# Reliability vs. Power

Kimberly A. Barchard
University of Nevada, Las Vegas

## Abstract

If we want to make relative comparisons between individuals, the Reliability Coefficient should be high. However, in many circumstances, statistical power is more important. For example, we may want to compare a treatment group with a control group. To maximize statistical power, we must minimize within group variance. To do this, we must recruit participants for each group who are very similar to each other and treat them all the same. Ideally, the only differences between people who are in the same group will be due to random error. Minimizing within group variance will result in low reliability coefficients in each of the two groups. Thus, reliability and power are in conflict. Psychologists must recognize when they want to distinguish between individuals (and thus want to maximize reliability) and when they want to distinguish between groups (and thus want to maximize power). When we want power, we must use test design methods that will maximize power (e.g., design tests with high content validity, and select items based upon their ability to discriminate between groups) rather than reliability (e.g., using corrected item-total correlations and factor analysis), and we must use test evaluation methods that focus on power (e.g., did this test result in significant results in past studies?) rather than reliability (e.g., was this test reliable in past studies?). The use of reliability-maximizing statistics (e.g., factor analysis) during test design sacrifices the power of future studies by throwing out content areas with little variability in the derivation sample.

## Reliability

The American Psychological Association's Code of Conduct states "Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested" (2002, section 9.02.b). This is bad advice. Validity is good, but reliability is only sometimes good. Sometimes power is more important, and if we focus on high reliability we will unintentionally reduce power.

The Reliability Coefficient is defined as the ratio of True Score variance to Observed Score variance (Allen & Yen, 1979; Lord & Novick, 1968; see Figure 1). For this ratio to be high, we need high variability in True Scores (so that people are different from each other) and low variability in Error Scores (so that Observed Scores for each person are similar to each other). When the Reliability Coefficient is high, test scores allow us to distinguish between test takers. However, as psychologists, this is not always our primary goal. Quite often, statistical power is more important.

## Power

Power is the probability that we will reject a false null hypothesis. The null hypothesis always says that nothing interesting happened: For example, this one group is average, these two groups are the same, or these variables have no relationship. We usually think that the null hypothesis is false, and so we want high statistical power.

The formulas for power depend upon what hypothesis testing procedure we are using. In this paper, I will discuss how to maximize power for the independent sample t-test, which is used to compare the means from two groups. These might be pre-existing groups (e.g., men and women; bipolar and major depression) or groups that were created by the researchers (e.g., treatment and control; stimuli 1 and 2). I will discuss the independent sample t-test because most readers will have seen the formulas before. Many of the techniques that increase power for the independent sample t-test will also increase power for other statistical techniques. See Cohen (1988) for a discussion of power for additional hypothesis testing techniques.

There are five ways to increase power in an independent-sample t-test. Three ways to increase power are unrelated to reliability: We can use a one-tailed test rather than a two-tailed test, increase alpha, and increase sample size. However, one-tailed tests are frowned upon and alpha is limited to .05. Thus, the only way to increase power without reducing reliability is to increase sample size. The remaining two ways to increase power are in direct conflict with reliability. I will discuss these in more detail.

The first way is to increase $\bar{X}_1 - \bar{X}_2$, the observed difference in the means. Figure 2 shows a comparison of Group A and Group B. In the left-hand diagrams (2a and 2c), the difference in group means is small. In the right-hand diagrams (2b and 2d), it is large. If we increase the difference in the means, then the two groups are further apart vertically and we obtain more power.

There are two methods to increase the difference in group means. One method is to study groups who are very different from each other on the dimension of interest. For example, if we are interested in individualism and collectivism, we will have more power if we compare Americans and Japanese than if we compare Americans and Canadians (Hofstede, 2001). The other method is to treat the two groups very differently. For example, if we want to compare a treatment for depression against a control group that gets no treatment, we will have more power if the treatment group gets six months of therapy than if it gets two weeks.

**Figure 1**
*Classical Test Theory*

The Observed Score

$$X_{ij} = T_i + E_{ij}$$

where $X_{ij}$ = the Observed Score for person i on measurement j
$T_i$ = the True Score for person i, and
$E_{ij}$ = the Error Score for person i on measurement j.

The Reliability Coefficient

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where $\sigma_X^2$ = the variance of Observed Scores across all test takers,
$\sigma_T^2$ = the variance of True Scores across all test takers, and
$\sigma_E^2$ = the variance of Error Scores across all test takers.

However, it is not enough to study two groups that are different from each other. We must also focus our measurement on the areas in which they differ. This is one of the key differences between reliability and power. To have a high Reliability Coefficient, we must measure a dimension where there are large differences between the individuals we are studying. Any dimension that has a lot of variability can have a high Reliability Coefficient. However, to have high power, we must measure a dimension where the groups differ.

The other way to increase power is to reduce the variance in each group. In Figure 2, the bottom graphs (2c and 2d) have smaller within group variances than the top graphs (2a and 2b). When we decrease the within group variances, it is easier to distinguish between Groups A and B because there is less overlap in the scores.

There are two things we need to do to make the within-group variance small. First, we need to recruit people for each group who are very similar to each other on the dimension of interest. If we are studying depression, for example, everyone should have equal levels of depression. Second, within each group, we need to treat everyone the same, so that we do not introduce differences between them. This is referred to as standardizing our experimental conditions. For example, if we are studying depression, everyone should receive the same kind of treatment.

This is the second difference between reliability and power. To obtain a high Reliability Coefficient in Group A, we need high True Score variance. However, if we want high power when we compare groups, we want True Score variance to be low. Ideally, everyone who is in the same treatment would have an identical (or nearly identical) True Score and an Error Score that is zero (or close to zero), so that the Observed Scores are all identical (or nearly identical). If we calculate the Reliability Coefficient in each group, it will be very low.

It is possible to design (or select) a test with high reliability that will also have high power. However, we do not want reliability within the groups we are comparing. Instead, we want reliability in the *general population*, where we expect a lot of variability on the construct of interest. In a study that compares groups who are different from each other, we want people within the groups to be homogeneous. Do not calculate reliability within those groups.

## Sacrificing Power for Reliability

Consider a treatment for depression. Depression includes affective, somatic, and behavioral symptoms. In the general population, a measure of depression will have a high Reliability Coefficient if it can distinguish people who are moderately depressed from people who are not depressed. To maximize the Reliability Coefficient, test designers should use many items that focus on one sub-part of depression. In contrast, if we want to compare two treatments for depression, we want a test that will capture the largest differences between those two treatments. If the primary difference in outcomes is that one treatment is more effective in treating behavioral symptoms (e.g., overeating, excessive sleeping), then our study will have the greatest power if we focus our measurement on that specific area. If we are not sure where the treatments will differ or if we are interested in all treatment differences, we should compare the treatments on all aspects of depression.

**Figure 2**
*How to Increase Power When Comparing Two Groups*

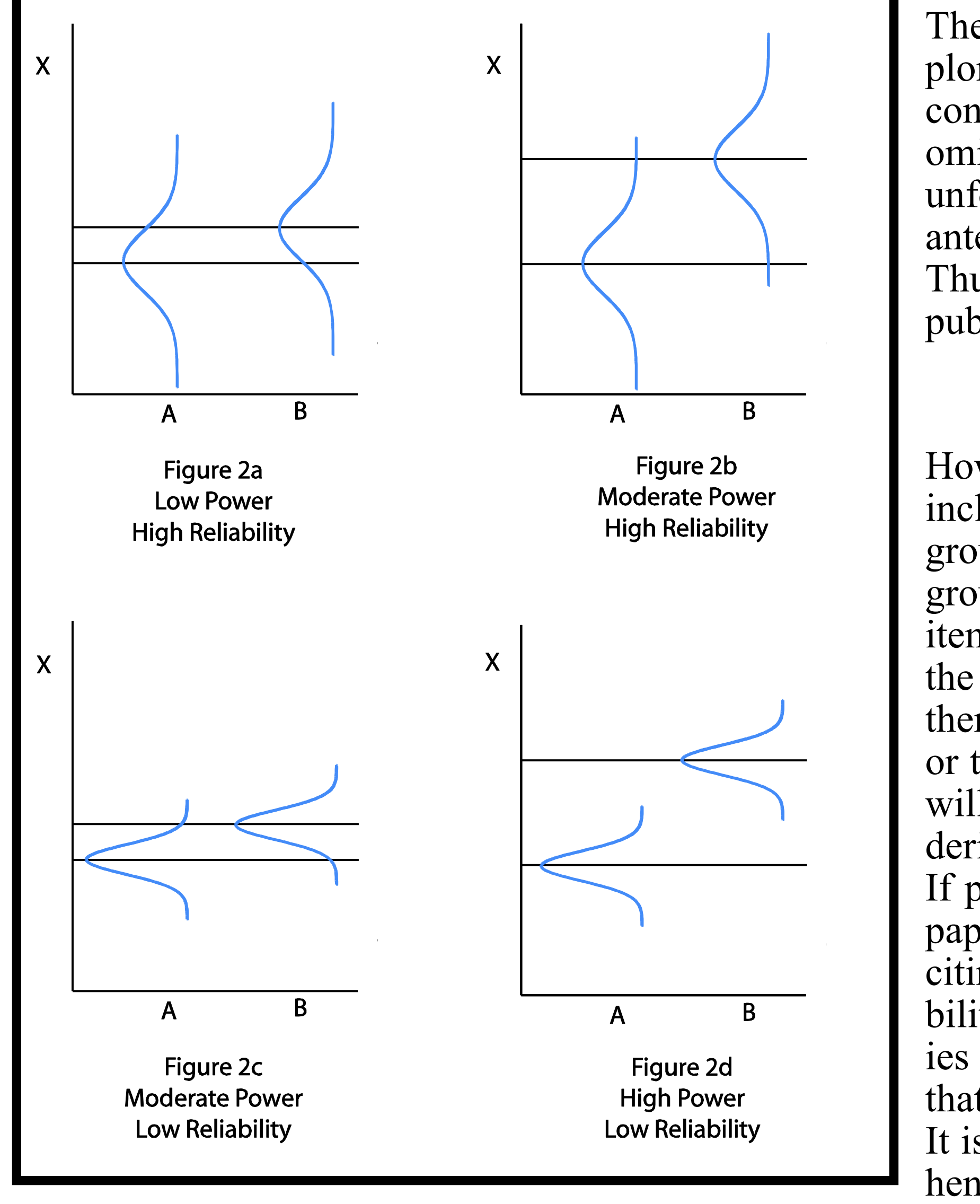

Figure 2a
Low Power
High Reliability

Figure 2b
Moderate Power
High Reliability

Figure 2c
Moderate Power
Low Reliability

Figure 2d
High Power
Low Reliability

Imagine that there is a treatment that reduces depression. When we measure depression well, the treatment has a very large effect (d = 1.5). The measure of depression has almost no variability at pretest or at post-test – everyone is very depressed before treatment and no one is depressed afterwards. Because depression has low variability at each time, the scores have low reliability. At pretest, the reliability is .03 and at posttest it is .04. The treatment is universally successful, so that every participant has their scores go from nearly the bottom of the scale to nearly the top of the scale. Because the treatment is universally and equally effective, the reliability of the changes score is very low, .02. Now imagine that someone thought that reliability was more important than power and required us to use a reliable test. We use a test of depression that shows a lot of variability at both pretest and posttest, so that not everyone is really depressed at the beginning and not everyone got better at the end. At pretest, the reliability is .75 and at posttest it is .80. We require change scores to have reliability, so that the treatment works better for some people than others (reliability of change scores = .65). The effect of the treatment decreases substantially; now it only has a moderate effect (d = .25). Our power is greatly reduced, so that we need a very large sample to detect this effect.

Does this ever happen? Would a researcher require a test to demonstrate a high Reliability Coefficient within a group which – theoretically – should demonstrate little variability on the construct? No. Although many researchers believe that reliability is universally good and understand that reliability depends upon the population, they would not make the mistake of requiring reliability *within* groups that they think differ on the construct of interest. Would a researcher select a heterogeneous group of people in order to get a high Reliability Coefficient, when power (and the theoretical importance of the research) is increased by having homogeneous samples? No. Researchers know that they cannot make conclusions about a certain kind of person if their study includes a wide variety of people, and that they cannot make conclusions about a certain type of treatment or procedure unless that procedure is standardized for everyone in the group. Novice researchers might make these kinds of errors, but experienced researchers know how to design a decent study, even though they may not realize the mathematical conflict between within-group reliability and statistical power.

Experienced researchers might none-the-less sacrifice power to obtain reliability, because the most commonly used test design methods (and test evaluation methods) emphasize reliability and neglect power. Most psychological constructs are complex and can be conceptualized as having multiple inter-related subcomponents. Researchers might use tests that have high Reliability Coefficients (or high Reliability Coefficients for each subscale), even though the test does not measure *all* aspects of the construct. This probably happens all too often. Unless the construct is very narrow (e.g., typing accuracy) or the researchers provide a compelling rationale for focusing upon just some subcomponents of the construct, it is quite likely that the researchers have not included those subcomponents upon which the groups have the largest differences.

We cannot assume that our measures include *all* aspects of the construct unless (a) a comprehensive examination of the full content domain was the basis for the measure, and (b) no statistics were used to exclude content areas upon which there was little variability in the derivation sample. Unfortunately, when comprehensive examinations of the content domain have been undertaken, these kinds of statistics have almost always been used. These statistics include corrected-item-total correlations, alpha-if-item-deleted, and exploratory and confirmatory factor analysis. When we use these statistics, we will retain content areas where there is a lot of variability in the derivation sample and we will omit all content areas where there is little variability in the derivation sample. This is unfortunate, because lack of within-group variation in the derivation sample is no guarantee that there is little between-group variation for the groups we want to compare. Thus, we sacrifice the power of future studies to obtain high reliability in our initial publication of the test.

## Designing Tests for High Power

How can we design tests that will have high power? First, we must design the test to include content from the full content domain. To ensure we measure the areas where group differences are the largest, we must either design a new measure for each set of groups we want to compare or else we must measure *all* areas. Second, we can select items from a larger item pool based upon statistical power. For example, if the goal of the test is to distinguish between two groups in order to make differential diagnoses, then test items can be selected on the basis of significant discriminant function analyses or t-tests. Third, we must rigorously reject the use of any item selection procedure that will maximize reliability and will discard areas upon which there is little variation in the derivation sample.

If power is important to us, we also need to change how we evaluate tests. Empirical papers should justify the selection of measures that will be used in significance tests by citing previous studies that found significant results, not by citing evidence about reliability. For example, if a study will compare groups, the paper should cite previous studies that found significant differences on this test when comparing the kinds of groups that will be used in the new study.

It is possible to obtain both reliability and power. But to do so, we must use a comprehensive measure that taps all aspects of the construct. If the scope of that undertaking is unrealistic for a particular test design project, psychologists need to be aware of the conflict between reliability and power, and focus on the goal that is the most important to them.

## References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Prospect Heights, IL: Waveland Press.

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060-1073. Available from http://www.apa.org/ ethics/code/index.aspx

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). California: Sage Publications.

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

Presented at the April 2011 Western Psychological Association conference, Los Angeles, CA