

True versus Measured Information Gain

Robert C. Luskin
University of Texas at Austin
March, 2001

Both measured and true information may be conceived as proportions of items to which the respondent knows the correct answer, in the first case of items actually posed, in the second of the incomparably larger number of items that could have been posed. The algebra to follow reflects three key assumptions. First, measured information is distorted by *item sampling bias*. The universe of all possible items includes many too hard to be worth posing in a survey of the whole public, indeed many too hard to be answered by anyone but a handful of specialists. Thus measured greatly overestimates true information, especially among the truly better informed. Surely no one knows more than say 10% of the universe, but those who do will surely answer 100% of the information items in any mass survey correctly. Second, *the rich get richer*. Other things being equal, the more information people (truly) have, the more of any given batch of new information they tend to acquire. Learning is easier when the material falls effortlessly into context. And third, measured information gains are constrained by *ceiling effects*. Those who start at or near 100% can show little if any gain.

What follows is more “demonstration” (said to convince any reasonable person) than “proof”(said to convince even the unreasonable), because linear equations are only approximate for proportions. We are also implicitly assuming away the distinction between “known” and “answered correctly,” i.e., the existence of guessing, but the conclusions are undisturbed if guessing is either (a) corrected for or (b) either uncorrelated or positively correlated with true information.

Equations and Assumptions

The backbone of the set-up consists of equations depicting the dependence of measured on true information at times 1 and 2 (affected by item sampling bias), the dependence of time 2 on time 1 true information (affected by the rich getting richer), and, derivatively, the dependence of both measured and true information gain on time 1 true information (affected by ceiling effects). Our assumptions can be represented by ordinal stipulations as to the values of certain parameters. We take these three sets of equations and stipulations in order:

Item Sampling Bias: Let $x_{i1} = X_{i1} + \varepsilon_{i1}$ and $x_{i2} = X_{i2} + \varepsilon_{i2}$, where x_{i1} , x_{i2} , X_{i1} , and X_{i2} denote the i th participant's measured (lower case) and true (upper case) information at times 1 and 2, and ε_{i1} and ε_{i2} are errors. Further, let $\varepsilon_{i1} = c_1 + d_1 X_{i1} + u_{i1}$ and $\varepsilon_{i2} = c_2 + d_2 X_{i2} + u_{i2}$, where c_1 , c_2 , d_1 , and d_2 are parameters, and u_{i1} and u_{i2} are the random portions of ε_{i1} and ε_{i2} , assumed to have zero means and be uncorrelated with both true information at both times and with each other: $E(u_{i1}) = E(u_{i2}) = E(u_{i1}X_{i1}) = E(u_{i1}X_{i2}) = E(u_{i2}X_{i1}) = E(u_{i2}X_{i2}) = E(u_{i1}u_{i2}) = 0$.

Substituting for ε_{i1} and ε_{i2} brings us to

$$(1) \quad x_{i1} = a_1 + b_1 X_{i1} + u_{i1}$$

$$(2) \quad x_{i2} = a_2 + b_2 X_{i2} + u_{i2},$$

where $a_1 = c_1$, $a_2 = c_2$, $b_1 = 1 + d_1$, and $b_2 = 1 + d_2$.

Item sampling biases like those described above may be viewed as a matter of positive covariance between each X and its ε : the more information you truly have, the more it is overestimated. Since the relevant covariances are $E(\varepsilon_{i1}X_{i1}) = d_1 \sigma_{X_1}^2$ and $E(\varepsilon_{i2}X_{i2}) = d_2 \sigma_{X_2}^2$,

where $\sigma_{x_1}^2$ and $\sigma_{x_2}^2$ are the variances of X_{i1} and X_{i2} , the item sampling bias effects can be represented by the stipulation that $d_1, d_2 > 0$. Equivalently, in terms of (1) and (2), the stipulation is $b_1, b_2 > 1$. Two individuals differing by say 1% in true information will differ by more than 1% in measured information.

A rough way of seeing this stipulation's plausibility is to imagine the slopes b_1 and b_2 as being determined by the two points consisting of the minimum and practical maximum values of true information and the corresponding expectations for measured information, denote them at time 1 as $(X_{m1}, E(x_{m1}))$ and $(X_{M1}, E(x_{M1}))$. What are these points? Obviously, $X_{m1} = 0$, and $E(x_{M1}) = 1.0$. I have argued that X_{M1} cannot be high. For argument's sake, set it, generously by perhaps an order of magnitude, at 0.1 (as suggested above). That leaves $E(x_{m1})$, whose value depends on assumptions about guessing, but if either nobody guesses or there is a correction for guessing, $E(x_{m1}) = 0$. Now, forcing (1) through the points (0, 0) and (0.1, 1.0) yields $b_1 = 10.0$. This precise number depends on where we set X_{M1} , but since X_{M1} necessarily ≤ 1 , b_1 necessarily ≥ 1 , and the only way for b_1 to equal rather than exceed 1 is for X_{M1} , inconceivably, to equal 1.0.

Granted, this last conclusion hinges on $E(x_{m1}) = 0$. Thus consider what happens when guessing occurs but goes uncorrected. In the most extreme case, where everyone who doesn't know guesses, $E(x_{m1})$, for an index composed of true-false items like ours, $= 0.5$. Now $X_{M1} = 0.1$ implies $b_1 = \text{"only" } 5.0$, still far above 1.0, and although it can now > 1.0 for $X_{M1} > 0.5$, the threshold is still unthinkably high.

The Rich Get Richer: Let

$$(3) \quad X_{i2} = \alpha + \beta X_{i1} + v_i,$$

where v_i is a purely random disturbance having zero mean and zero correlation with X_{i1} , u_{i1} , and u_{i2} : $E(v_i) = E(v_i u_{i1}) = E(v_i u_{i2}) = E(v_i X_{i1}) = 0$. The rich getting richer can be represented by the

stipulation that $\beta > 1$. A given difference between two individuals in true information at time 1 will result in a bigger difference between them at time 2.

Ceiling Effects: From (1) - (3), the measured and true information *gains* can be written as

$$(4) \quad \Delta_i \equiv x_{i2} - x_{i1} = (a_2 - a_1) + (b_2\beta - b_1)X_{i1} + (u_{i2} - u_{i1}) + b_2v_i$$

and

$$(5) \quad \Delta_i^* \equiv X_{i2} - X_{i1} = \alpha + (\beta - 1)X_{i1} + v_i.$$

Ceiling effects may be viewed as a matter of X_{i1} 's having a negative slope in (4): the more one knows (and thus also appears to know), the less one can appear to gain. The stipulation is thus $b_2\beta - b_1 < 0$ or, equivalently, $b_1 - b_2\beta > 0$.

Results

Given this set-up, it can be established, first, that the correlation between Δ_i and Δ_i^* will always be less than the correlation between x_{i2} and Δ_i^* and, second, that the correlation between Δ_i and Δ_i^* can even be negative. I number these results below as 1 and 2.

To see these results, note that the correlations between measured and true information gain and between the latter and measured time 2 information are $\rho_{\Delta\Delta^*} \equiv \sigma_{\Delta\Delta^*} / \sqrt{\sigma_{\Delta}^2 \sigma_{\Delta^*}^2}$ and $\rho_{x_2\Delta^*} \equiv \sigma_{x_2\Delta^*} / \sqrt{\sigma_{x_2}^2 \sigma_{\Delta^*}^2}$, where as usual single-subscripted, squared σ 's represent variances, and double-subscripted, unsquared ones covariances. Trivially, we add the assumption that true time 1 information and all the error terms do actually vary from person to person, making the relevant variances not merely nonnegative but positive: $\sigma_{X_1}^4, \sigma_{u_1}^2, \sigma_v^2, \sigma_{u_2}^2 > 0$.

Under our assumptions about u_{i1} , u_{i2} , and v_{i2} , (1) - (5) can be manipulated, chiefly by a combination of substitutions and multiplications by X_{i1} , to yield

$$(6) \quad \sigma_{\Delta\Delta^*} = (\beta - 1)(b_2\beta - b_1)\sigma_{X_1}^2 + b_2\sigma_v^2$$

$$(7) \quad \sigma_{x_2\Delta^*} = (\beta - 1)b_2\beta\sigma_{X_1}^2 + b_2\sigma_v^2$$

$$(8) \quad \sigma_{\Delta}^2 = (b_2\beta - b_1)^2\sigma_{X_1}^2 + \sigma_{u_1}^2 + \sigma_{u_2}^2 + b_2^2\sigma_v^2$$

$$(9) \quad \sigma_{x_2}^2 = b_2^2\beta^2\sigma_{X_1}^2 + b_2^2\sigma_v^2 + \sigma_{u_2}^2.$$

Result 1: Starting with $\rho_{x_2\Delta^*} > \rho_{\Delta\Delta^*}$, squaring both sides, and multiplying both by σ_{Δ}^2 yields $\sigma_{x_2\Delta^*}^2 \sigma_{\Delta}^2 > \sigma_{\Delta\Delta^*}^2 \sigma_{x_2}^2$, which following some tedious algebra can be reexpressed as

$$[(b_1^2 - 2b_1b_2\beta)\sigma_{X_1}^2 + \sigma_{u_1}^2][(\beta - 1)b_2\beta\sigma_{X_1}^2 + b_2\sigma_v^2] > \\ [(1 - \beta)(b_1)\sigma_{X_1}^2][b_2^2\beta^2\sigma_{X_1}^2 + b_2^2\sigma_v^2 + \sigma_{u_2}^2],$$

and in turn as

$$(\beta - 1)b_1b_2\beta[b_1 - b_2\beta]\sigma_{X_1}^4 + b_1b_2[(b_1 - b_2\beta) + b_2(b_1 - 1)]\sigma_{X_1}^2\sigma_v^2 + \\ (\beta - 1)b_2\beta\sigma_{X_1}^2\sigma_{u_1}^2 + b_2\sigma_{u_1}^2\sigma_v^2 + (\beta - 1)b_1\sigma_{u_2}^2 > 0.$$

Since, by assumption, $b_1, b_2, \beta > 1$, $b_1 - b_2\beta > 0$, and all the variances are positive, all six addends on the left are positive, which establishes the result.

Result 2: The same assumptions establish that the first addend on the lefthand side of (6) is negative (since $b_2\beta - b_1 < 0$), while the second is positive, so $\sigma_{\Delta\Delta^*}$ and thus $\rho_{\Delta\Delta^*}$ will be negative when the first outweighs the second, i.e., when

$$-(\beta - 1)(b_2\beta - b_1)\sigma_{X_1}^2 > b_2\sigma_v^2.$$

In practice, this condition is hard to satisfy unless the gap between b_1 and b_2 is extremely large or X_{i2} is very highly predictable from X_{i1} (σ_v^2 is small). The more relevant lesson, therefore, may be that even when positive $\sigma_{\Delta\Delta^*}$ and thus $\rho_{\Delta\Delta^*}$ are driven toward zero as the configuration of parameters approaches this condition.