February, 2005

Reliability, effect size, and responsiveness of health status measures in the design of randomized and cluster-randomized trials

Paula Diehr, *University of Washington*

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Contemporary Clinical Trials 26 (2005) 45–58

Contemporary Clinical Trials

www.elsevier.com/locate/conclintrial

Reliability, effect size, and responsiveness of health status measures in the design of randomized and cluster-randomized trials

Paula Diehr[a,b,*], Lu Chen[a], Donald Patrick[b], Ziding Feng[a,c], Yutaka Yasui[d]

[a]*Department of Biostatistics University of Washington, Box 357232, Seattle, WA 98195, United States*
[b]*Department of Health Services University of Washington, Seattle, WA 98195, United States*
[c]*Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, United States*
[d]*Department of Public Health Sciences, Faculty of Medicine and Dentistry, University of Alberta, Edmonton Alberta, Canada T6G 2G3*

Received 17 June 2004; accepted 16 November 2004

Abstract

Background: New health status survey instruments are often described by their psychometric (measurement) properties, such as Validity, Reliability, Effect Size, and Responsiveness. For cluster-randomized trials, another important statistic is the Intraclass Correlation (ICC) for the instrument within clusters. Studies using better instruments can be performed with smaller sample sizes, but better instruments may be more expensive in terms of dollars, opportunity cost, or poorer data quality due to the response burden of longer instruments.
Methods: We defined the psychometric statistics in terms of a mathematical model, and examined the power of a two-sample test as a function of the test–retest Reliability, Effect Size, Responsiveness, and Intraclass Correlation of the instrument. We examined the "cost-effectiveness" of using a one-item versus a five-item measure of mental health status.
Findings: Under the standard model for measurement error, the psychometric statistics are all functions of the same error term. They are also functions of the setting in which they were estimated. In randomized trials, power is a function of Reliability and sample size, and a less reliable instrument can achieve the desired power if N is increased. In cluster-randomized trials, adequate power may be obtained by increasing the number of clusters per

DOI of original article:10.1016/j.cct.2004.11.006.
* Corresponding author. Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, United States. Tel.: +1 206 543 8004; fax: +1 206 543 3286.
E-mail address: pdiehr@u.washington.edu (P. Diehr).

1551-7144/$ - see front matter
doi:10.1016/j.cct.2004.11.014

treatment group (and often the number of persons per cluster), as well as by choosing a more reliable instrument. The one-item measure of mental health status may be more cost-effective than the five-item measure in some situations.
Conclusion: If the goal is to diagnose or refer individual patients, an instrument with high Validity and Reliability is needed. In settings where the sample sizes are large or can be increased easily, any valid instrument may be cost-effective. It is likely that many published values of psychometric statistics are accurate only in settings similar to that in which they were estimated.

Keywords: Sample size; Power; Reliability; Effect Size; Responsiveness; Intraclass Correlation

1. Introduction

New survey instruments are developed frequently, along with documentation that reports their psychometric (measurement) properties, such as Validity, Reliability, Effect Size, and Responsiveness. (In the following, we shall sometimes refer to the latter three characteristics jointly as "Reliability", 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. The definitions of these statistics are often difficult to understand and compare.

These statistics may be used to help a clinical trial designer choose among the valid instruments available. More reliable instruments can reduce the necessary sample size, but the most reliable instrument may be too "expensive" for the intended use. The cost of using a particular instrument may be thought of in terms of time (patient or interviewer time), length (if total length of the survey is fixed, there is an opportunity cost of not measuring the other facets of health as well), dollars (costs of proprietary instruments, highly trained interviewers or interpreters), or the completeness 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 need not be the instrument with the highest Reliability. An instrument used to diagnose, treat or refer an individual should have high Reliability. Nunnally suggests a goal of 0.90 or higher [1]. However, most research involves the comparison of groups of persons, often on their average 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 0.70 has been recommended as a minimum Reliability to be used in comparing two groups [1,3,4].

Behavior change interventions are often conducted as cluster-randomized trials when the intervention naturally occurs at the cluster level; the cluster could be a clinic, a community, or a workplace [5–12]. 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 use a statistical model to define the psychometric statistics, to show how they are related to one another, and to show how they are related to the power and required sample size of a study. A simple example is presented.

2. Methods

We begin this paper with a review of several psychometric statistics, first defining 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.

3. 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 μ_z=50 and σ_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) 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 μ_{trend}=1 and standard deviation σ_{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 (Δ) 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 in the intervention group). 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).

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 describes 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 SD is assumed independent of Z. If M is zero, then Y is an unbiased estimate of Z. We will let M equal zero, without loss of generality, since the psychometric measures we will discuss all remove the mean. In the following we

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	+	$N(\mu_{\text{trend}} = 1, \sigma_{\text{trend}} = 1)$ Secular Trend		+ Δ = Treatment effect = 3 Treatment	
Y = instrument							
y_0	=	Z_0	+	ϵ ;	ϵ ~	$N(M, \text{SD})$	$M = 0$; SD = 0 – 17.3
y_1	=	Z_1	+	ϵ ;	ϵ ~	$N(M, \text{SD})$	
y_2	=	Z_2	+	ϵ ;	ϵ ~	$N(M, \text{SD})$	

Table 2
Distributions of true and measured health variables

Variable	Mean	Variance
Z_0	μ_z=50	σ_z^2=100
Z_1	μ_z=50	σ_z^2=100
Z_2	$\mu_z+\mu_{trend}$=51 (control), $\mu_z+\mu_{trend}+\Delta$=54 (treatment)	$\sigma_z^2+\sigma_{trend}^2$=101
Y_0	μ_z=50	$\sigma_z^2+SD^2$=100+SD^2
Y_1	μ_z=50	$\sigma_z^2+SD^2$=100+SD^2
Y_2	$\mu_z+\mu_{trend}$=51 (control), $\mu_z+\mu_{trend}+\Delta$=54 (treatment)	$\sigma_z^2+\sigma_{trend}^2+SD^2$=101+$SD^2$
Y_2-Y_1	μ_{trend}=1 (control), $\mu_{trend}+\Delta$=4 (treatment)	σ_{trend}^2+2*SD^2=1+2*SD^2

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

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.

3.1. 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_{yz}^2 is the proportion of variation in Y that is explained by Z. From Tables 1 and 2, it is clear that this proportion is $\sigma_z^2/(\sigma_z^2+SD^2)$=100/(100+$SD^2$). If SD is 2.29, then R_{yz}^2=0.95, and its square root is r_{yz}=0.975. The correlation between Y_0 and Y_1 (r_{yy}) can be thought about in two parts. Y_0 explains R_{yz}^2 of the variability in Z_0 which also, because Z_0=Z_1, explains R_{yz}^2 of the variability in Y_1. The percent of variability in Y_1 explained by Y_0 is then the product of these two proportions, or $(R_{yz}^2)^2$, and r_{yy}=R_{yz}^2. These correlations are shown in Table 3, for several values of SD. Correlations all decrease as SD increases.

Table 3
Correlations among true and measured health variables

SD	R_{yz}^2	r_{yz}	R_{yy}^2	R_{yy}
2.29	0.950	0.975	0.902	0.950
3.33	0.900	0.949	0.810	0.900
5.00	0.800	0.894	0.640	0.800
6.55	0.700	0.837	0.490	0.700
10.00	0.500	0.707	0.250	0.500
17.30	0.250	0.500	0.063	0.250

R_{yz}^2 is the correlation between the true (Z) and measured (Y) values at T_0.
R_{yy}^2 is the correlation between Y_0 and Y_1.

4. Psychometric characteristics of the instrument *Y*

Four commonly cited properties of an instrument are its Validity, test–retest Reliability, Effect Size, and Responsiveness [2–4,13–16]. These will be discussed in turn. Another property, Intraclass Correlation within clusters, is discussed in the section on cluster-randomized trials.

4.1. Validity

Validity has to do with whether an instrument measures what it is supposed to measure. Validity is demonstrated in several ways, such as whether the items on the instrument appear to be appropriate (face validity), agreement with a "gold standard" (criterion validity), or correlation in the expected direction with other measures (construct validity). In our example, *Y* has criterion validity because it is correlated with the "gold standard", *Z*. Table 3 shows that the correlation, r_{yz}, and thus the criterion Validity, becomes lower when SD is large. We know of no minimum acceptable level for criterion Validity, probably because a gold standard is almost never available. We will not discuss Validity further, but will assume that investigators will consider only valid measures.

4.2. Reliability

Reliability is a measure of whether the same person, under the same conditions, would give the same response. (Unfortunately, the internal consistency of an instrument, as measured by Cronbach's alpha statistic, is also referred to by some as reliability. We do not use reliability in that sense.) Reliability 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 does not 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 $\mathrm{Var}(Z_0)/\mathrm{Var}(Y_0)=R^2_{yz}=\sigma_z^2/(\sigma_z^2+\mathrm{SD}^2)=100/(100+\mathrm{SD}^2)$, but ordinarily Reliability must be estimated, traditionally from an analysis of variance table [17,18]. In a two-way random model, if there are *N* persons and *K* occasions (K=2 in a test–retest situation), and the relevant mean square errors are calculated for the effect of person (MSP), occasion (MSO), and residual (MSE), then the ratio of the variance among persons divided by total variance is estimated as:

$$\text{Reliability} = \frac{\text{MSP} - \text{MSE}}{\text{MSP} + \text{MSE}(K-1) + K(\text{MSO} - \text{MSE})/N}$$

When K=2, the estimate reduces to a simpler formula, all the terms of which are conveniently available from the output of a paired *t*-test of the test–retest data [19]. In our notation, Reliability would be estimated from Y_0 and Y_1, as follows:

$$\text{Reliability} = \frac{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_0} - S^2_{(Y_1-Y_0)}}{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_0} - \left[S^2_{(Y_1-Y_0)}/N\right] + \left(\bar{Y}_1 - \bar{Y}_0\right)^2}$$

Table 4
True psychometric characteristics of the instrument Y as a function of measurement error (SD)

SD	Reliability $\frac{\sigma_z^2}{\sigma_z^2 + \mathrm{SD}^2}$	Effect Size $\frac{\Delta}{\sqrt{\sigma_z^2 + \mathrm{SD}^2}}$	Responsiveness $\frac{\Delta}{\sqrt{\sigma_{\mathrm{trend}}^2 + 2\mathrm{SD}^2}}$	N per group for 80% power $\frac{2(1.96 + .84)^2}{\mathrm{Responsiveness}^2}$	ICC for change from T_1 to T_2 $\frac{\sigma_c^2}{\sigma_c^2 + \sigma_{\mathrm{trend}}^2 + 2\mathrm{SD}^2}$
0.00	1.0	0.300	3.00	"3.5"	0.5000
2.29	0.95	0.292	0.89	22	0.0800
3.33	0.90	0.284	0.62	42	0.0414
5.00	0.80	0.268	0.42	91	0.0192
6.55	0.70	0.250	0.32	153	0.0114
10.00	0.50	0.212	0.21	352	0.0050
17.30	0.25	0.151	0.12	1046	0.0017

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

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. Note that Reliability in Table 4 is the same as R_{yz}^2 in Table 3, and that we had suggested that R_{yz}^2 was a measure of criterion Validity. In this sense, Validity and Reliability are the same thing when a criterion or gold standard is available (i.e., when Z is known, or when Y is known to be valid).

4.3. Delta (Δ)

The next two psychometric measures, Effect Size and Responsiveness, require specification 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 [13]. 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 causes an increase 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 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 [19]. 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 clearly important to specify the source of the Δ used in calculations.

4.4. 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 pre-intervention value for a particular instrument, with a particular value of SD. Under our model this would be:

$$\mathrm{ES} = \frac{\Delta}{\sigma_{Y_1}} = \frac{\Delta}{\sqrt{\sigma_z^2 + \mathrm{SD}^2}} = \frac{3}{\sqrt{100 + \mathrm{SD}^2}}$$

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.

4.5. 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 [13]. Under our model, the stable subjects are the controls, and

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

Values of Responsiveness for different values of SD are in Table 4. The highest possible Responsiveness, when SD=0, is Δ=3.0.

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=$2(1.96+0.84)^2$/Responsiveness^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 it was based on the normal approximation rather than the t-statistic.) Instruments with a larger SD require larger sample sizes to achieve 80% power.

4.6. 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 [8–10,22,23]. 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 difference over time $D_k=(\bar{Y}_{2k}-\bar{Y}_{1k})$, calculated for each cluster k. The $2C$ cluster means (C for treatment and C for control), will then be analyzed using a t-test. (Randomization tests [24] and multi-leveling model are alternate 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 [25, p. 7], where our clusters are his groups and our treatment groups are his study conditions). The Intraclass Correlation

(within a treatment group) is the variance among clusters divided by that variance plus the variation among people within clusters:

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

$\mathrm{ICC_{CRT}}$ (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 σ_C^2=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. Murray [25, p. 81] suggests that an ICC value of 0.02 is typical in large community CRTs. Campbell et al. report ICCs for cost data as large as 0.47 [26].

Feng and Grizzle [27] 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. Since we wish to consider smaller numbers of clusters, we present different calculations here. In the following, we selected values of σ_C^2 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 ($\sigma_{\mathrm{trend}}^2$+2SD^2) divided by the number of persons (NC) plus the variance among clusters (σ_C^2) divided by the number of clusters (C) [28].

$$\mathrm{Var}\left(\overline{D}\right) = \frac{\sigma_C^2}{C} + \frac{\sigma_{\mathrm{trend}}^2 + 2\mathrm{SD}^2}{\mathrm{NC}}$$

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 $1-\beta$, 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

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

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 $\bar{D}$, the necessary number of clusters (C) for a fixed value of N is:

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

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).

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

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=0.01, Reliability of the instrument=0.25, and C=20 clusters per treatment group, then a study with 124 persons per cluster will yield 80% power with alpha=0.05, and only 23 persons per cluster are needed if the Reliability is 0.50. Table 5 shows that higher Reliability, more clusters, and lower ICC are all associated

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 group)	C=2	C=5	C=10	C=15	C=20
0.01	0.25				310	124
	0.50		1156	65	34	23
	0.70		65	20	12	9
	0.80		30	11	7	5
	0.90	297	12	5	3	2
	0.95	59	6	2	1	1
0.025	0.25					
	0.50				70	35
	0.70			34	15	10
	0.80		57	15	8	5
	0.90		14	6	3	2
	0.95	668	6	3	1	1
0.05	0.25					
	0.50					605
	0.70			154	25	14
	0.80			21	10	6
	0.90		24	6	3	2
	0.95		7	3	2	1

σ_C^2 is different for each line; σ_C^2=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.

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 0.01 and Reliability is 0.95, 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.

5. 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 other 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 achieve the desired power through an increase in sample size. There are many situations in which the sample size is fixed based on some other criterion. For example, many studies are powered to detect mortality differences, which usually provides more subjects than needed to detect differences in quality of life [22]. Large, simple trials are designed especially to study very large numbers of persons, and their success depends on using extremely simple data collection instruments [30]. A study's sponsor may require inclusion of all people in a certain class, such as all primary care patients in a clinic. In those situations, the most reliable instrument might not be needed. For example, if the sample size was fixed at 352 per group for some reason, Table 4 shows that an instrument with Reliability of only 0.50 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.

6. Example: data from the LIDO study

The Longitudinal Investigation of Depressive Outcomes (LIDO) study was an observational study of depression in six international cities [31]. Primary care patients who met eligibility criteria were assessed for depression using the Composite International Diagnostic Interview (CIDI) [32]. There were 981 persons who had clinical depression at baseline and a valid CIDI assessment 9 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 six 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

Table 6
Descriptive and psychometric statistics from the LIDO study

	Five-item score (MHI-5)	Single-item score
Baseline mean	43.3	1.23
Baseline SD	18.4	0.42
Correlation (baseline, 6 weeks)	0.51	0.38
Reliability	0.48	0.37
Effect Size	1.06	0.83
Responsiveness	0.94	0.74
Cost (length)	6	2
Sample Size needed (N)	18	29
Total cost~Cost×N	108	58
σ_C^2 of change among cities	13.99	0.0037
ICC of change	0.025	0.012
ICC of baseline Y	0.017	0.066
ICC of 9 month Y	0.042	0.061

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 five-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 $2(1.96+0.84)^2/\text{Responsiveness}^2$. 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 choice would need to balance these marginal costs with the costs of obtaining an additional person and the much larger fixed costs of the study.

We also estimate σ_C^2 from the six 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 0.02, as suggested by Murray, and three were substantially higher. Because our estimates were based on only six clusters, they are not, of course, definitive.

7. Discussion

We used the standard measurement 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 found to be a function of SD, the measurement error, and so behaved in similar ways. We noted that Reliability can be considered a type of criterion Validity, implying that a less reliable instrument is also less valid. Since this relationship holds only when Y is known to be a valid instrument, however, its practical significance is unclear. We also noted some inconsistencies in how parameters were defined (particularly

Δ, 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 Δ.

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_z^2), as is well known. If σ_z^2 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 a property not only of the instrument, but also of the population in which the Reliability is estimated. 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_z^2, 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 will clearly be more valuable when calculated in a similar context to the new planned investigation.

Valid instruments should be chosen to meet the purposes of the investigation. This does not always mean choosing the most reliable 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 valid instruments under consideration could achieved the desired power if 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. One might also choose a less reliable instrument if it was the standard instrument in the particular field of investigation. Again, this would probably require an increase in sample size to account for the lower Reliability.

The LIDO example suggested that a single item might sometimes be chosen in preference to a five-item question on the basis of cost-effectiveness if the costs of finding additional subjects are not high. We doubt anyone would actually choose the one-item measure based on this analysis, because the "cost" of the five-item scale is likely to be a tiny fraction of the research budget, and also because a single item might not be considered a valid measure of mental health status. We rather intended the example to show how such a decision might be reached in a more critical situation. For example, we would like to have compared the cost of using the resource-intensive CIDI, which had to be administered in person twice and scored by trained professionals, to a valid self-administered depression scale. We could not do this with the LIDO data because the CIDI was needed to define the outcomes. A general cost analysis is beyond the scope of this paper, but would require specification of fixed and variable costs, which would include the cost of patient recruitment as well as the cost of the interview, of processing the data, and of non-monetary factors such as subject burden and resulting data quality. Use of an instrument that was not used by other investigators would also have non-monetary costs in terms of lack of comparability with other studies.

Recent published systematic reviews of instruments have the laudable aim of encouraging investigators to use the best instruments. Such reviews do not usually differentiate between instruments to be used for diagnosis and treatment, which must be highly reliable, and those to be used in randomized trials, where lower Reliability may be acceptable. An investigator could use such a review to

identify the most valid instruments, and then choose the one with Reliability/cost most suitable to the study needs.

7.1. Limitations

These findings should not depend very much on the exact form of the model we 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 variations are likely to have affected the results appreciably. We let Y be an unbiased estimate of Z. The findings also apply if $E(Y)=a+bZ$, since a and b cancel out 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 measurement error (SD).

7.2. Conclusion

The psychometric qualities of survey instruments are often used to select the best instrument for a randomized trial. It is important to estimate these psychometric characteristics in a variety of settings to yield estimates that are appropriate for different types of patients and settings. In selecting an instrument, an investigator should distinguish between the highly reliable instruments needed at the patient level, and the valid but less reliable instruments that may be appropriate and "cost-effective" for the research study at hand.

Acknowledgement

Supported in part by NIH grant CA84079. The sponsor of the LIDO project was Eli Lilly and Company, Indianapolis, IN, USA.

References

[1] Nunnally JC. Psychometric theory. New York: McGraw-Hill Book; 1967.
[2] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Second edition. New York: Oxford University Press; 1995.
[3] Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
[4] Patrick DL, Erickson P. Health status and health policy: allocating resources to health care. New York: Oxford University Press; 1993.
[5] Wagner EH, Wickizer TM, Cheadle A, Psaty BM, Koepsell TD, Diehr P, et al. The Kaiser Family Foundation Community Health Promotion Grants Program: findings from an outcome evaluation. Health Serv Res 2000 (Aug);35(3):561–89.
[6] 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;32:230–8.
[7] Green SB, Carle DK, Gail MH, Mark SD, Pee D, Freedman LS, et al. Interplay between design and analysis for behavioral intervention trials with the community as the unit of randomization. Am J Epidemiol 1995;142:587–93.
[8] Koepsell T, Martin D, Diehr P, Psaty B, Wagner E, Perrin E, et al. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. J Clin Epidemiol 1991;44:701–13.

[9] Koepsell T, Wagner E, Cheadle A, Patrick D, Kristal A, Allan-Andrilla CH, et al. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. Annu Rev Public Health 1992;13:31–57.
[10] Koepsell T, Diehr P, Cheadle A, Kristal A. Commentary: symposium on community preventive trials. Am J Epidemiol 1995;142:594–9.
[11] Feng Z, Diehr P, Yasui Y, Evans B, Koepsell TD. Explaining community-level variance in group randomized trials. Stat Med 1999;18:539–56.
[12] Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. Annu Rev Public Health 2001;22:167–87.
[13] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171–8.
[14] Fairclough DL. Design and analysis of quality of life studies in clinical trials. New York: Chapman and Hall; 2002.
[15] Fayers PM, Machin D. Quality of life: assessment, analysis, and interpretation. West Sussex (England): John Wiley and Sons; 2000.
[16] Staquet MJ, Hays RD, Fayers PM. Quality of life assessment in clinical trials: methods and practice. New York: Oxford University Press; 1998.
[17] Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3–11.
[18] Fleiss JL. Reliability of measurement. In: Fleiss JL, editor. The design and analysis of clinical experiments. New York: Wiley; 1986. p. 1–32.
[19] Deyo R, Diehr P, Patrick D. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12:142S–58S.
[20] Kristal AR, Beresford SA, Lazovich D. Assessing change in diet-intervention research. Am J Clin Nutr 1994;59:185S–9S [Suppl.].
[21] Cohen J. Statistical power analysis for the behavioral sciences. Second edition . Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
[22] Diehr P, Patrick DL, Burke G, Williamson J. Survival versus years of healthy life: which is more powerful as a study outcome? Control Clin Trials 1999;20:267–79.
[23] Donner A, Klar A. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
[24] Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). Control Clin Trials 1992;13:6–21.
[25] Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
[26] Campbell K, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. Stat Med 2001;20:391–9.
[27] Feng Z, Grizzle J. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculations. Stat Med 1992;11:1607–14.
[28] Dunn OJ, Clark V. Applied statistics: analysis of variance and regression. New York: John Wiley and Sons; 1974.
[29] Moinpour CM. Costs of quality-of-life research in southwest oncology group trials. J Natl Cancer Inst Monogr 1996;20:11–6.
[30] Yusuf S, Collins R, Peto R. Why do we need some large, simple trials? Stat Med 1984;3:409–22.
[31] Herrman H, Patrick DL, Diehr P, Martin M, Fleck M, Simon G, et al., 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. Psychol Med 2002;32:889–902.
[32] 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.