

Including Measurement Error When Assessing Statistical Robustness

By:

Theodore Micceri, Ph.D.

Assistant Dean's Office, College of Engineering,
University of South Florida, Tampa, FL 33620

Scott W. Campbell, Ph.D.

Department of Chemical Engineering, College of Engineering,
University of South Florida, Tampa, FL 33620

Internal Technical Report

Thursday, May 30, 1991

Abstract

Gregoire and Driver (1987) evaluated the robustness of several statistics by using nonreversible transformations from several simulated continuous populations to create Likert-scale data . They took the Likert-scale equivalent of the true population mean as the location parameter of interest. The authors thereby attempted to include measurement error in their assessment of statistical robustness. Rasmussen (1989) contended the parameter of interest was truly the population mean of the transformed data not that of the underlying population. Stigler (1977) appraised the robustness of several location estimates using values accepted today for several physical properties as "true" scores, and values produced by crude and biased 19th century measurement as the score distribution measuring those "true" values. Eisenhart (1977) took Rasmussen's tack and criticized this selection of "true score", contending that a statistic's robustness should be evaluated external to measurement error. The current paper contends that measurement error is often the greatest impedance to accurate research results, and that both the absolute and relative influence of measurement error is of great interest. An analysis of comparative findings from two prior studies suggests that although confidence intervals about a mean are not greatly disturbed by the distributional anomalies of bounded Likert-type data, measurement bias has disastrous effects on such tests and these effects are not revealed by studies that examine statistical error alone. A method that allows the exploration of this issue using Monte Carlo techniques is proposed.

Including Measurement Error When Assessing Statistical Robustness

The concern here involves the influence of different error sources on statistical robustness and the impact this may have on both the selection of statistics and interpretation of findings. Historically, most robustness research has avoided consideration of measurement error (validity). Researchers routinely assume that their measures are “consistently and validly” measuring the trait of interest. Although limiting statements abound in technical papers, these tend to be lightly dismissed or forgotten when decisions are made. The current discussion centers on the controversy triggered by two studies that sought to include both statistical and measurement errors in their assessments of statistical robustness. Further, we assert that this commingling of errors may be appropriate for all save purely theoretical inquiries. A method to investigate the differential influence of statistical and measurement error is described.

Gregoire & Driver (1987) conducted a robustness study that focused “...on the ability of selected procedures to uncover population characteristics using Likert-scale response format data.” (p. 159) The authors evaluated the robustness of certain statistics by using nonreversible transformations from several simulated continuous populations to create Likert-scale data. They took the Likert-scale equivalent of the true population mean as the location parameter of interest. Rasmussen (1989, p. 168) claims their study to be fundamentally flawed for the location tests because

...they transformed from the continuous distribution to the Likert distribution and then incorrectly used the Likert-scale equivalent of the mean of the continuous population in calculating the Likert-scale statistics. Their attempt to calculate statistics on the second distribution using the mean of the first distribution is not logically justified.

A similar debate occurred as a result of Stigler’s (1977) article in which he declared the arithmetic mean to be a reasonable estimator and the 10 percent alpha trimmed mean the optimal statistic in a study of real-world distributions. His findings, like those of Gregoire & Driver (1987), were in some disagreement with several prior studies. Selecting only data collected by reputable scientists, Stigler (1977) used 20 data sets of size 20 from 18th and 19th century measurements on the parallax of the sun, the mean density of the earth and the speed of light. He defined the “true score” for these data sets as that value accepted by physicists today. Often this value was located far in the tails of distributions generated by crude and biased 19th century measurement instruments. Eisenhart (1977) took Rasmussen’s tack and criticized this selection of “true score”, contending that values located at percentiles such as the 88th, 32nd and 8th “...are clearly unsuitable for judging the relative merits of a group of estimates that are ‘arguing’ over which value in a central core of a set of data ‘best’ summarizes the evidence of the set as a whole.” (p. 1086) Stigler suggested that the long tails found in some of the distributions occurred because the experimenters were aware of their instrument's biases and adjusted

observations to account for this. Therefore, the sample mean, which is influenced by these values in the distribution's tails, better represents the underlying "true value" than so-called robust estimators which reduce the weight of outlying values to zero or near zero. Stigler took great pains to find data that he consider relatively "pure" and collected only by the most trusted scientists. He notes that much scientific data in physics, chemistry and biology cannot be trusted. This may be due in part either to scientists' wishes to not encounter "different" results, or because of the driving influence of preconceptions on perceptions and reporting.

Both Stigler and Gregoire & Driver investigated the relationship of a statistical estimate to a "true" population value and thereby include some effects of measurement error in their results. Eisenhart and Rasmussen concern themselves solely with a statistic's reliability by recommending that the only error of interest to statistics is statistical error. We contend that both are of interest and that a comprehensive set of analyses is possible that evaluates the impact of both statistical and measurement error on research findings. Rasmussen's reanalysis considerably expands the information available from the earlier work and should be considered in conjunction with Gregoire & Driver's findings. Alternatively it may be said that Stigler and Gregoire & Driver expand on the information contained in studies that limit themselves to statistical error. Three important questions may be addressed by comparing the results of Gregoire & Driver (1987) with those of Rasmussen (1989): 1) how reliably do the various statistics attain nominal alpha for obtained distributions (the typical question of robustness studies), 2) how reliably do the various statistics attain nominal alpha for the underlying population parameters and 3) of all error identified, what proportion is attributable to measurement and which to statistical factors.

The repercussions of measurement error on research findings differs substantially from field to field. Few will deny that measures of invisible socio- and psychometric traits must contain more errors of both the systematic (bias) and random types than measures of more readily accessible physical properties. Biometrics and other measures of similar ilk probably fall somewhere between the preceding extremes. In such disciplines as psychology, education, sociology, business, public health and medicine, among others, measures may be both biased and unreliable estimates of an underlying trait. This may result from variability or bias inherent in the measurement instrument itself, such as that found in psychometric projective techniques or performance evaluation measures. It also may be due to variability in the subject. For example, in medicine, measures of characteristics such as blood pressure and chemistry vary greatly within the same subject from hour to hour and day to day. In either situation, measurements vary either symmetrically or asymmetrically about some relatively central point that may be called a "true" score.

For some of the more theoretically inclined, such measurement error may be of little interest aside from its influence on the shape of score distributions. However, for those who must decide what statistics to use on decision-influencing data, the consistency and accuracy with which statistical estimates represent a population's "true" value is of paramount interest. It would be extremely useful to know approximately what proportion of the error in our

inquiries results from statistic(s), what proportion from measurement, and what types of measurement error distort which statistics to what degree. As an example of such an investigation, we may compare Table 4 in Gregoire & Driver (1987) with Table 2 in Rasmussen (1989) regarding the Type I error rates of confidence intervals about respective population means. Type I error in the first case was defined as the percentage of confidence intervals that failed to include the Likert-scale equivalent of the true population mean and in the second case, as the percentage of confidence intervals that failed to include the actual mean of the Likert-scale values. Most values in Rasmussen's Table 2 are close to the nominal alpha of .05 (ranging from .03 to .08). This tells us that the distributional nature of the transformed data, which involved several rather extreme characteristics, did not substantially affect Type I error for the ultimate Likert-scale distributions. Confidence intervals about the mean proved fairly reliable estimates of the transformed parameter. Therefore, almost all of the error in Gregoire & Driver's Table 4 results from measurement error.

Table 1 shows the absolute error associating with each situation in the two tables, where error is defined as the difference between the obtained proportion of sample means for which a 95 percent confidence interval fails to capture the population parameter and nominal alpha (.05). It also shows a rather simplistic approach to separating measurement from statistical error by again subtracting the difference in these absolute errors. Although probably inappropriate when error values are small, for most of the cells in Table 1, the magnitude of the error contained in difference scores pales in comparison to the error associating with the measurement transformations used by Gregoire & Driver. Evaluation of Table 1's first three columns clearly shows that statistical error is minimal and the second three show that statistical plus measurement error is often complete (.05 + .95 = 1.00 or 100% of confidence intervals fail to include the population parameter). For large samples, this is not surprising because as sample size increases, confidence intervals shrink and a location estimate biased by measurement error remains distant from μ but not from a score distribution's mean.

Table 1

Type I Error Rates for Gregoire & Driver's Experiment I and Rasmussen's Corrections

n	Rasmussen			Gregoire & Driver			Measurement Error Only*		
	Statistical Error Only			Measurement & Statistical Error			Measurement Error Only*		
	Mapping			Mapping			Mapping		
	A	B	C	A	B	C	A	B	C
Population I									
10	2	0	0	2	0	25	0	0	25
25	-2	0	0	1	2	72	3	2	72
50	1	-1	-1	0	7	93	-1	8	94
100	0	0	-1	-1	16	95	-1	16	96
300	0	-1	0	1	52	95	1	53	95
500	0	1	-1	-1	75	95	-1	74	96
Population II									
10	-1	0	1	-1	1	-2	0	1	-3
25	0	-1	0	0	4	70	0	5	70
50	1	0	0	-1	11	94	-2	11	94
100	0	-1	0	0	17	95	0	18	95
300	3	0	2	-1	46	95	-4	46	93
500	0	1	0	1	71	95	1	70	95
Population III									
10	0	0	0	7	0	5	7	0	5
25	1	0	0	19	1	26	18	1	26
50	0	-2	0	30	0	54	30	2	54
100	0	0	-1	62	1	86	62	1	87
300	0	1	0	93	3	95	93	2	95
500	0	-1	1	95	4	95	95	5	94
Population IV									
10	-1	1	0	0	2	41	1	1	41
25	-1	1	-1	0	7	88	1	6	89
50	1	-1	-1	0	20	95	-1	21	96
100	-1	-1	0	1	45	95	2	46	95
300	0	1	-1	-1	86	95	-1	85	96
500	-1	0	0	-1	94	95	0	94	95

* Column 2 Minus Column 1

Note. Each entry is the difference between the percentage of 1,000 confidence intervals failing to include either the Likert-scale mean or the mean of the transformed data and nominal alpha (.05).

Regarding the influence of different biases on these errors, the last three columns of Table 1 show that Mapping C consistently inflates error for all four

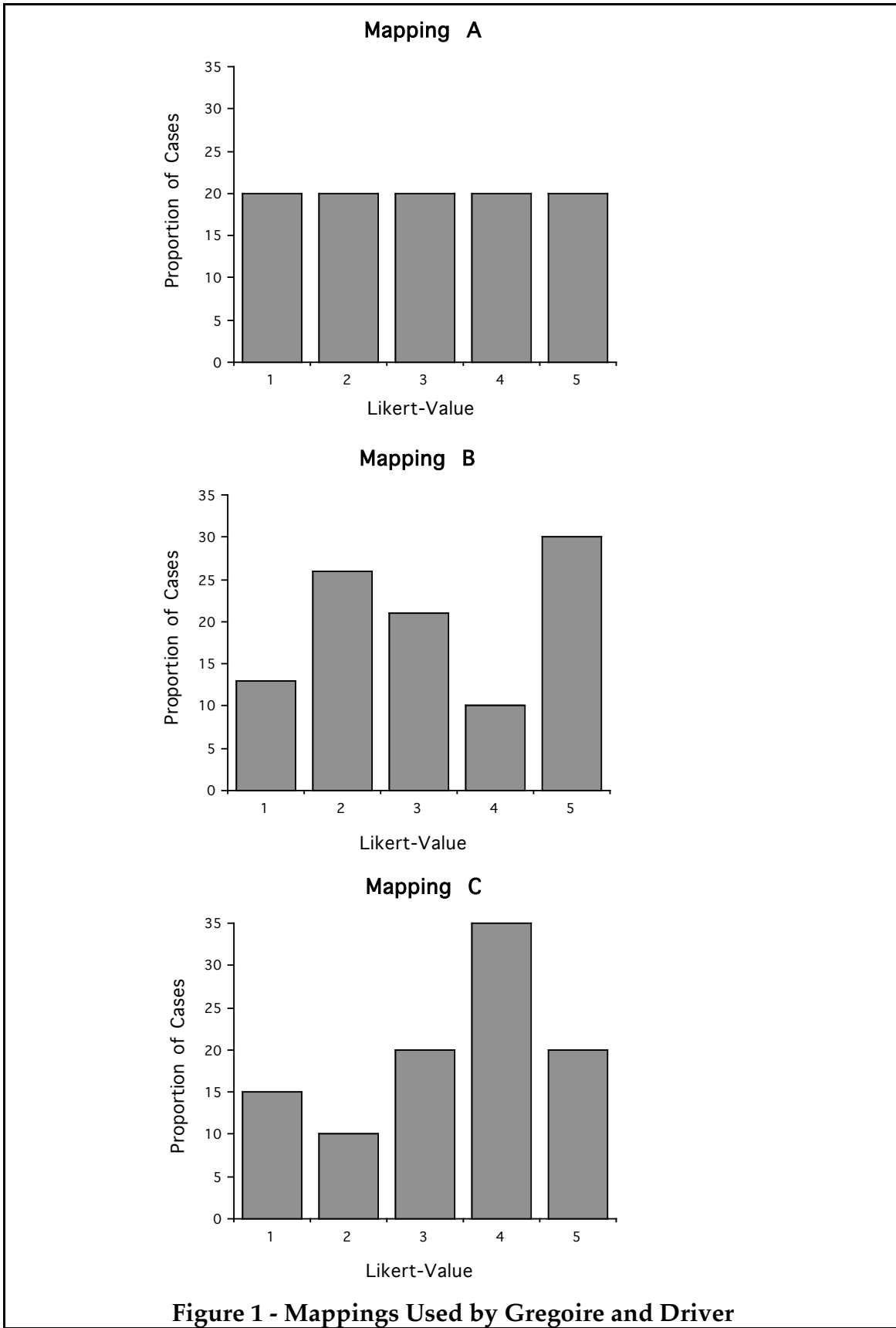
population types and for every sample size, frequently to a large extent. Mapping B causes problems for populations II and IV, particularly for samples of size 50 or greater, and Mapping A only for population III. Table 2 shows the populations and mappings used in the described studies.

Table 2

Populations and Mappings Used in Gregoire & Driver and Rasmussen

Popu- lations	Distribution Characteristics	Mapping	1	2	3	4	5
I	Uniform: range = 0-100	A	0-20	21-40	41-60	61-80	81-100
II	Normal: N(M=50, SD=17)	B	0-13	14-39	40-60	61-70	71-100
III	Normal: N(M=25, SD=17)	C	0-15	16-25	26-45	46-80	81-100
IIII	Bimodal normal mixture: N(M=30, SD=10), N(M=70, SD=10)						

Figure I provides a visual representation of the proportion of scores assigned to each Likert-scale value for each mapping. Viewed in conjunction, it is obvious that the second mode (value 5) of Mapping B compensates for the floor effect produced by population I. Mapping C, with its asymmetry, causes problems for all population situations, and Mapping A (Uniform) only for Population III, where asymmetry is produced by the underlying population floor effect. Thus, clearly, asymmetric measurement bias causes inflated error rates for confidence intervals about the mean of an obtained score distribution.



These comparative findings suggest that although confidence intervals about a mean are not much influenced by several distributional anomalies of bounded Likert-type data, measurement bias has disastrous effects on such tests and these effects are not revealed by studies that examine statistical error alone. However, these hypothetical data in no way indicate to what extent or when such biases might occur among real data. We suggest that the investigation of this matter could prove extremely enlightening. Asymptotic theory provides no clues for the researcher interested in this issue. Such investigations are feasible only using empirical Monte Carlo studies.

It is possible, by employing several measures of the same physical property, to assess "real" measurement bias. The measurement of physical properties has advanced to a point where errors for the best instruments are minute and several different measures of varying precision may be used to estimate the same true value. One example of this is the measurement of time. The most precise measure available is the cesium clock. Estimates of its accuracy suggest that two such clocks will differ by no more than one second every 3,000 years. A somewhat less accurate measure is the digital quartz crystal clock, the most precise of which should differ by no more than 5 minutes per 3,000 years (Halliday & Resnick, 1974, p. 6). Although not perfect, for the purposes described here, such a clock could provide a reasonable "true" value. Other, more variable measures could be used to appraise measurement error. Less accurate quartz crystal watches (particularly used ones) should provide both bias and variability. Mechanical wrist watches (particularly used ones) should exhibit similar and perhaps greater errors. Although not amenable to computer-based recording methods, wrist watches having marks only every five minutes would introduce observer error, as would those having marks only at six and twelve o'clock. Sundials could introduce yet greater error and the least precise measure would be estimates produced by an individual looking at the sun. Recording of numerous estimates of the same time interval by several measures of differing precision (or several observers) would provide a series of score distributions each including measurement error of different type and quality.

Another example is temperature. The ideal gas thermometer is used to determine "fixed points" or temperatures of reproducible phenomena (e.g. the temperature of water at its triple point). The most precise practical method to measure temperature is the platinum resistance thermometer, which is used to interpolate between these fixed points. Error may arise due to an improper interpolation function or incorrect temperatures of the fixed points. A mercury-in-glass thermometer introduces both random error and bias resulting from the different vision levels of observers. An individual's perception of temperature would provide estimates similar to those found in the social and behavioral sciences. A third example is the measurement of pressure, where dead weight gauges provide the primary standard and such instruments as mercury manometers, Bourdon tube gauges and electronic strain gauges introduce various sources of error.

For such readily accessible physical properties, a relative "true" value and measures including various degrees of both random error and bias are feasible. Distributions of such measures could be used to provide a "real" estimate of the

impact measurement error has on statistics. Obviously, a comparatively large number of different subjects (whether clocks or people) will have to make estimates at specified times in order to develop distributions. These may show that hypothetical errors differ substantially from those of reality. Although more costly and not amenable to electronic recording, the introduction of human measurement errors would more closely approximate scales in many fields. Limited only by a researcher's creativity, statistical analyses of almost any type imaginable could be conducted on score distributions generated by these techniques and the relative influence of both measurement and statistical error evaluated based upon known true scores, bias and error sources.

References

- Ansell, M. J.G. (1973). Robustness of location estimators to asymmetry. Applied Statistics, 22, 249-254.
- Bickel, P. J. and Lehmann, E. L. (1975). Descriptive statistics for nonparametric models II. Location. The Annals of Statistics, 3, 1045-1069.
- Doksum, K. A. (1975). Measures of location and asymmetry. Scandinavian Journal of Statistics, 2, 11-22.
- Eisenhart, E. (1977). Discussion (of Stigler, 1977). Annals of Statistics, 5, 1085-1087.
- Gregoire, T.G. & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. Psychological Bulletin, 101, 159-165.
- Hill, M. and Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. Biometrics, 38, 377-396.
- Halliday D. & Resnick, R. (1974). Fundamentals of Physics. Wiley: New York.
- Rasmussen, J. L. (1989). Analysis of Likert-Scale Data: A Reinterpretation of Gregoire and Driver. Psychological Bulletin, 105, 167-170.
- Stigler, S. M. (1977). Do robust estimators work with real data? The Annals of Statistics, 5, 1055-1098.