

## Why Does Measurement Error Propagate in the Physical Sciences, but not in the Social Sciences?

Here I am basically raising a question regarding how most people don't properly consider uncertainty in the social science measures of Education, Business, Psychology and the like, while in the physical sciences, this is almost always considered.

Taylor (1982, p. 40) notes, "...almost all interesting measurements involve ... two distinct steps, of direct measurement followed by calculation." He gives the example of computing speed – take the distance (Y) and the time (X), divide Y by X<sup>1</sup> to determine speed (q). Comparable social science analyses would be summing the items in a scale or analyses involving perhaps one independent variable and one dependent variable such as correlation, t-tests, Oneway ANOVA, Chi Square, etc., and becomes even more important in situations involving two independent variables and one or more dependent variables Factorial ANOVA, Multiple Regression, Discriminant Analysis, etc.

When estimating uncertainty, which must be present in any measure, Taylor<sup>2</sup> (1982) notes: "When a measurement ... of uncertainties ... involves two steps. One must first estimate the uncertainties in the quantities that are measured directly, and then find out how these uncertainties "propagate" through the calculation to produce an uncertainty in the final answer." Failure to properly consider uncertainties and tolerances in engineering results in such as collapsed bridges and buildings (Of course, a well-engineered bridge can collapse if it is not maintained properly as it ages. Did you know, for example, that something like 50% of all bridges in the U.S. are unsafe for traffic from a conservative engineering perspective.). Of course in the social sciences (Political Science, Politics<sup>3</sup>, Business, Education, Psychology, Anthropology, Medicine, Epidemiology, Criminology, Women's Studies, Africana Studies, etc., etc., etc....) improper estimation of uncertainty may merely result in giving the wrong treatment to a subject or the wrong medicine to a patient or the wrong recommendations to a prison warden, which don't have the same individual impact as a collapsed bridge, but cumulatively, can and do have vast consequences across all of the decisions and beliefs systems influenced by these fields.

### Computing Uncertainty

Taylor (1982) shows that when a physical measurement is made, say for example, the length of a board,<sup>4</sup> if it is measured to the tenth of a centimeter as 15.1 cm, then one assumes that the "true" value of this length falls somewhere in the interval between 15.05 cm and 15.15 cm. If one takes several estimates of the length of this board, with the shortest at 14.8 and the longest at 15.3 and a midpoint (mean, median, mode, etc.) of 15.1, then one could say that we assume the "true" value to be 15.1 with an uncertainty Of 0.3 cm (One basically uses the range of measures divided by two if the distribution is symmetric or 0.3 cm if it is asymmetric, as in this case, and taking the maximum deviation from the center as the uncertainty.). Another way of estimating the

---

<sup>1</sup> Note that I use capitals only to more clearly show the variables of interest, they should be small letters.

<sup>2</sup> I recommend this text for any researcher – excellent piece of work, particularly the pieces on precision and the inevitability of uncertainty. Taylor is/was a physicist, the most precise of all physical sciences.

<sup>3</sup> Speaking of Politics, if you haven't read *Peacemaking Among Primates* by De Wall, you should (see the end note).

<sup>4</sup> I won't deal here with the impossibility of ever having a "True" measurement of the phenomenon on which he expends considerable and well-thought out ink.

uncertainty of a specific measurement is to compute the variability of all individual measures, and take either the 95<sup>th</sup> (1.96  $\sigma$ ) or the 99<sup>th</sup> (2.576  $\sigma$ ) confidence interval as an estimate of uncertainty in the measurement of the “true” value.

### **Error Propagation**

Taylor shows that when two variables are added or subtracted (as for example when one sums the items in a scale), then the propagated uncertainty  $p$  is the addition of the error in measure  $X$  plus that of measure  $Y$  ( $\sigma_p = \sigma_x + \sigma_y$ )<sup>5</sup>. Thus, for a two item scale with 0.3 cm error in one and 0.2 cm in the other, the summed value would have an uncertainty of 0.5 cm and one could say with some confidence that two boards measured to say 15.3 cm and 15.8 cm have a combined length somewhere in the range of 30.1 plus or minus 0.5 cm (a true value somewhere between 29.6 and 30.6).

In the case where two variables are either multiplied or divided, the uncertainty of the estimate is computed by adding the fractional uncertainty of each number multiplied or divided (Where fractional uncertainty =  $\sigma/x$ ). Thus if we were computing momentum ( $p$ ) we could use a body’s mass ( $m$ ) and velocity ( $v$ ) multiplied together to estimate momentum ( $p=mv$ ). If  $m = 0.53$ , with  $\sigma = 0.01$  kg, and  $v = 9.1$  m/sec, with  $\sigma = 0.3$  m/sec, the respective fractional uncertainties would be  $m = 0.01/0.53 = 2\%$ , and  $v = 0.3/9.1 = 3\%$ . The combined uncertainty would be  $2\% + 3\% = 5\%$ . The estimate of momentum would then be:  $p = 0.53 \times 9.1 = 4.82$  plus or minus  $4.82 \times 5\% = 0.241$ , and we conclude that the momentum of the body is 4.8 kg\*m/sec plus or minus 0.2 kg\*m/sec, or somewhere between 4.6 and 4.8 kg\*m/sec (Note that one cannot estimate a final value to a greater number of significant digits than the measure having the least in the computation. In this case, velocity at 9.1 m/sec to one decimal place, has the least significant digits.).

Thus, adding or subtracting, the uncertainty in the total equals the summed uncertainties in each component. When multiplying or dividing, the uncertainty in the total equals the summed fractional uncertainties of each component. Note that although this final uncertainty is slightly greater than the standard error ( $\sigma$ ), it is usually only a miniscule difference except when uncertainty is more than perhaps 5%. This almost never occurs in physical science measurements, but does occur with some frequency in social science measurements. Thus, one would think that consideration of this would be of paramount importance in the social sciences, but it isn’t.

### **Why Do We See Such Little Use of Uncertainty Estimates in the Social Sciences?**

Social science measurement texts<sup>6</sup> teach that errors/uncertainties are smoothly distributed and cancel each other out. Although this may hold true sometimes, the physical sciences and engineering conservative assumptions is that they will sometimes cumulate. Fuller (1987) explains: “The models described are littered with assumptions regarding the nature of error. ...The true nature of error distributions is almost completely unknown. Assuming that errors are random and symmetric is no more reasonable than assuming that empirical score distributions are Gaussian in nature.”

---

<sup>5</sup> Note, this is approximately equal to, not actually = to in all cases of uncertainty measurement, I just don’t have the symbol available in any of my fonts.

<sup>6</sup> Realize that this means all of the fields mentioned earlier, plus many others.

The propagation of error becomes particularly important when error components of measures are large. Few will argue that many social science measures contain comparatively large quantities of uncertainty. Therefore, one must raise the question of why they are so rarely used in the social sciences as they are in the physical sciences to avoid false positives and negatives in decision processes. As an example of how important this is, in Micceri (2001)<sup>7</sup>, I show that admissions decisions based on standardized test scores will be incorrect more than 99% of the time for students within six percentile points below the cut off point. Even using Grade Point Averages (GPA), a far better predictor of college success, that error rate will be between 96% and 97%. We know that most, if not all social science measures (measures of humans) are heavily loaded with error/uncertainty. Abou-Sayf (1999) notes that differences in racial/ethnic responses from one survey to the next at the University of Hawaii ranged “...from 42% for Hawaiian/part-Hawaiian to 16% among Koreans.” Sex (gender for those who are politically correct) and race/ethnicity are two very commonly used variables in social sciences research, and they are used because they associate with a wide variety of important social variables such as salary, jobs, health, etc., etc. Now sex usually has a very small error component (circa 1%), but as Abou-Sayf notes, race/ethnicity contains massive quantity of error. Talkalkar, Waugh & Micceri (1993)<sup>8</sup> show that biases in varied measures range in a collection of such studies from 10% to about 50%. Error is not a trivial factor in the social sciences.

As an example of how uncertainty should be reported and used, let us take SAT standardized test scores as an example. Say, for example that student X had a combined SAT score of 950 and student Y, a 1030. Now the Standard Error of Measurement (SEM) for SAT quantitative and verbal subtests is 30 points each. The combined score uncertainty would therefore be 60 points. However, ETS (2001) has come up with a rather interesting (and I think questionable error estimate) called the Standard Error of Difference (SED) which they claim makes the combined SAT error a mere 40 points. Now they say that differences in scores of 60 points should indicate a “true” difference between subjects on the domain of interest (whatever that may be). This is 1.50 SED, which, assuming they use *z* tables for their calculations<sup>9</sup>, suggests that there is a two-tailed probability of error in that statement of about 13.5%. Now, I am not familiar with any statisticians who recommend using the 86.5% confidence intervals, most suggest a 90%, 95% or 99% interval (even Karl Pearson, the originator of these criteria). For the SED, the 95% would be (1.96 times 40 = 78), thus a 90% confidence using the SED would be an 80-point difference, using the SEM it would be 120. A 99% confidence interval<sup>10</sup> would be respectively 105 for the SED and 155 for the SEM. Thus, a conservative approach, using the SEM and a 99% confidence interval would say that student X’s SAT score was somewhere between 795 and 1105 and that student Y’s score was somewhere between 875 and 1185. Obviously, these ranges overlap substantially, and no reasonable person would be foolish enough to make a statement like: “Student Y has more academic ability than student X.”

---

<sup>7</sup> Available on this Web site under Empirical Evidence – Testing and College Admissions.

<sup>8</sup> Also on this Web site under Empirical Evidence, Statistics and Measurement, Measurement Error in Surveys.

<sup>9</sup> I am not getting into the fact that error is not likely to distribution in a Gaussian fashion to avoid complication.

<sup>10</sup> Failure to admit a student to college can have detrimental effects on both their and America’s future, so one would think that only extremely stringent error estimates would be used for these purposes.

I strongly recommend that rather than listing individual values on the various error-loaded human variables that are so commonly used in the social sciences, one lists the mid-point estimate (best guess of true value) with the confidence or uncertainty specified so that one eliminates the false sense of precision typically held by people who use such measures. Even variables like income,<sup>11</sup> even when reported to the IRS includes a considerable amount of error due to such factors as privacy issues, social desirability, and tax consequences.

### End Note

From Franz De Wall's *Peacemaking Among Primates* – Note that his earlier book, *Chimpanzee Politics* was recommended as reading for all congress people by Tip O'Neil, but De Wall said he wished he had written at that time "Peacemaking..." because "...Politics" was early and comparatively erroneous. On p. 105, De Wall states:

Some modern textbooks have it that science starts with a battery of hypotheses, dispassionately tested for acceptance or rejection. I believe that science starts with fascination and wonder. Charles Darwin did not sail away on the *Beagle* to test a theory; he returned with the ingredients for one. The exploratory phase, as it is called, is indispensable to creative research. The first tasks of an ethologist beginning research on an unfamiliar species are to try to get into its skin; to think at its level; and as stressed by the maestro of observation, Konrad Lorenz, to actually love the new species.

### References

- Abou-Sayf, F.K. (1999). *The integrity of ethnicity data*. Paper presented at the Association for Institutional Research Annual Forum, Seattle, WA, May 30-June 2, 1999.
- ETS (2001). *Test characteristics of the SAT I: reliability, difficulty levels, completion rates*. <http://www.collegeboard.org/sat/cbsenior/stats/stat002.html>. Downloaded from the WWW in February, 2000.
- Fuller, W.A. (1987). *Measurement error models*. John Wiley & Sons: New York.
- Micceri, T. (2001). *Facts and fantasies regarding admissions standards*. Paper presented at the AIR Annual Forum, Long Beach, CA, June 3-6, 2001.
- Taylor, J.R. (1982). *An introduction to error analysis: The study of uncertainties in physical measurements*. University Science Books, Oxford University Press, Mill Valley, CA.

---

<sup>11</sup> Note that there is a huge black market in the U.S. and that many employers pay part or all of an employee's wage using cash to reduce such as workman's comp and social security taxes.

Takalkar, P., Waugh, G., & Micceri, T. (1993). A search for TRUTH in student responses to selected survey items. Paper presented at the AIR Annual Forum, Chicago, IL, May 15-19, 1993.