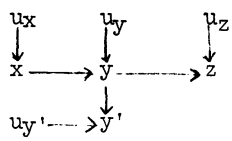
CONSEQUENCES OF NON-RANDOM MEASUREMENT ERROR IN PERCENTAGE DIFFERENCE TABLES

Henry Jay Becker
Johns Hopkins University

In the social sciences, attention to the precise measurement of individual variables on individual cases has been supported in principle far more than in practice. However, in several of the disciplines efforts have been made both to educate other researchers of the biasing effects of measurement error and to improve the way in which necessarily "soft" data may be employed in analysis. There have been several approaches taken to the problem reflecting the different academic disciplines and corresponding traditions of research methodology involved.

Sociologists involved in the complexities of multivariate causal models have begun employing both measured but fallible variables and their corresponding unmeasured but "true" variables in their analytic efforts. The causal modeling approach to measurement error is typified by the following causal diagram which postulates measurement error for variable y only, measurement error which is random with respect to x and z, and no direct causal relationships between variables x and z.



These assumptions allow one to estimate the causal paths p_{yx} and p_{zy} (equivalent in this case to the "true" correlation coefficients) from the observed correlations $r_{xy'}$, $r_{zy'}$ and r_{xz}. (Heise, 1969)

In fact, $p_{yx} = \sqrt{\frac{r_{xz'}\, r_{xy'}}{r_{y'z}}}$ and $p_{zy} = \sqrt{\frac{r_{y'z}\, r_{xz}}{r_{xy'}}}$.

For example, suppose $r_{xy'} = .2$, and $r_{y'z} = .4$ and $r_{xz} = .1$. If the above causal model is assumed, then $p_{yx} = r_{xy} = .22$ and $p_{zy} = r_{yz} = .45$. Also $r_{yy'} = .89$.

Most work by Sociologists in this area has followed this approach or a companion "multiple indicators" approach (Hauser and Goldberger, 1971). The research for the most part has been limited to situations where measurement error for a given variable can be assumed to be randomly distributed with respect to other variables in the causal system and other error terms as well. (For an exception, see Sullivan, 1974.) This is primarily because these methodologists, dealing with many variables simultaneously, have had their hands full just incorporating random measurement error effect into their measurements. In the end, however, problems produced by non-random measurement error will necessarily have to be attended to.

The consequences of non-random measurement error have been addressed by the economists, Lansing and Morgan (1971). They have shown, for simple bivariate correlations between continuous variables, the consequences of a number of different types of non-random measurement error. Measurement error is discussed in the context of evaluating its effects on the regression coefficient $\beta_{y'x'}$ where the primes denote observed variables. If the observed values are composed of a "true" value plus an "error" term (e.g., $y' = y + v$), the estimating equation for β in terms of "true" scores and "error" terms becomes:

$$\beta_{y'x'} = \frac{\text{cov } xy + \text{cov } xv + \text{cov } yu + \text{cov } uv}{\text{var } x + 2\text{cov } xu + \text{var } u} \qquad (1)$$

and that for the correlation coefficient:

$$r_{x'y'} = \frac{\text{cov } xy + \text{cov } xv + \text{cov } yu + \text{cov } uv}{\sqrt{\text{var } x + 2\text{cov } xu + \text{var } u}\ \sqrt{\text{var } y + 2\text{cov } yv + \text{var } v}}. \qquad (2)$$

An examination of equations (1) and (2) shows the role played by three types of non-random measurement error as well as the role of random measurement error:

(a) Association between the error deviate terms u and v. (cov uv) Positive association between error terms spuriously raises the reported association between x and y. Negative assocation between error terms attenuates the reported x:y relationship from the true value. A large enough negative assocation between error terms could reverse the reported association between x and y.

(b) Association between the true value of one variable and the error term of the other. (cov xv + cov yu) This has the same kinds of effects as (a); e.g., positive associations spuriously raise correlation and regression coefficients while negative associations reduce them or reverse them.

(c) Association between the true value of one variable and its error deviate. (cov xu; cov yv) With the placement of these factors in the denominator of Eq. (1) and (2), a positive association here would lower the β or r term from its true value while a negative association would increase the values of these statistics.

(d) Random error (var u; var v) Since these terms, also found in the denominator, can only be positive, their presence always causes regression and correlation coefficients to be underestimated.

Categorical variables can also be examined in the above framework. Consider the association between two dichotomies each coded "0" and "1." In this case, the true value is always negatively correlated with its own error term, since a true "0" score will have an error term which is either "0" (accurately measured true score) or "+1" (measured in error) while a true "1" score will have a "-1" error term when falsely reported as "0." Alone, this factor would cause reported associations to be higher than the true association. At the same time, however, the magnitude of the errors--or more precisely, the total

variance of the errors of observation in each variable considered separately--is acting to spuriously lower the reported association from the true value. This latter factor is universally more powerful, as can be shown by extensive algebraic manipulation.

Thus, whenever there is no association between the error term of each variable and the true score of the other and where error terms themselves are uncorrelated, the net effect of the other two factors is to spuriously reduce measured association from its true value.

Conversely, wherever error terms are sufficiently positively correlated or where there is a large enough positive correlation between true values of one variable and error terms of the other, these factors can overcome the general tendency for attenuation of relationships. In particular, for the regression coefficient to be spuriously high, the offending factors (cov xv + cov yu + cov uv) must be $\hat{\zeta}_{yx}$ times as great as the factors tending to diminish the measured relationship (cov xu + var u). The inequality with respect to the correlation coefficient is somewhat more complex: (cov xv + cov yu + cov uv)2 > r_{xy}^2 [var y (cov xu + var u) + var x (cov yu + var v) +(cov xu + var u) (cov yv + var v)] .

One of the principal advantages of working with dichotomies and categorical variables in general is the clarity with which statistical phenomena can be exemplified. To examine the effects of non-random measurement error on reported associations between variables, consider the data in Tables 1a and 1b.

Each table shows both the true and reported values of a variable X scored "+" and "-" cross-tabulated by a variable Y, whose "types" represent population subgroups. Variable Y is assumed to be know without error. For each "type" (i.e. population subgroup), 90% of the x = "+" cases are known without error to be "+." (The other 10% are called "false negatives.") For two types (B and C), 95% of the x = "-" cases are known without error; only 5% are "false positives." In the other two types (A and D), false positives number 20% of the true x = "-" cases. Overall, more than 88% of the cases are accurately scored on both variables X and Y.

However, in neither case is the result similar to what it would have been if only random measurement error (i.e. equal error rates) were present. If, for example, the error rate were 10% in each cell, the reported percent difference would have been 16% instead of 20%, making the reported result a "conservative" estimate of the true relationship. Instead, in Table 1a, the reported association slightly exaggerates the true association between the variables. (In percentage difference terms, 20% points true difference is increased to 26%.) In Table 1b, the association is markedly understated by the reported data (the 20% true difference is reduced to 5%).

Several earlier papers by biostatisticians involved in public health research have discussed the

TABLE 1: EFFECTS OF DIFFERENTIAL FALSE POSITIVE RATES ON REPORTED ASSOCIATIONS (HYPOTHETICAL DATA)

Table 1a

TRUE VALUES

	Type A	Type B
+	40%	20%
-	60%	80%
Total	(100)	(100)

%difference = 20 percentage points
gamma = .45
odds ratio = 2.7

ERROR RATES

True Values	Type A	Type B
+(FN)	10%	10%
-(FP)	20%	5%

REPORTED VALUES

	Type A	Type B
+	48%	22%
-	52%	78%
Total	(100)	(100)

% difference = 26 percentage points
gamma =.53
odds ratio = 3.3

Table 1b

TRUE VALUES

	Type C	Type D
+	40%	20%
-	60%	80%
Total	(100)	(100)

% difference = 20 percentage points
gamma = .45
odds ratio = 2.7

ERROR RATES

True Values	Type C	Type D
+(FN)	10%	10%
-(FP)	5%	20%

REPORTED VALUES

	Type C	Type D
+	39%	34%
-	61%	66%
Total	(100)	(100)

% difference = 5 percentage points
gamma = .11
odds ratio = 1.2

issue of non-random error's effects on measured association using cross-classified dichotomies, but their work needs more widespread attention and follow-up.

Keys and Kihlberg (1963) graphed the expected bias on individual variables (true vs. reported "prevalences"--i.e. percent "+") for a variety of false positive and false negative rates. They also graphed the bias of an associational measure (proportionate deviation from a true "relative risk" of 1.0) against that common true risk given particular combinations of the four error rates involved. However, the authors did not calculate bias in the measured assocation under the condition that different (true) rates were present (i.e., true relative risk $\neq$ 1.0). They did explore the complexities of calculating estimated bias with fallibility on both measured variables, but produced no graphical examples as in the case of a single variable measured in error. (Instead of only four parameters--two false positive and two false negative rates--that situation involves 16 parameters.)

Goldberg (1975), examining the same type of hypothetical data, found that subgroup differences in false positive rates produces far larger average bias than differences between false negative rates for "prevalences" under 50%; coversely for higher prevalences.

One other point shown in the above example is that the direction of bias depends on whether the group with the greater or the lesser proportion of "+" cases is the one with the larger proportion of false positives. If false positives are more frequent in the group with more "+" cases, generally the reported relationships will be an exaggeration of the true relationship. Conversely major underreports of association tend to occur where false positives are concentrated in the group with fewer "+" cases. These results are reversed and apply to false negative rates where "+" proportions are above 50%.

In general, the reported percentage difference is related to the true percentage difference and the various error factors as $(p_1'-p_2') = p_1 (1-FN_1 - FP_1) - p_2 (1-FN_2 - FP_2) + (FP_1 - FP_2)$ where for subgroups i = 1, 2, p_i' are reported proportions "+," p_i are true proportions "+", FN_i are false negative rates (proportions of true "+" reported as "-") and FP_i are false positive rates (proportion of true "-" reported as "+"). Each p_i is also related to the reported p_i' and the error terms, thusly:

$$p_i = \frac{p_i' - FP_i}{1-FN_i-FP_i}.$$

Under the assumption that all error rates, FN_i and $FP_i \leq e$, where $p_1 - p_2 \geq 0$, the maximum and minimum value for $(p_1 - p_2)$ are given by

$$\left(\frac{1}{1-e}\right)(p_1' - p_2') \pm \left(\frac{e}{1-e}\right).$$

As shown in Table 2, extremely large variations in the magnitude and direction of bias can be observed due to non-random measurement error even when the absolute proportions of cases in error is rather modest, for example e = .2.

In most cases, although not for the maxima and minima shown in Table 2, the magnitudes of p_1' and p_2' as well as their difference $(p_1' - p_2')$ affect the value expected for $(p_1 - p_2)$ under specific error rates. (See Table 3.) The smaller the magnitude of p_1' and p_2' (given $p_1' - p_2'$ = k), the greater the difference between true and reported associations.

TABLE 2: RANGES OF POSSIBLE TRUE PERCENTAGE DIFFERENCES GIVEN CERTAIN LIMITS ON MEASUREMENT ERROR

REPORTED VALUES[2] $(p_1'-p_2')$	MAX & MIN TRUE VALUES $(p_1 - p_2)$ FOR MAXIMUM ERROR RATE e[1]					
	e=.3		e=.2		e=.1	
	MAX	MIN	MAX	MIN	MAX	MIN
+.3	.86	.00	.62	.12	.44	.22
+.2	.71	-.14	.50	.00	.33	.11
+.1	.57	-.29	.38	-.12	.22	.00
+.05	.50	-.36	.31	-.19	.17	-.06
.00	.43	-.43	+.25	-.25	.11	-.11

1. excluding effects of sampling error
2. with $p_2' > p_1'$, results are the same except minima and maxima are reversed and signs are reversed also.

Even rather minor differences in the two false positive rates can produce rather major differences in measured association. The example in Table 4 shows that under rather ordinary conditions, a reversal of false positive rates changes the result from nearly no measurement error effect (situation (1)) to a very strong attenuation of the true relationship (situation (2)).

The important questions that are raised by this discussion of non-random error concern not the possible effects of differential error rates, but the actual effects caused by real differences in error rates. Here our knowledge is hampered by the paucity of validation studies that report validity for subgroups rather than solely for the complete sample studied. Two sources have been found in the public opinion literature which do report on validity of responses for subgroups.

One is the Denver validity study done in 1949 but most recently analyzed in (Cahalan, 1969). Reworking the results on the basis of the published statistics, interesting and quite major differences between subgroups can be seen for both false negative and false positive rates, with the consequence that many relationships would have been reported erroneously without the validating information.

ABLE 3: EFFECTS OF MAGNITUDE OF p_i' AND $(p_1' - p_2')$ ON BIAS OF $(p_1'-p_2')$ GIVEN FOLLOWING ERROR RATES: $FN_1 = .10$ $FN_2 = .10$ $FP_1 = .20$ $FP_2 = .10$

REPORTED VALUES			TRUE VALUES	BIAS
$p_1'-p_2'$	p_1'	p_2'	p_1-p_2	$(p_1-p_2)-(p_1'-p_2')$
.30	.5	.2	.30	.00
	.4	.1	.29	-.01
.20	.5	.3	.18	-.02
	.3	.1	.14	-.06
.10	.5	.4	.05	-.05
	.3	.2	.02	-.08
.05	.5	.45	-.01	-.06
	.3	.25	-.05	-.10
.00	.5	.5	-.07	-.07
	.3	.3	-.11	-.11
-.05	.45	.5	-.14	-.09
	.25	.3	-.18	-.13
-.10	.4	.5	-.21	-.11
	.2	.3	-.25	-.15
-.20	.3	.5	-.36	-.16
-.30	.2	.5	-.50	-.20

TABLE 4: EFFECTS OF SMALL DIFFERENCES IN FALSE POSITIVE RATES

TRUE SCORES

	A	B
%+	25%	15%

% difference = 10 percentage points
gamma = .31
odds ratio = 1.9

ERROR RATES (1)

	A	B
FN	12%	12%
FP	15%	10%

ERROR RATES (2)

	A	B
FN	12%	12%
FP	10%	15%

REPORTED RESULTS (1)

	A	B
%+	33%	22%

% difference = 11 percentage points
gamma = .27
odds ratio = 1.7

REPORTED RESULTS (2)

	A	B
%+	30%	26%

% difference = 4 percentage points
gamma = .10
odds ratio = 1.2

The Cahalan data is only approximately reproduced in Table 5. (The absence in the original article of sufficient information about the magnitude of "don't knows" for men and women separately made it impossible to exactly reproduce the data.) The data as estimated here, though, exhibit some remarkable differences between cells in error rates for some of the measured associations. Notice that in both cases of large reported associations by sex (the Community Chest contributions and the drivers' licenses), the true relationships are of smaller magnitude--one of them being less than 40% as large as the reported association.

The second article from Public Opinion Quarterly reporting validity coefficients for subgroups is Weiss's (1968) report of interviews with mothers receiving welfare. The results given in Table 6 illustrate the tendency for upward bias in reported percentage differences to occur only where both $p_1 - p_2$ is small and differences in false positive rates exist. For cases where $p_1 = p_2$, bias is also greater when the p_i approach zero. (See Keys and Kihlberg, 1963.)

Finally, the question remains: given the clearly major consequences of non-random measurement error on reported associations between variables, what now needs to be done to better take account of this problem?

Journal articles most often report findings with due respect for sampling error. Confidence limits, for example, are reported for percentage differences and the result is reported as statistically significant if the null hypothesis of "no difference" lies outside the range of 2 standard errors of the sample statistic. It would seem to make some sense to include in this statement the range of uncertainty that is attributable to possible measurement errors.

That is, the analyst might propose two sets of error rates for a given two-by-two table that would have opposite effects on the bias of the reported result and which are just extreme enough to be plausible given the possible error-producing causes present in the particular data. Those two sets of error rates, in turn, would be applied to the reported data in order to produce "extreme" but plausible "true" sets of sample data. Confidence intervals would then be calculated for these extreme but hypothetically true sample data tables. The summary confidence intervals reported would thus reflect both sampling error and plausible ranges of error caused by measurement inadequacies.

Consider an example, more or less randomly chosen from Rosenberg's The Logic of Survey Analysis (1968). (See Table 7.) In our use of his re-examination of the relationship between vote intention and respondent's education (1948 data), we percentage on education (% with some high school) taking education as the variable possibly reported in error. Vote intention is assumed to be reliably and validly known. Let us make two contrasting suppositions: (t_1) that persons claiming an intention to vote are more

TABLE 5: MEASUREMENT ERROR EFFECTS IN THE DENVER VALIDITY STUDY (1949)

	(N)	REPORTED[1]	TRUE	ERROR RATES[2] FN	FP	Gammas
Community Chest Contributions (percent "yes")						
MEN	327	85%	35%	0%[3]	76%	r=.65
WOMEN	404	55%	24%	0%	41%	t=.26
△		30%	11%			
Driver's License (percent "yes")						
MEN	423	78%	66%	3%	41%	r=.76
WOMEN	497	33%	32%	12%[4]	8%	t=.61
△		45%	34%			
Voting in 1946 Cong. Elections (percent "yes")						
MEN	382	55%	32%	5%	37%	r=.14
WOMEN	447	48%	33%	9%	27%	t=-.02
△		7%	-1%			
Voting in 1948 Pres. Elections (percent "yes")						
MEN	423	74%	61%	1.6%	36%	r=.05
WOMEN	497	72%	61%	1.3%	31%	t=.00
△		2%	0%			

1. Percentage recalculated from published data omitting "don't knows."
2. FN = false negative FP = false positive
3. Not checked: assumed = 0.
4. Unable to reproduce marginals from published data: doubtful accuracy.

TABLE 6: MEASUREMENT ERROR EFFECTS IN THE WEISS (1968) STUDY OF WELFARE MOTHERS

	PERCENT VOTING IN 1964 ELECTION					
	(N)	REPORTED	TRUE	ERROR RATES[3] FN	FP	Gammas
Worked more than 10 years[1]	189	53%	29%	0%[2]	34%	r=.17
Worked less than 10 years	329	44%	28%	0%	22%	t=.02
△		9%	1%			
Expects children to continue educ. past high school	141	64%	42%	0%[2]	39%	r=.33
Does not expect...	191	47%	29%	0%	26%	t=.29
△		17%	13%			
Did not get as much education as wanted	372	48%	28%	0%[2]	27%	r=.02
Got enough education	147	47%	28%	0%	26%	t=.00
△		1%	0%			

1. Assumed known without error
2. Assumed to be zero
3. FN = false negative FP = false positive

likely to overreport their education than those not claiming an intention to vote (because of the good citizen image of voting and completing one's education); (t_2) alternatively, that persons not planning to vote, being more alienated from the culture, make more errors in their self-reports (both over-reporting and underreporting) than do those planning to vote. Table 7 shows the the four parameters assumed for each of these two basic models. The data in Table 7 show how the reported relationship (a) is affected by the measurement error assumptions (t_1) and (t_2) producing hypothesized "true" sample results (b) and (c). If the confidence intervals for (p_1-p_2) are calculated from the "extreme" results (b) and (c) instead of from (a), they change the estimate of (p_1-p_2) from $8\% \leq (p_1\text{-}p_2) \leq 20\%$, $p = .95$, to $5\% \leq (p_1\text{-}p_2) \leq 24\%$, $p = .95$.

TABLE 7: ILLUSTRATION OF THE USE OF MEASUREMENT ERROR HYPOTHESES IN THE ROUTINE PRESENTATION OF CONFIDENCE INTERVALS[1]

	REPORTED EDUCATION: % WITH SOME HIGH SCHOOL (% "+")		EXPECTED UNDER ERROR HYPOTHESES	
			t_1	t_2
Vote Intention:	(a)	(N)	(b)	(c)
Positive	59%	(2515)	52%	58%
Negative	43%	(297)	41%	40%
△	16%		11%	18%
Confidence Intervals:				
p=.95	$10\% \leq (p_1-p_2) \leq 22\%$		$5\% \leq (p_1-p_2) \leq 24\%$	
p=.68	$13\% \leq (p_1-p_2) \leq 19\%$		$8\% \leq (p_1-p_2) \leq 21\%$	

t_1		Education FN	FP
Vote Int:	Pos	.01	.15
	Neg	.03	.05

t_2		Education FN	FP
Vote Int:	Pos	.02	.05
	Neg	.07	.10

1. Data from Rosenberg, 1968

One standard error confidence limits for the data based on the measurement error assumptions do not differ much from the two-standard error limits based solely on sampling error: $8\% \leq (p_1-p_2) \leq 21\%$.

Another procedure that could be more widely utilized is a version of Schuman's suggested "random probe" method of clarifying the meaning of information obtained through interview procedures. Schuman (1966) proposed that respondents to a given survey be asked to give reasons, interpretations, or explanations for their answer to a subset of all closed-ended questions in the interview. These extended responses are used to calculate the proportion of closed-ended responses that were "accurate." A different subset of respondents is used to "validate" in this way each of several subsets of questions.

This random probe technique has the potential of providing information concerning response error rates for various subgroups being compared in the statistical analysis. Thus real data can be employed as "expected values" in the error rate table, and variances for these error rates can be computed. These variances can be used to produce the "extreme" values of the "true" results which can then be used, along with the sampling distribution of the statistic, to compute reasonable confidence limits for the data.

Obviously, much additional work needs to be done in order to make it possible for measurement error effects to be incorporated into routine statistical procedures to the extent that sampling error has been. But the expanding application of complex statistical methodologies to the "soft" data of the social sciences has not been accompanied by sufficient appreciation of the magnitude and direction of effects caused by unreliability and invalidity of measurement. This is a problem that requires major attention if we are to avoid perpetuating the publication of technically flawless conclusions invalidated by the use of extremely soft and fallible data.

References

Cahalan, Don, "Correlates of Respondent Accuracy in the Denver Validity Survey," Public Opinion Quarterly, Vol. 32, (Winter, 1968), 607-621.

Goldberg, Judith D., "The Effects of Misclassification on the Bias in the Difference Between Two Proportions and the Relative Odds in the Fourfold Table," Journal of American Statistical Association, Vol. 70 (September, 1975), 561-567.

Hauser, Robert M. and Goldberger, Arthur S., "The Treatment of Unobserved Variables in Path Analysis," in Sociological Methodology 1971, Herbert L. Costner, ed. Jossey-Bass, 1971.

Heise, David R., "Problems in Path Analysis and Causal Inferences," in Sociological Methodology 1969, Edgar F. Borgatta, ed. Jossey-Bass, 1969.

Keys, Ancel and Kihlberg, Jaakko K., "Effect of Misclassification on Estimated Relative Prevalence of a Characteristic," American Journal of Public Health, Vol. 53 (October 1963), 1656-1665.

Lansing, John B. and Morgan, James N., Economic Survey Methods. Institute for Social Research, University of Michigan, 1971.

Rosenberg, Morris, The Logic of Survey Analysis. Basic Books, 1968.

Schuman, Howard, "The Random Probe: A Technique for Evaluating the Validity of Closed Questions," American Sociological Review, Vol. 31 (1966), 218-222.

Sullivan, John L., "Multiple Indicators: Some Criteria of Selection," in Measurement in the Social Sciences, H.M. Blalock, Jr., ed. Aldine, 1974.

Weiss, Carol H., "Validity of Welfare Mothers' Interview Responses," Public Opinion Quarterly, Vol. 32 (Winter, 1968), 622-633.