and measurement becomes essentially zero in very large samples, the equating error variance remains unchanged.

Conclusion

This paper discusses how sample design effects, measurement error, and equating error affect power analysis. It has been shown that all three sources of error reduce power and increase sample size requirements to detect the target effect size. All three sources of error reduce power by increasing the variance of the sampling distribution of our test statistic. Complex sampling increases the variance of the sampling distribution through the design effect of the sample. Measurement error increases the variance of the sampling distribution by increasing the observed variance of the population. Finally, equating error reduces power by adding a constant error component to the sampling distribution. In general, equating error has been an unrecognized source of instability in many statistical research studies and has received minimal attention in the research literature (Phillips, Doorey, Forgione, & Monfils, 2011).

Reviews of the literature show that many published research studies have low statistical power. For

example, Cohen (1962) found a median power of .48 for a medium effect sizes, and Sedlmeier & Gigerenzer (1989) found a median power of .37 for medium effect sizes. Low power means the study has a reduced chance of detecting a true effect (Type II error). This paper shows that when the outcome variables are unreliable and contain equating error, power can be substantially lower than we think. This implies that the research literature may be populated by even more Type II errors than we realize. If important error components in the variance of the sampling distribution are ignored, the researcher may substantially underestimate sample size requirements.

Practitioners should realize that the components of error variance covered in this paper are generally present in most research studies in the social sciences. Ignoring the components of error variance does not make them go away. For example, let's say a researcher does an a priori power analysis and concludes he/she needs a sample size of 394 students in both the control group and the treatment group to have a power equal to .80 to detect a small effect (see Table 1). The researcher may not be aware that the sample has a design effect of 4.0, the test scores have a reliability of .75 and the test scale has an equating error equal to .05 (these are the examples used in this paper). In order to have a power equal to .80 to detect a small effect size, the researcher actually needs a sample size equal to 3,345, not 394 (see Table 4). Ignoring the components of error variance in an a priori power analysis can result in a substantially under powered research study.

References

Chow, S. C., Shao, J., & Wang, H. (2002). A note on sample size calculation for mean comparisons based on non-central t-statistics. *Journal of Biopharmaceutical Statistics*, 12(4), 441–456.

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.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley and Sons.

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

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). New York: Academic Press.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

Cornfield, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*, 41, 654–661.

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

Hedges, L., & Rhoads, C. (2009). *Statistical Power Analysis in Education Research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Kish, L. (1965). *Survey sampling*. New York: John Wiley and Sons.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Konstantopoulos, S. (2009, May). Using power tables to compute statistical power in multilevel experimental design. *Practical Assessment, Research & Evaluation*, 14(10), 1–9.

Levy, P. S. (1999). *Sampling population: Methods and applications* (3rd ed.). New York: John Wiley and Sons.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press, Brooks/Cole Publishing.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

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

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Phillips, G. W., Doorey, N. A., Forgione, P. D., & Monfils, L. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers.

Raudenbush, S. (1997). Statistical analysis and optimal test design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114.

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

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, 116(4), 359–369.

Wolter, K. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Citation:

Phillips, Gary W., and Jiang, Tao. (2016). Measurement Error and Equating Error in Power Analysis. *Practical Assessment, Research & Evaluation*, 21(9). Available online: http://pareonline.net/getvn.asp?v=21&n=9

Corresponding Author

Gary W. Phillips
Vice President and AIR Institute Fellow
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835

gwphillips [at] air.org