3-24-2006

Reliability, Effect Size, and Responsiveness and Intraclass Correlation of Health Status Measures Used in Randomized and Cluster-Randomized Trials

Paula Diehr
University of Washington, pdiehr@u.washington.edu

Lu Chen
University of Southern California, lchen@childrensoncologygroup.org

Donald L. Patrick
University of Washington, donald@u.washington.edu

Ziding Feng
University of Washington & Fred Hutchinson Cancer Research Center, zfeng@fhcrc.org

Yutaka Yasui
Fred Hutchinson Cancer Research Center, yyasui@ualberta.ca

Introduction

New survey instruments are generated every day with documentation that reports their psychometric properties, such as Reliability, Effect Size and Responsiveness. These statistics may be used to choose among valid instruments whose content measures what the researcher intends to measure. (In the following we shall sometimes refer to these three characteristics jointly as "Reliability" for short, since they will all be seen to be related). Another property, relevant for cluster-randomized trials, is the Intraclass Correlation (ICC) among persons within a cluster. Reliable instruments can reduce the necessary sample size, but there are settings where the most reliable instrument is too "expensive" for the projected use.

The cost of using a particular instrument may be thought of in terms of time (patient or interviewer time), length (space required on a survey designed to examine multiple facets of health, and the related opportunity cost of not measuring the other facets as well), or dollars (costs of proprietary instruments, highly trained interviewers or interpreters), and the accuracy of the resulting data (lower response rates or accuracy due to increased patient burden).

The best instrument is one that fits the study needs, but is not necessarily the most reliable instrument. An instrument used to diagnose, treat or refer an individual should have high Reliability; Nunnally suggests that a value of .90 be used.[1] However, most research involves the comparison of groups of persons, often on their mean score. It is well known, [2] but perhaps less well understood, that less reliable instruments can have high power to detect difference in the means of two groups, if the sample size per group is high enough. For example, a Reliability of .70 has been recommended as a minimum Reliability to be used in comparing

two groups. [1, 13, 3]

Behavior change interventions are often conducted as cluster-randomized trials when the intervention naturally occurs at the cluster level. [4 5 6 7 8 9 10 11] In this situation, the study design requires choosing the number of clusters per treatment and the number of persons to survey per cluster, as well as which instrument to use. There has been, to our knowledge, no discussion of how to choose the best instrument for a cluster-randomized trial.

The purpose of this paper is to define a statistical model to define the psychometric statistics, to show how they are related to one another, to show how they are related to the power and necessary sample size of a study, and to discuss the data needed to estimate them accurately. In addition, we develop approximate equations to calculate the sample size needed to assess the Reliability (etc.) of new psychometric instruments.

Methods

We begin this paper with a review of several psychometric statistics, starting first with a statistical model for the true values, and then introducing the usual "true value plus error" model for an instrument. We begin at the person level and then consider randomized trials and cluster-randomized clinical trials. We use a combination of exact calculations, simulation, and real data to describe the psychometric properties. For the simulations we generated data described by the model in Table 1, with varying values of SD (1,2,5,10) and N per treatment group (10, 20, 50, 100, 250, 500, 1000), with 151 replicates for each set of parameters. For each combination of N and SD we estimated Delta, Reliability, Effect Size, and Responsiveness (defined below) from each sample. We calculated the standard error of the 151 estimates. We found from exploratory analyses that the logarithm of the sample size was linearly related to the other parameters, and

so regressed the log of the sample size on the standard error and the true value of the parameter. We used the resulting equation to estimate the sample size needed to estimate each psychometric statistic to a given level of precision. This is explained more detail in the section under Reliability, below.

True State

Consider a construct, perhaps a person's true health, denoted as Z, and described in Table 1, where a higher value denotes better health. For this example, Z is assumed to be normally distributed in the population, with $\mu_z = 50$ and $\sigma_z = 10$. We will consider three time points, T_0, T_1, and T_2, where T_0 and T_1 are "close together" (perhaps a week apart) and T_2 is perhaps a year later. The true values of Z (health) for those times will be denoted Z_0, Z_1, and Z_2. T_0 and T_1 are close enough together that health has not changed ($Z_0=Z_1$). From T_1 to T_2 there is some natural change (secular trend) over time, which is normally distributed with mean $\mu_{trend} = 1$ and standard deviation $\sigma_{trend} = 1$. Half of the people will thus improve 1 or more points, and half will improve less; in fact, 16% will be sicker at T_2 than they were at T_1. Finally, half of the hypothetical people are assigned to an experimental treatment which raises each person's Z value exactly 3 points at T_2. The true change from T_1 to T_2 is the secular trend (mean 1, standard deviation 1) plus the treatment effect (3). Z_2 thus has mean 54 in the treatment group and 51 in the control group, and in both groups the variance is 101 (because it includes the variance of the secular change).

[Table 1 about here]

Although different parameter values could have been chosen for the true situation, our

interest is only in how well a particular instrument measures truth. The bottom half of Table 1 deals with an instrument that estimates Z, with some error. We refer to the value from the instrument as Y. Y (given Z) is equal to the true value of Z plus error, where the error has mean M and standard deviation SD, and M is independent of Z. If M is zero, then Y is an unbiased estimate of Z. We will let M equal zero, without loss of generalizability, since the psychometric measures we will discuss all remove the mean. In the following we will consider the characteristics of Z as fixed, but will examine the effect of changing SD. (We use Greek symbols to denote the values that will be held constant). For future reference, the distributions of the Z's, the Y's and of the change score $Y_2 - Y_1$ are summarized in Table 2. These distributions can be derived from the information in Table 1. In the following we consider all parameters in Table 1 to be fixed except SD, and will examine the effect of varying SD.

[Table 2 about here]

Correlations among measures.

Some correlations among the various measures are summarized here. Consider first the correlation between Y_0 and Z_0, which we will refer to as r_{yz}. Although these correlations can be calculated algebraically, a more mnemonic way is to recall that R^2_{yz} is the proportion of variation in Y that is explained by Z. From Table 1 and Table 2, it is clear that this proportion is $\sigma^2_z /(\sigma^2_z + SD^2) = 100 / (100 + SD^2)$. If SD is 2.29, then $R^2_{yz} = .95$, and its square root is $r_{yz} = .975$. The correlation between Y_0 and Y_1 (r_{yy}) can be thought about in two parts. Y_0 explains R^2_{yz} of the variability in Z_0 which also, because $Z_0 = Z_1$, explains R^2_{yz} of the variability in Y_1. The percent of variability in Y_1 explained by Y_0 is then the product of these two percentages, or

$(R^2_{yz})^2$, and $r_{yy} = R^2_{yz}$. These correlations are shown in Table 3, for several values of SD. Correlations all decrease as SD increases.

[Table 3 about here]

Psychometric Characteristics of the Instrument Y

Three commonly cited properties of an instrument are its test-retest Reliability, Effect Size, and Responsiveness. [2, 19 12 13 14 15 16] These will be discussed in turn. Another property, Intraclass Correlation within clusters, is discussed in the section on cluster-randomized trials.

Reliability. Reliability is a measure of whether the same person, under the same conditions, would give the same response. It is usually estimated from test-retest data (ideally two measures taken close enough in time that the true value has not changed, but far enough apart that the previous response doesn't affect the current answer), which are used to estimate the intraclass correlation *within a person*. We will refer to this intraclass correlation only as "Reliability", to avoid confusion with the Intraclass Correlation within a cluster, discussed under cluster-randomized trials. Reliability is the proportion of the variance among people's scores that is accounted for by their true values. In our situation Reliability is clearly $Var(Z_0) / Var(Y_0) = r_{yy} = \sigma^2_z / (\sigma^2_z + SD^2) = 100 / (100 + SD^2)$, but ordinarily Reliability must be estimated. Although usually defined in terms of analysis of variance, estimates of test-retest reliability can be obtained more simply from a paired t-test of the test-retest data, as demonstrated elsewhere. [17] In our setting, Reliability would be estimated from Y_0 and Y_1, as follows:

$$\mathrm{Re}liability = \frac{s_{Y1}^2 + s_{Y0}^2 - s_{(Y1-Y0)}^2}{s_{Y1}^2 + s_{Y0}^2 - [s_{(Y1-Y0)}^2 / N] + (\overline{Y}_1 - \overline{Y}_0)^2}$$

Values of the Reliability of Y for selected values of SD are shown in Table 4. Perfect Reliability (1.0) is achieved if SD = 0, and it becomes lower for larger values of SD.

[Table 4 about here]

We also simulated person-level data (Y's) according to the model in Table 1 and estimated the psychometric statistics from each sample. We varied SD and N, the number of persons per treatment group, and examined the distributions of the resulting estimates to examine the properties of estimated Reliability. For example, with samples of size 20, where the true Reliability was .80, the estimated Reliability ranged from .45 to .95 with a mean of .80 and a standard deviation of .09. The histogram of estimates from this simulation is shown as Figure 1. Clearly, Reliability estimates based on a sample of only 20 subjects are likely to be inaccurate.

[Figure 1 about here]

To explore this variability we generated 151 such estimates of Reliability for each combination of N (from 20 to 2000), and Reliability (from .5 to .99). For each set of parameters we calculated the standard deviation of the 151 estimates of Reliability ($s_{Reliability}$), which decreased as either N or the true Reliability increased. We regressed ln(N) on true Reliability, ln (Reliability), and ln($s_{Reliability}$). The fit was excellent (R^2= .92). From this regression, the N needed to achieve a specified accuracy for the estimated Reliability can be calculated as:

$$N = \exp[69.4 - 1.75*\ln(s_{Reliability}) - 75.1*Reliability + 45.6*\ln(Reliability) + .619^2/2],$$

where .619 is the standard error of the regression, and is included to adjust for the re-

transformation bias in the lognormal distribution. [18] For example, suppose the instrument is expected to have Reliability about .8, and the investigators wish the estimate of Reliability to fall between .7 and .9 with 95% probability; that is, the confidence interval will have length 0.20. Assuming normality of the estimates, about 95% of the estimated Reliability values will fall in the range Reliability $\pm$ 2* $s_{Reliability}$, meaning that the value of $s_{Reliability}$ must be 0.05. From the equation, the required N is 100; that is, about 100 people would be needed for a test-retest study to estimate the Reliability to this level of accuracy. This number would be lower if the true Reliability was higher, or if a wider range of confidence was permitted. For example, if the range of 0.6 to 1.0 was permissible, the desired $s_{Reliability}$ would be .10, and the required N would be 30. For $s_{Reliability}$ = 0.025, a sample of 336 would be needed. If there is no information on the probable Reliability of a new instrument, it is conservative to assume a smaller value of Reliability for this calculation.

Delta (Δ). The next two psychometric measures (Effect Size and Responsiveness) require Δspecification of the change in Y in a specific situation, which we shall refer to as **Δ**. The quantity **Δ** is defined in several different ways. It may be defined as the minimum clinically important difference or change, which is not usually well specified. [19] If **Δ** is defined as the change associated with a treatment of known efficacy, it is obvious that **Δ** = 3 for our situation, since the treatment makes a change of 3 points for each person in the treatment group. However, if we had a different treatment in mind, which changed the treatment group by 10 points, then **Δ** for that situation would be 10. Thus, an instrument could have many values of **Δ**,depending on the intervention effect that was assumed. For practical reasons, **Δ** is often estimated from available data. It has been estimated as the mean change over time in a group of patients in the

treatment group of an RCT, or in patients who seemed improved by some other standard.[17] Under our model, such an estimate would include the secular trend, and the estimated **Δ** would be 4 rather than 3. Others have subtracted the change in the "control" group from the change in the "treatment" group, which would provide an estimate of **Δ**= 3. [20] It is important to specify the source of the **Δ** used in calculations.

Effect Size (ES). Although Effect Size has many definitions, [21] the most common estimate of ES is **Δ** divided by the standard deviation of the "before" value for a particular instrument, with a particular value of SD. Under our model this would be:

$$ES = \frac{\Delta}{\sigma_{Y1}} = \frac{\Delta}{\sqrt{\sigma_Z^2 + SD^2}} = \frac{3}{\sqrt{100 + SD^2}}$$

For SD=0, ES = 0.3, and ES approaches zero as SD becomes larger. Values of ES are shown in Table 4 for different values of SD.

We used the simulated data to explore the variability of the estimated Effect Size as explained in the Reliability section. (We estimated Δ as the difference in change between the treatment and control groups). Variability was lower for larger N, and for instruments with larger Effect Size. We regressed ln(N) on Effect Size, $\ln(s_{\text{Effect Size}})$, and ln (Effect Size). The fit was excellent (R^2= .996). The N needed to achieve a specified accuracy for the estimated Effect Size is:

$$N = \exp[116.2 - 1.90*\ln(s_{\text{Effect Size}}) - 205.5*\text{Effect Size} + 46.1*\ln(\text{Effect Size}) + .143^2/2].$$

For example, the necessary N if the expected Effect Size = .25 and the desired $sd_{\text{Effect Size}}$ is .05 is

thus 689 (343 per group); if $sd_{\text{Effect Size}} = .10$, only 92 persons per group are needed. It is conservative to assume a smaller Effect Size.

Responsiveness. Responsiveness is the ability of an instrument to detect minimal clinically important differences, which is defined as the expected change in Y under a treatment of known efficacy divided by the standard deviation of change in stable subjects. [19] Under our model, the stable subjects are the controls, and

$$\text{Responsiveness} = \frac{\Delta}{\sqrt{\sigma^2_{trend} + 2 * SD^2}} = \frac{3}{\sqrt{1 + 2 * SD^2}}$$

Values of Responsiveness for different values of SD are in Table 4. The highest possible Responsiveness, when SD = 0, is $\boldsymbol{\Delta} = 3.0$.

In the simulated data, the estimates of Responsiveness were less variable for larger N, but surprisingly became more variable as Responsiveness increased. We regressed ln(N) on Responsiveness, $\ln(s_{\text{Responsiveness}})$, and ln (Responsiveness). The fit was excellent ($R^2 = .994$). We used the regression equation to calculate for the N needed to achieve a specified accuracy for the estimated Responsiveness, with results as follows:

$$N = \exp[-0.354 - 1.870*\ln(s_{\text{Respons}}) + 1.479*\text{Respons} - 0.354*\ln(\text{Respons}) + 0.161^2/2].$$

The necessary N if the expected Responsiveness= 1.0 and the desired $sd_{\text{Respons}} = .05$ would be 846 (423 per group); for $sd_{\text{Respons}} = .10$, only 116 persons per group would be needed. Unlike the other statistics, it is more conservative to assume that Responsiveness is larger.

One attractive feature of the Responsiveness statistic is that it can be used directly to

estimate the necessary sample size per group for detecting a difference of **Δ** in the treatment and control change scores. For 80% power, for example, N per group = 2* $(1.96+.84)^2$/$Respons^2$. The necessary sample sizes per group to achieve 80% power in our hypothetical clinical trial are shown in Table 4 for various values of SD. (The tabled sample size for SD=0, (3.5), is not accurate because we used the normal approximation rather than the t-test).

Cluster Randomized Trials (CRTs)

To this point we have compared the means of two groups of persons. Cluster randomized trials (CRTs) are conducted when the intervention is performed at the cluster level, but the effects are measured on individuals. [7,8,9, 25, 22] Investigators must choose both the number of clusters (C) to be randomized to treatment or control, and the number of persons per cluster (N) to be evaluated, in addition to choosing the instrument to be used in the assessment. For simplicity, we assume that the N persons in each cluster will be evaluated at times T_1 and T_2, the change in the two scores calculated, and the mean change calculated for each cluster. The 2*C cluster means (C for treatment and C for control), will then be analyzed using a t-test. (Randomization tests [23] and multi-leveling model are alternative approaches).

In the hypothetical CRT, then, the same person-level model applies, but the people were assigned to treatment/control by cluster, with N persons per cluster. We further assume that the true mean change is different in each cluster (independent of any intervention), with the differences distributed as Normal $(0, \sigma^2_C)$. That is, the true mean change is different in each community, but the clusters in the treatment group will have an average change 3 points higher than the controls. Intraclass correlation is the correlation among persons within the same cluster. The ICC is also the fraction of the total variation in the data that is attributable to the unit of

assignment (the cluster—in Murray, page 7, where our clusters are his groups and our treatment groups are study conditions). [24] The Intraclass Correlation (within a treatment group) is the variance among clusters divided by that variance plus the variation among people within clusters:

$$ICC_{CRT} = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{change}^2} = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{trend}^2 + 2 * SD^2}$$

ICC_{CRT} (henceforth referred to simply as ICC) is near to 1 if there is high variability among clusters in the mean amount of change, and is smaller if there is a good deal of variability in the trend over time. Table 4 shows some values of ICC assuming that $\sigma^2_C = 1.0$. As the instrument becomes less reliable (SD increases), the ICC decreases because there is relatively more variation within the clusters than among them.

Feng and Grizzle provide sample-size formulas for the situation in which the number of clusters per group is 10 or greater, parameterized in terms of the ICC. [25] Since we wish to consider smaller numbers of clusters, we present different calculations here. In the following we selected values of σ^2_C to yield nice values of the ICC and also varied SD, C, and N to determine the sample size per cluster needed to obtain 80% power. This design can be thought of as an analysis of variance for a nested design, with clusters nested in treatment and persons nested in clusters. The variance of the mean for a treatment group is the variance among persons ($\sigma^2_{trend}+2SD^2$) divided by the number of persons (NC) plus the variance among clusters (σ^2_C) divided by the number of clusters (C). [26]

$$Var(\bar{D}) = \frac{\sigma_C^2}{C} + \frac{\sigma_{trend}^2 + 2SD^2}{NC}$$

Since the number of clusters per group is usually small, the number of clusters needed to achieve a power of β must be specified in terms of the percentiles of the t-distribution instead of the normal distribution. Following the usual derivation of sample size in the normal case, and assuming the variance of D is the same in both treatment groups, we need to find a value of C such that under the alternative hypothesis the probability of rejecting the null hypothesis is 1-β, when the difference is actually Δ, or

$$\Pr\left(\frac{\overline{D}_1 - \overline{D}_2}{s\sqrt{2/C}} > t_{2C-2,1-\alpha/2}\right) = 1-\beta$$

The quantity in parentheses on the left does not have a central t-distribution under the alternative hypothesis, but subtracting $\frac{\Delta}{s\sqrt{2/C}}$ from both sides of the inequality yields

$$\Pr\left(\frac{\overline{D}_1 - \overline{D}_2 - \Delta}{s\sqrt{2/C}} > t_{2C-2,.1-\alpha/2} - \frac{\Delta}{s\sqrt{2/C}}\right) = 1-\beta$$

where the left side does have a central t distribution. The equality holds only if

$$t_{2C-2,.1-\alpha/2} - \frac{\Delta}{s\sqrt{2/C}} = t_{2C-2,\beta},$$

or the number of clusters per group is

$$C = \frac{(t_{2C-2,1-\alpha/2} + t_{2C-2,1-\beta})^2 2s^2}{\Delta^2}$$

Letting T_{2C-2} be the term in parentheses, and setting s^2 to a single community's variance, i.e., C times the variance of $\overline{D}$, the necessary number of clusters (C) for a fixed value of N is:

$$C = \frac{2T_{2C-2}^2(\sigma_C^2 + (\sigma_{trend}^2 + 2SD^2)/N)}{\Delta^2}$$

As T_{2C-2} is different for different values of C, this equation must be solved iteratively. We

solved instead for the number of persons needed per cluster (N), for a fixed number of clusters per treatment group (C).

$$N = \frac{2T^2_{2C-2}(\sigma^2_{trend} + 2SD^2)}{\Delta^2 C - 2T^2_{2C-2}\sigma^2_C}$$

Although this does not reduce to a convenient function of the ICC, the sample sizes needed per cluster for different values of C and ICC can be calculated, as shown in Table 5. For example, if ICC=.01, Reliability of the instrument= .25, and C = 20 clusters per treatment group, then a study with 124 persons per cluster will yield 80% power with alpha=.05, and only 23 persons per cluster are needed if the Reliability is .50. Table 5 shows that higher Reliability, more clusters, and lower ICC are all associated with smaller required sample sizes. There are many different configurations of Reliability, C, and N that will allow a trial with 80% power, and these configurations are different depending on the ICC of the instrument. For example, if there are only two clusters per treatment group but the ICC is .01 and Reliability is .95, the CRT will have 80% power with 59 persons per cluster. Blank cells means that it is not possible to achieve 80% power with this configuration. When the number of clusters is small, a more reliable instrument may be needed.

[Table 5 about here]

Feng and Grizzle found that an asymptotic equation for the variance of estimated ICC was accurate for situations with 10 or more clusters and 30 or more persons per cluster. [24] Because we wanted to study the behavior of estimates of the ICC for smaller numbers of clusters, we created 200 random datasets following the CRT model, and estimated the ICC_{CRT} from each dataset. Murray (p. 81) suggests that an ICC value of .02 is typical. [27] (Campbell et

al report ICC's for cost data as large as 0.47). [28] For the simulation we considered ICC = .02 and .05; SD = 2.29, 5, and 10; and N = 50 and 100. As there are many possibilities, we present only some typical cases. Suppose the goal is to estimate the ICC to within plus or minus .01 with 80% probability; (that is, 80% of all estimates will be within $\pm$ 0.01 of the true ICC). This can be achieved for ICC = .02 and N = 100 with C = 25; for ICC = .02 and N = 50 with C = 40; and for ICC = .05 and N = 50 more than 80 clusters are required. Since the number of available clusters (C) is usually much smaller that these calculated values, most published ICC estimates are probably inaccurate. The asymptotic formulas for the variance of the ICC of Feng [24] and Donner [22] were accurate for 10 or more clusters, but underestimated the variance for 2 and 5 clusters.

Cost of Using a Particular Instrument

The cost of including Quality of Life measures in clinical trials has been considered, [29] but the cost associated with a particular instrument was not discussed. Proprietary instruments have license fees. An instrument that requires highly trained professionals to administer and interpret it is more expensive, and a more detailed instrument may require more of their time. If the total length of the survey is constrained, use of a particular instrument has opportunity costs, in that using a long instrument to measure one patient characteristic could preclude measuring different characteristics well or at all. Other non-monetary costs of a longer instrument include subject burden, and the likelihood that subject fatigue will lead to lower quality data.

In a randomized clinical trial (RCT) a more reliable instrument would usually be preferred because it permits smaller sample sizes, as shown in Table 4. However, if there were large differences between the costs of the most reliable instrument and an alternative, it could be

more cost-effective to use the less reliable instrument and achieved the desired power through an increase in sample size. There are many situations in which the sample size is determined based on some other criterion. Many studies are powered to detect mortality differences, which usually provides more subjects than needed to detect differences in quality of life.[30] Large, simple trials are designed especially to study very large numbers of persons, and their success depends on using extremely simple data collection instruments. [31] The study's sponsor may require including all people in a certain class, such as all primary care patients in a clinic. In those cases, the most reliable instrument might not be needed. For example, if the sample size was fixed at 350 per group for some reason, Table 4 shows that an instrument with Reliability of only .50 would have sufficient power to detect the difference of interest in our example. Such a choice might reduce respondent burden and other costs. This also holds true for cluster-randomized trials.

Example: Data from the LIDO Study

The Longitudinal Investigation of Depressive Outcomes (LIDO) study was an observational study of depression in 6 international cities. [32] Primary care patients who met eligibility criteria were assessed for depression using the Composite International Diagnostic Interview (CIDI). [33] There were 981 persons who had clinical depression at baseline and a valid CIDI assessment nine months later, which was the "gold-standard" for whether their depression had remitted. Here we use these data at the person level and also use the mean change for each of the 6 cities, to illustrate the points made earlier. We compare the MHI5 mental health subscale of the SF-36 to the single item (from that scale) "Have you accomplished less than you would like as a result of any emotional problems (such as feeling

depressed or anxious)?" Table 6 shows information about the two mental health instruments, including the sample mean, standard deviation, the estimated Reliability (calculated from the baseline and 6 week measures), Effect Size, and Responsiveness. The 5-item scale had better psychometric characteristics than the single item, but the difference was not large.

We assumed the marginal cost of using the two instruments was proportional to their length, assuming a stem question plus 5 or 1 additional questions, for costs of 6 or 2, respectively. In planning a new study, the necessary sample size per group to achieve 80% power is $2*(1.96+.84)^2$ / Responsiveness^2 . The shorter instrument would require a larger sample size, but the total cost (unit cost * sample size) would be only about half as high for the shorter instrument (58 versus 108). This is an example in which the less reliable instrument might be preferred, if it included the content that was necessary for the investigation at hand. The decision would need to balance these marginal costs with the fixed costs of obtaining an additional person.

We also estimate σ^2_C from the 6 cities (clusters), and calculated the ICC at baseline, at 9 months, and for the average change. The cross-sectional ICC estimates were higher for the single item at baseline and 9 months, but the ICC for change was slightly smaller for the single item. Three of the estimates were near .02, and three were substantially higher. Based on the simulation results described above, it is unlikely that the estimated ICC's are close to the true value.

Discussion

We used the usual true score plus error model to describe a valid instrument that measures true health with a certain amount of error. Under this model, all of the psychometric statistics were a function of SD, the measurement error, and so behaved in similar ways. We also noted some inconsistencies in how parameters were defined (particularly), the true treatment effect), which would often restrict the usefulness of the Effect Size and Responsiveness estimates in the literature. An instrument has one Reliability, but may have a variety of Responsiveness and Effect Size values, since they depend on). Estimates in the literature are also sensitive to the values of σ^2_C , σ^2_Z , or σ^2_{Trend} in the study.

Instruments should be chosen to meet the purposes of the investigation. This does not always involve choosing the "best" instrument, which may be too expensive and have features that are not needed. In a clinical trial setting, where the object is to achieve a specified power to detect a specified treatment effect, any of the instruments under consideration could be suitable if only the sample size were high enough; that is, you can make it up in volume. If only a few subjects are available, a highly reliable instrument can increase the power. An example using data from the LIDO study suggested that a single item might sometimes be chosen in preference to a 5-item question on the basis of cost-effectiveness if the costs of finding additional subjects are not high.

There is little in the literature about the sample sizes needed to estimate the psychometric statistics accurately. Our simulations show that rather large samples are required to provide accurate estimates. Nunnally suggested N = 300 for estimating Reliability, but our results suggest that smaller N's may be sufficient for highly reliable instruments. Precise estimates of Effect Size and Responsiveness will require larger N's. We recommend that re-sampling

methods such as bootstrap be used when Reliability is estimated, to provide future users with the degree of accuracy. This is easily feasible but rarely done. The problem of calculating ICC_{CRT} is even more acute because it requires a large number of clusters and is probably different depending on the nature of the cluster, as well as on the other parameters.

It is interesting to compare the psychometric statistics (see Table 4). In every case, an instrument with smaller SD will result in a larger value of the statistic, because SD is in the denominator of each statistic. Note, however, that Reliability is also strongly related to the variance among people (σ^2_z), as is well known. If σ^2_z is large relative to SD (perhaps in a general population), then Reliability will approach 1, and if it is small (perhaps in patients recovering from the same surgical procedure), Reliability will be primarily a function of SD, or the instrument. Reliability is thus not the property only of the instrument, but also of the norming sample. Effect Size is a function of the instrument, the population, and also Δ, which can vary substantially as noted above. Responsiveness is not related to σ^2_z, but is related to Δ, SD^2, and also to the variance of the secular trend over time. An evaluation setting with substantial variability in how subjects changed over time would show less Responsiveness than one in which all people moved in the same direction by about the same amount. Finally, ICC is a function of all of these factors plus variation in the type of cluster; the ICC is likely to be higher in interacting units such as families and workplaces than in counties and states. Published values of the psychometric statistics are clearly most valuable when calculated in a similar context to the new planned investigation.

The calculations above give some guidance about the sample sizes needed to obtain good estimates of the psychometric statistics. Once the data have been collected, a bootstrap approach

could be used to provide approximate confidence intervals for the estimate, as suggested by Feng and Grizzle. [25]

Limitations. These findings should not depend very much on the exact form of the model we assumed. The model did not include a person-level error term, but it would have cancelled out in the analysis of change that was assumed. We assumed that the error terms were independent of Z and of one another, to make the calculations more straightforward. We could have made the treatment effect random instead of fixed. None of these are likely to have affected the results much. We let Y be an unbiased estimate of Z. We could alternatively have let $E(Y) = a+bZ$, since a and b would have disappeared in the calculation of the psychometric statistics. If Y was a non-linear function of Z, results would have been similar in kind but would not be exact. We examined a reasonable range of model parameters. In the simulations we estimated the value of Delta separately from each sample. If we had used a fixed value for Delta, the variances of estimated effect size and responsiveness would have been lower.

Conclusion. In RCTs and CRTs, there are situations in which the best instrument in terms of psychometric qualities is not the most cost-effective instrument for the study at hand. Efforts to decrease the complexity of current instruments, such as the SF-36, are likely to pay off in many situations.

Table 1

The Models for Z (Truth) and Y (Instrument)

Z = TRUE STATE

Z_0 $N(\mu_z = 50, \sigma_z = 10)$

Z_1 Z_0

Z_2 Z_1 + $N(\mu_{trend} = 1, \sigma_{trend} = 1)$ + **Δ**= Treatment effect = 3

Secular Trend Treatment

Y = INSTRUMENT

y_0	Z_0	+	**ε ;**	**ε** ~	N(M, SD)	**M=0; SD = 0,1,2,5,10**
y_1	Z_1	+	**ε ;**	**ε** ~	N(M,SD)	
y_2	Z_2	+	**ε ;**	**ε** ~	N(M, SD)	

Table 2

Distributions of True and Measured Health Variables

Variable	**Mean**	**Variance**
Z_0	$\mu_z = 50$	$\sigma^2_z = 100$
Z_1	$\mu_z = 50$	$\sigma^2_z = 100$
Z_2	$\mu_z + \mu_{trend} = 51$ (control) $\mu_z + \mu_{trend} + \Delta = 54$ (treatment)	$\sigma^2_z + \sigma^2_{trend} = 101$
Y_0	$\mu_z = 50$	$\sigma^2_z + SD^2 = 100+SD^2$
Y_1	$\mu_z = 50$	$\sigma^2_z + SD^2 = 100+SD^2$
Y_2	$\mu_z + \mu_{trend} = 51$ (control) $\mu_z + \mu_{trend} + \Delta = 54$ (treatment)	$\sigma^2_z + \sigma^2_{trend} + SD^2 = 101+SD^2$
$Y_2 - Y_1$	$\mu_{trend} = 1$ (control) $\mu_{trend} + \Delta = 4$ (treatment)	$\sigma^2_{trend} + 2*SD^2 = 1 + 2*SD^2$

* The distribution of the Y's is unconditional; that is, not conditioned on Z.

Table 3

Correlations among True and Measured Health Variables

SD	R^2_{yz}	r_{yz}	R^2_{yy}	r_{yy}
2.29	.950	.975	.902	.950
3.33	.900	.949	.810	.900
5.00	.800	.894	.640	.800
6.55	.700	.837	.490	.700
10.00	.500	.707	.250	.500
17.30	.250	.500	.063	.250

$\mathbf{R^2_{yz}}$ is the correlation between the true (Z) and measured (Y) values at a particular time.

$\mathbf{R^2_{yy}}$ is the correlation between Y_0 and Y_1.

Table 4

True Psychometric Characteristics* of the Instrument Y As a Function of Measurement Error (SD)

	Reliability	**Effect Size**	**Responsive-ness**	**N per group for 80% power**	**ICC for Change from T_1 to T_2**
	$\frac{\sigma_z^2}{\sigma_z^2 + SD^2}$	$\frac{\Delta}{\sqrt{\sigma_z^2 + SD^2}}$	$\frac{\Delta}{\sqrt{\sigma_{trend}^2 + 2*SD^2)}}$	$\frac{2*(1.96+.84)^2}{Responsiveness^2}$	$\frac{\sigma_C^2}{\sigma_C^2 + \sigma_{trend}^2 + 2*SD^2}$
0.00	1.0	.300	3.00	"3.5"	.5000
2.29	.95	.292	.89	22	.0800
3.33	.90	.284	.62	42	.0414
5.00	.80	.268	.42	91	.0192
6.55	.70	.250	.32	153	.0114
10.00	.50	.212	.21	352	.0050
17.30	.25	.151	.12	1046	.0017

* Assumes σ_z =10, σ_{trend} =1, Δ= 3, and σ_C = 1.

Table 5
CRT Sample sizes needed per cluster to achieve 80% power by ICC, Reliability, and C (# clusters/tx group)* **

ICC	Reliability \ C (#Clusters/tx grp)	C=2	C=5	C=10	C=15	C=20
.01	.25				310	124
	.50		1156	65	34	23
	.70		65	20	12	9
	.80		30	11	7	5
	.90	297	12	5	3	2
	.95	59	6	2	1	1
.025	.25					
	.50				70	35
	.70			34	15	10
	.80		57	15	8	5
	.90		14	6	3	2
	.95	668	6	3	1	1
.05	.25					
	.50					605
	.70			154	25	14
	.80			21	10	6
	.90		24	6	3	2
	.95		7	3	2	1

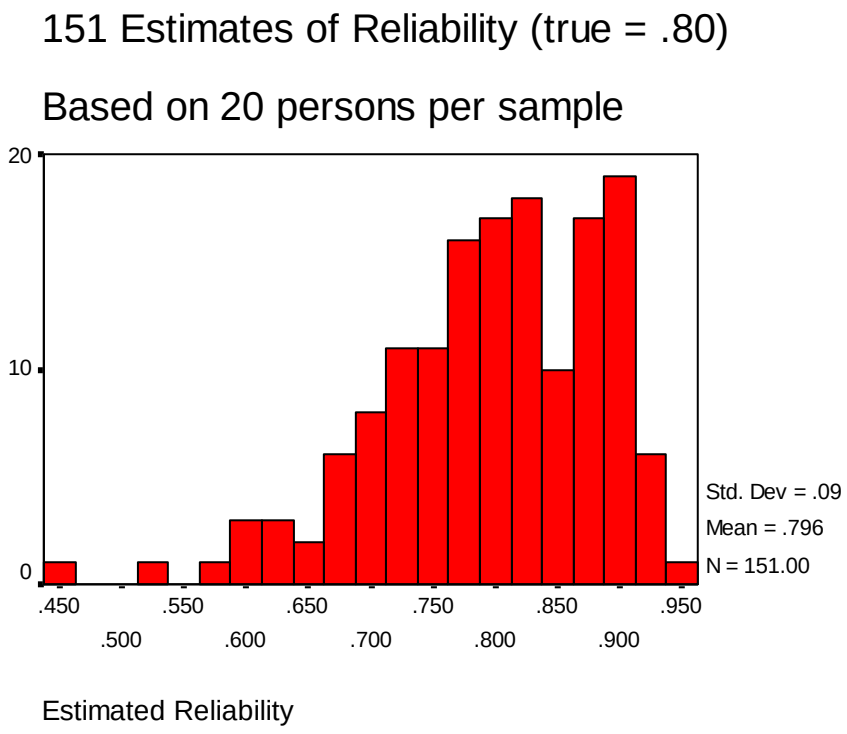
* Note: σ^2_C is different for each line; σ^2_C = ICC/(1-ICC)*[1 + 200*(1-Reliability)/Reliability], assuming σ_z =10 and σ_{trend}=1.

**A blank cell indicates that it is not possible to achieve 80% power with the specified configuration.

Table 6
Descriptive and Psychometric Statistics from the Lido Study

	5-Item Score (MHI-5)	**Single Item Score**
baseline mean	43.3	1.23
baseline s.d.	18.4	.42
Corr (baseline, 6 weeks)	.51	.38
Reliability	.48	.37
Effect Size	1.06	.83
Responsiveness	.94	.74
Cost (length)	6	2
Sample Size Needed (N)	18	29
Total cost ~ Cost * N	108	58
σ^2_C of change among communities	13.99	.0037
ICC of change	.025	.012
ICC of baseline Y	.017	.066
ICC of 9 month Y	.042	.061

Figure 1
Distribution of Reliability Estimates, N=20, True Reliability = .80



References

1. Nunnally JC. Psychometric theory. 1967. McGraw-Hill Book Company. New York.

2. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Second edition. 1995. Oxford University Press, New York.

3. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research 2002; 11:193-205.

4. Wagner EH, Wickizer TM, Cheadle A, Psaty BM, Koepsell TD, Diehr P, Curry SJ, Von Korff M, Anderman C, Beery WL, Pearson DC, Perrin EB**.** The Kaiser Family Foundation Community Health Promotion Grants Program: findings from an outcome evaluation. Health Serv Res. 2000 Aug;35(3):561-89.

5. Beresford SA, Thompson B, Feng Z, Christianson A, McLerran D, Patrick DL. Seattle 5 a Day worksite program to increase fruit and vegetable consumption. Prev Med 2001; 32:230-238.

6. Green SB, Carle DK, Gail MH, Mark SD, Pee D, Freedman LS, Graubard BI, Lynn WR. Interplay between design and analysis for behavioral intervention trials with the community as the unit of randomization. Am J Epidemiol. 1995; 142:587-593.

7. Koepsell T, Martin D, Diehr P, Psaty B, Wagner E, Perrin E, Cheadle A: Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: A mixed-model analysis of variance approach. *Journal of Clinical Epidemiology* 44:701-713, 1991.

8. Koepsell T, Wagner E, Cheadle A, Patrick D, Kristal A, Allan-Andrilla CH, Dey L, Martin DC, Diehr P. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. Annual Review of Public Health 1992. 13:31-57

9. Koepsell T, Diehr P, Cheadle A, Kristal A. Commentary: Symposium on Community Preventive trials. Amer J of Epidemiol 142:594-599. 1995.

10. Feng Z, Diehr P, Yasui Y, Evans B, Koepsell TD. Explaining community-level variance in group randomized trials. Statistics in Medicine. 18:539-556. 1999.

11. Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. Annual Review of Public Health 2001. 22:167-187.

12. Scientific Advisory Committee, Medical Outcomes Trust. Assessing health status and quality of life instruments: Attributes and review criteria. Quality of Life Research 2002;11(3):193-205

13. Patrick DL, Erickson P. Health status and health policy: allocating resources to health care. 1993. Oxford University Press. New York.

14. Fairclough DL. Deesign and analysis of quality of life studies in clinical trials. 2002. Chapman and Hall. New York.

15. Fayers PM, Machin D. Quality of life: assessment, analysis, and interpretation. 2000. John Wiley and Sons, West Sussex, England.

16. Staquet MJ, Hays RD, Fayers PM. Quality of life assessment in clinical trials: methods and practice. 1998. Oxford University Press. New York.

17. Deyo R, Diehr P, Patrick D: Reproducibility and Responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clinical Trials 12:142S-158S,* 1991.

18. Duan N. Smearing estimate: a nonparametric retransformation. J. Am. Stat. Assoc. 1983. 78:605-610.

19. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40:171-178.

20. Kristal AR, Beresford SA, Lazovich D. Assessing change in diet-intervention research. Am J Clin Nutr 1994;59 (suppl):185S-189S.

21. Cohen J. Statistical power analysis for the behavioral sciences. Second ediдtion. Lawrence Erlbaum Associates. 1988. Hillsdale, New Jersey.

22. Donner A, Klar A. Design and analysis of cluster randomization trials in health research. Arnold. London. 2000.

23. Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials* 1992; **13**:6-21.

24. Murray DM. Design and analysis of group-randomized trials. Oxford University Press New York. 1998.

25. Feng Z, Grizzle J. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculations. Statistics in Medicine 1992; 11:1607-1614.

26 Dunn OJ, Clark V. Applied statistics: analysis of variance and regression. 1974. John Wiley and Sons, New York.

27. Murray DM. Design and analysis of group-randomized trials. Oxford University Press New York. 1998.

28. Campbell K, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. Statistics in Medicine 2001; 20:391-399.

29. Moinpour CM. Costs of quality-of-llife research in Southwest Oncology Group trials. J Natl Cancer Inst Monogr. 1996;20:11-6.

30. Diehr P, Patrick DL, Burke G, Williamson J. Survival versus years of healthy life: which is more powerful as a study outcome? Controlled Clinical Trials. 1999. 20:267-279.

31. Yusuf S, Collins R, Peto R. Why do we need some large, simple trials? Statistics in Medicine 1984; 3:409-422.

32. Herrman H, Patrick DL, Diehr P, Martin M, Fleck M, Simon G, Buesching D, the LIDO group. Longitudinal investigation of depression outcomes in primary care in six countries: the LIDO study. Functional status, health service use and treatment of people with depressive symptoms. Psychological Medicine, 2002. 32:889-902.

33. Kessler RC, Andrews G, Mroczek D, Ustun B, Wittchen HU. The World Health Organization Composite International Diagnostic Interview Short-Form (CIDI-SF). *Int J Methods Psychiatr Res* 1998;7:171-85.