

HOW RELIABILITY CONSTRAINS OUR ABILITY TO DETECT RELATIONSHIPS

or, Why You Should Care About Reliability

Rob MacCoun

Why should you care about measurement reliability? One reason might be the desire for precision and accuracy in your measures. In basic scientific research, engineering, medical practice, accounting, and a few other domains, precision and accuracy are highly valued attributes. I have found that policy analysts and policy makers are sometimes much more casual about precision and accuracy. Thus we have the saying "good enough for government work."

"Methodological fascists" are people who obsess over precision for its own sake, and dismiss any measures that aren't perfect. But rather than reflecting ignorance or naivete or sloppiness, casual measurement standards can indicate pragmatism. Good measurement is costly in money, labor, and time. We should only be as precise as we really need to be under our resource constraints -- measurement should be have a positive benefit-cost ratio like any other tool.

So how much should we care about reliability? Here's one reason why it is often sensible to care about reliability -- even in situations where precision in and of itself isn't all that important: *If our measures are unreliable, we increase the risk that we will underestimate (or even fail to detect) true relationships among variables; e.g., between predictors and outcomes, or between interventions and outcomes.*

The relationship between a predictor variable (also called independent variable) and an outcome variable (also called dependent variable) is called the predictive validity of the predictor. Examples might include the correlation between GRE scores and graduate school GPA, the correlation between number of training sessions and success on the job, or the correlation between number of prevention sessions and future drug use.

Reliability places an upper bound on predictive validity--i.e., it limits the maximum possible predictive validity we can observe. As a result, if one or both of the variables are unreliable, we are likely to underestimate their association. Most seriously, we may conclude there's no relationship between the variables when there is in fact a relationship. (A Type II statistical error.)

THE MATHEMATICS

Here's some mathematics that show why this occurs (e.g., Ghiselli, Campbell, & Zedeck's *Measurement Theory for the Behavioral Sciences*, 1981). First, some notation:

$r_{ox,oy}$ is the observed correlation between x and y.

$r_{tx,ty}$ is the true correlation between x and y--the "*true predictive validity*" of y.

r_{xx} is the reliability of the x scores (e.g., Cronbach's coefficient alpha)

r_{yy} is the reliability of the y scores (e.g., Cronbach's coefficient alpha)

Here's the maximum observable predictive validity as a function of true predictive validity and reliability:

$$r_{ox,oy} = r_{tx,ty} \sqrt{r_{xx} \cdot r_{yy}}$$

I find it easier just to think about one variable at a time. Let's take as an example an independent variable that is, "for all practical intents and purposes," measured with near certainty (e.g., years of imprisonment). The dependent measure might be "number of days employed in first year out of prison," something that might well be measured by the parole officer's estimate, and hence be unreliable. (The noise that reduces reliability is that some will guess high, others will guess low.) Then we can assume $r_{xx} = 1.0$ and just focus on the effects of r_{yy} . The formula can then be simplified to:

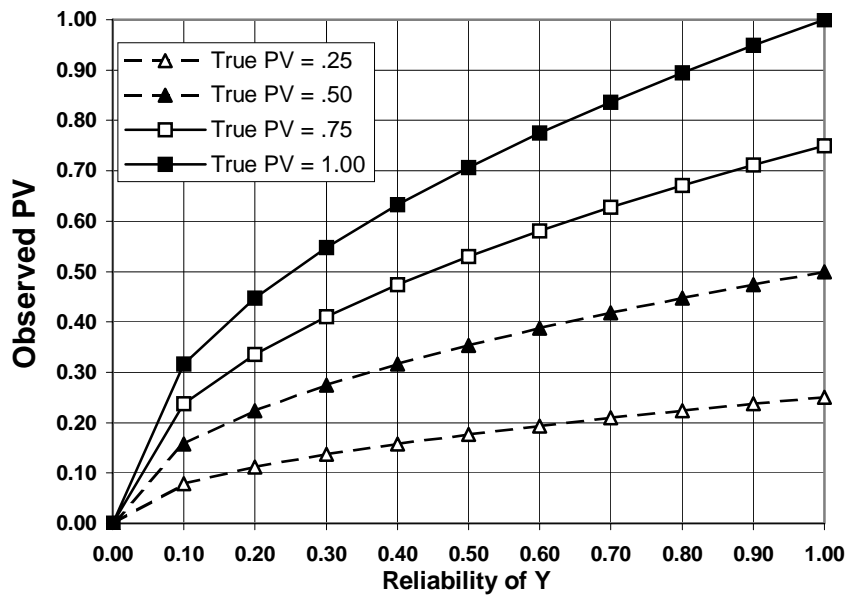
$$r_{ox,oy} = r_{tx,ty} \sqrt{r_{yy}}$$

In essence, this says that *the true predictive validity gets discounted by the square root of reliability*.¹

IMPLICATIONS

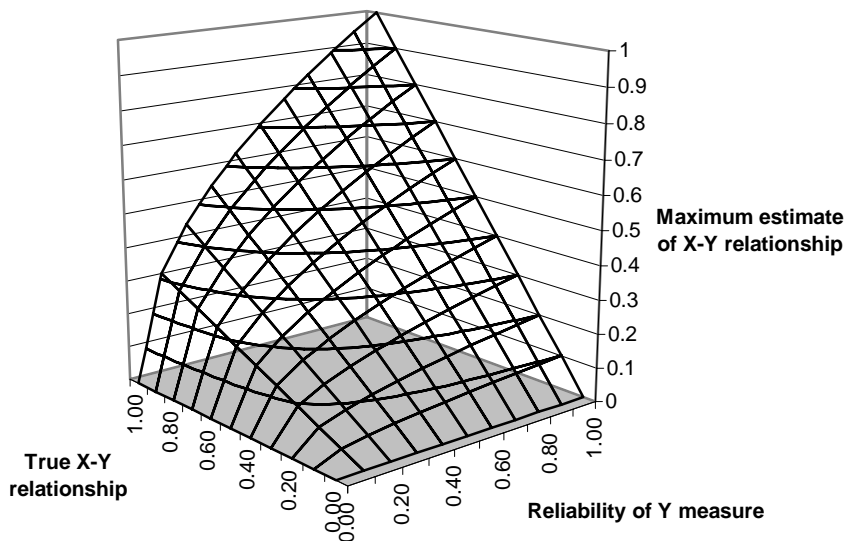
The following figure shows how well we'd measure the predictive validity of x (it's ability to predict y) as a function of the reliability of our outcome measure (the x axis), for four scenarios ranging from a true predictive validity of 0.25 (not unusual in policy research) to the maximum possible predictive validity of 1.00:

¹ Why the square root of reliability, rather than the reliability coefficient itself? It turns out that the square root of reliability is theoretically the correlation between observed and true scores, $r_{yt} = \sqrt{r_{yy}}$. So the observed predictive validity is the true predictive validity multiplied by the correlation between the y measure and what it's supposed to be measuring (Nunnally, 1978, p. 217).



The figure shows that if our outcome measure is unreliable, we will underestimate (drastically in some cases) the true x-y relationship. We are only likely to correctly estimate the true x-y relationship when we have perfectly reliable measures. Remember, the "observed PV" refers to upper bound estimates; we might even do worse than suggested by the figure.

For those who like the full story, here's a 3-D surface plot showing the full range of predictive validity possibilities (see next page).



The "right wall" of the figure shows how reliability constrains estimation of a perfect X-Y relationship. The foreground shows the effects of reliability on very weak X-Y relationships.

The "left wall" shows how reliability constrains the ability to detect a perfect X-Y relationship. A curious feature is that low reliability is much more damaging in cases where the true relationship is strong. An implication of the figure is that *low reliability might be one reason why we rarely observe strong X-Y correlations in policy research.*

A CAUTIONARY TALE

Imagine two new demonstration projects, Program A and Program B, each designed to increase English reading achievement among elementary school students from Spanish speaking homes. Program B is about 15% more expensive to implement. Your agency has funded evaluations of each program. The results are in.

The first evaluation estimated a correlation of .38 between number of Program A sessions and reading achievement scores. The second evaluation estimated a correlation of .36 between number of Program B sessions and reading achievement scores. Thus, it appears to be a tie. But recall that Program B is slightly more expensive. With a slightly stronger correlation and a slightly smaller cost, Program A looks like a clear winner.

Unfortunately, neither evaluation team bothered to assess the reliability of their measure of reading achievement. If they had, they'd have learned that the outcome measure for Program A had a very strong reliability coefficient of .90. The outcome measure for Program B, on the other hand, had a reliability coefficient of only .20 -- an extremely noisy measure.

Okay, so Program A was not only cheaper and slightly more "effective" (according to the data), it was also more reliably assessed. Even better, right?

Wrong. In fact, the true relationship between Program A and reading achievement was 0.40. This is a moderate effect, and better than many in policy research. But the true relationship between Program B and reading achievement was .80 -- an extremely strong effect. Thus we've mistakenly rejected a very powerful intervention in favor of a significantly weaker one. Why? Because the Program B evaluation team had sloppy measurement!

CORRECTION FOR ATTENUATION

By the way, if Program B's evaluators had bothered to measure reliability, they could have retrospectively *corrected their estimate of the x-y relationship* to take into account the low reliability of their outcome measures.

This is called "correction for attenuation"; the correction has the following formula:

$$r_{tx,ty} = \frac{r_{ox,oy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

or in our simplified case (with perfect measurement of the x variable):

$$r_{tx,ty} = \frac{r_{ox,oy}}{\sqrt{r_{yy}}}$$

Given an observed correlation of $r_{ox,oy} = .36$, and a reliability coefficient for the outcome measure of $r_{yy} = .20$, we can get a corrected correlation of $.36/\sqrt{.20} = .80$.

Correction for attenuation is widely accepted among psychometricians as a reasonable adjustment. So why not just go ahead and measure things sloppily (and presumably, more cheaply), since we can always correct later? Two reasons. First, these formulas are of the "everything else being equal" variety, and don't hold up if we have overestimated either reliability or predictive validity (because of bias, not noise). Second, I have observed that in the policy arena, *most people scoff at corrections for attentuation*. It comes across as gimmicky--a cheap statistical trick to manipulate the data. Those suspicions often have a grain of truth to them, but are based in part on ignorance. But the reality is that you are *vastly better off measuring things (interventions, outcomes) reliably in the first place*.

FROM CORRELATION TO MULTIPLE REGRESSION

For didactic purposes, I've focused on the correlation coefficient. But many sophisticated policy analyses use some form of multiple regression, rather than simple bivariate correlations. Why does that matter?

1. Unreliability (error, noise) in the INDEPENDENT variable biases the regression coefficient downward, as with the correlation coefficient.
2. But unreliability in the DEPENDENT variable does not bias the estimated slope – the regression line will look the same. On the other hand, it DOES inflate the standard error, and so it reduces statistical power (a topic we'll visit later) and increases the chances of a TYPE II statistical error – saying there's "no effect" when there really is one. So even here, reliability matters.