The Gulliver Effect: The Impact of Error in an Elephantine Subpopulation on Estimates for Lilliputian Subpopulations

by:

Theodore Micceri, Ph.D., Pradnya Parasher, Ph.D., Gordon W. Waugh, Ph.D. & Charlene Herreid, Ph.D.

Paper Presented at the Florida Association for Institutional Research Annual Conference, Cocoa Beach, FL. Feb 25-27, 2009

Abstract

An extensive review of the research literature and a study comparing over 36,000 survey responses with archival true scores indicated that one should expect a minimum of at least three percent random error for the least ambiguous of self-report measures. The Gulliver Effect occurs when a small proportion of error in a sizable subpopulation exerts a large influence on one or more Lilliputian/small subpopulations. As a simple example, if three percent random measurement error occurs for a sample of 1,000 on a race/ethnicity item having two groups (white and minority), where the majority group make up 90 percent of the total population, then 27 majority subjects (three percent) will be erroneously classified as minorities and three minorities (three percent) will be incorrectly classified as majority. Assuming no missing data, the study will therefore report 876 majority respondents (873 majority + 3 incorrectly classified minority respondents) and 124 minority respondents (97 minority + 27 incorrectly classified majority). Although the incorrect majority respondents percentage is small ($\frac{3}{873}$ = 0.3%), the erroneous minority classification is substantial ($\frac{27}{97}$ = 27.8%). Such a large proportion of incorrectly classified respondents may alter estimates of measured differences between groups. The Gulliver Effect may occur whenever small subpopulations are of interest, whether this involves endangered species, rare diseases or unusual crimes such as the kidnapping of juveniles.

The Gulliver Effect: The Impact of Elephantine Subpopulation Measurement Error on Lilliputian Subpopulation Estimates

Introduction

Much social science research relies upon survey data. Takalkar*, et. al, (*1993) found in a study of 258 research articles in five flagship journals from three different fields - (a) Industrial-Organizational Psychology; (b) Management; and (c) Institutional Research -that between 36 and 76 percent of the articles in these journals used either survey or other self-report measures. Error estimates for surveys, when provided, usually represent a strictly theoretical approximation of sampling error[i], when, in fact, measurement error is almost surely the greatest problem in all social and behavioral sciences' research (Olson, 2006, p. 739). Numerous and various sources of error occur in measurements as the literature indicates: Fuller, 1987; Goyder, 1987; Wentland and Smith, 1993, Groves, 1989, etc.. Groves (1989, p. 17) lists twelve error sources present in all non-interview social science instruments, of which only three involve sampling error. Interviews add two additional error sources. Most researchers acknowledge the tentative nature of survey data, yet few conduct empirical investigations of survey items' response validity. Herein we primarily address only two of these nine error sources present in addition to the typically reported sampling error. We attempt to show that one can always expect at least a small proportion of non-sampling error from any self-report measure. Because of the Gulliver Effect, even a small proportion of error among large subpopulations can substantially impact data and possibly alter differences between majority and minority, or rare subpopulations in many common situations.

For continuous measures, classical measurement theory views obtained estimates as a combination of a true value plus noise or error. In the dichotomous or multichotomous situation, mismeasurement can no longer be viewed in this way, but rather as the probability of correct and incorrect classifications. In the dichotomous situation, say males and females, these would be respectively, the probability of a correctly classified male and a correctly classified female. In the current discussion, for the dichotomous situation, any incorrectly classified male or female would necessarily be a misclassified case for the alternative value.

This paper primarily involves a strictly mathematical phenomenon, the effects of which we hope to make clear through explanation and examples and a 50 year review of related literature.

Some Relevant Literature

Below are some studies where researchers compared survey responses to a reasonably accurate correctly classified value. In general, studies with any direct comparison between observed measures and what may be called 'true' values are rare. This occurs due to an inability to obtain such values for most variables that are measured by self-report. However, sometimes, organizational records can provide values for certain variables, such as criminal offenses, race/ethnicity and salary. Test retest studies, where responses at time one are compared with those at time two, are a less valid source of information.

Cannell and Fowler (1963) asked patients (_N_ = 586) to report the number of times they had visited a doctor in a two-week time period. The researchers compared the patients' responses with doctors' records. Results showed that there was 30 percent error in reporting the number of visits to the doctor. Most patients underreported the number of visits. Additionally, some unknown degree of interviewer errors also contributed to the total measurement error in this study.

Walsh (1967) asked college students (_N_ = 270 males) to report demographic information using three methods: interviews and two forms of questionnaires, all of which exhibited approximately the same error rate against personal records. Students were least accurate in reporting their high school GPA to a single decimal point (52% incorrect) and college GPA (49% incorrect). Students were most accurate in reporting the number of legal violations (2% incorrect) and the number of visits to Student Affairs for disciplinary reasons (3% incorrect), but this resulted primarily because these last two items almost never occurred. The most frequent response was zero, with very little variance.

Wyner (1980) compared the self-reported number of arrests with the actual number of arrests obtained from police records for a group of drug addicts (_N_ = 79). The mean error was 3.2 percent below actual. Although the average bias was small, the number of errors was high. Wyner reported that 87.5 percent of all respondents (69 of 79) gave incorrect responses. Another finding showed the influence of recall: "For each arrest that occurred before 1960...the odds are almost even that it will be reported or that it will be omitted" (Wyner, 1980, p. 171).

Laing, Sawyer, and Noble (1987) examined the accuracy of student-reported extra-curricular activities and special accomplishments in high school by comparing students' responses with information obtained from the high school staff (_N_ = 477). The typical rate of incongruent responses was about ten percent. The authors found a seven percent error rate on a race/ethnicity variable. Lacking electronic records, the researchers had high school staff complete questionnaires to obtain the validating information. This step adds error to the measures against which student responses were compared.

Based on their review of the literature, Means, Habina, Swan, and Jack (1992) concluded that estimates suggest a one to four percent underreporting of smoking among the general population and that this underreporting increases significantly to 15-20 percent among those smokers who are being heavily pressured to quit.[1] The researchers used relatively "objective" chemical indicators of smoking behavior, such as the level of cotinine in the blood.

Several studies report on incomes matched from surveys to tax records. Cristia & Schwabish (2007) found an average of more than 15 percent error between actual and survey-reported incomes for 140,259 records compared with IRS records, with males exhibiting greater errors than females. Moore, Stinson, & Welniak (1999) report a net underreporting of roughly four percent across all tax data, with higher incomes tending to under-report while lower income individuals tended to over-report.

[1] Takalkar, Waugh and Micceri (1993) also found 20 percent error where emotional pressure appeared to influence responses.

Takalkar, Waugh, & Micceri (1993) with a sample of 4,596, investigating survey responses on 36,061 items for which archival data were available regarding application and admission to a university, found 1,522 (4.2%) of the responses to be incorrect. Errors on the race/ethnicity variable were three percent (a form of test-retest error). Regarding the accuracy of their true scores, the authors note: "The data in these files were created in the admissions offices of the universities. Therefore, these data *should* be extremely accurate and were treated as 'true scores'."

Wentland and Smith (1993) evaluated data from 37 studies where survey responses were compared with records of behavior or characteristics for 56,701 respondents on 258 items. They found error percentages ranging from zero percent to 96 percent ("...an outlying value..."), with few items having more than 70 percent error, but many having 10 to 15 percent error.

The proportion of responses that were incorrect in the research cited above ranged from two percent to about 96 percent. The measures investigated by this research ranged from extremely simple, factual and permanent items (e.g., sex) to either items with substantial memory requirements (e.g., number of arrests) or items involving both precision and memory (e.g., High School Grade Point Average). In general, items involving more complex tasks (e.g., greater memory, greater precision, more decisions, etc.) were associated with a greater proportion of measurement error. Supporting this, Groves (1989, p. 407) states: "Cognitive processing appears to be central to survey measurement error." The preceding studies indicate that one should expect at least a small amount of measurement error to be present in all survey and other self-report items, even those that are simple, factual and permanent. Olson (2006, p. 739) concludes: "Comparisons of overall nonresponse bias and measurement error bias on survey statistics often show that measurement error bias is at least as large as nonresponse bias, if not larger, and that these non-sampling errors often far outweigh any sampling errors."

The Gulliver Effect

The Gulliver Effect occurs when a small proportion of error in a large subpopulation exerts a large influence on one or more small subpopulations. One common situation in which the Gulliver Effect could have a major influence on findings involves studies that investigate differences among racial/ethnic groups. As a simple example, if 900 of 1,000 respondents to a survey are majority, cases where majority respondents erroneously mark themselves as minorities may represent such a great proportion of the "apparent" minority respondents as to attenuate or exaggerate estimates of differences between or among the groups. If three percent random error occurs among the 1,000 respondents (27 errors from the majority and 3 errors from the minority), this would mean that 27 of the apparent 124 minority respondents (97 correctly classified minority respondents + 27 erroneously classified majority) are actually majority that erroneously marked themselves as minority. Such a large proportion of incorrectly classified individuals may diminish response differences between groups (e.g. ethnic groups) thereby creating false negatives. This particular race/ethnic proportion is not uncommon in social and behavioral science researches (see review in the section on reducing the Gulliver Effect's influence).

False positives may also occur when apparently significant differences between majority and minorities result not from actual differences between members of populations, but rather when a comparatively large number of the majority group sloppily completes surveys and is erroneously included in the minority groups. One might expect that people who make one error are likely to make more than one error (see Wyner, 1980). This means that apparent differences between majorities and minorities may only represent differences between those who carefully fill out forms and those who are careless. As an extreme, yet instructive example, suppose a survey involving ethnic groups asked the question "Upon what planet were you born?" We might expect a high proportion of those who incorrectly answer "Mars" to be among those also incorrectly marking themselves as minorities. Because the "sloppy" majority respondents would represent a comparatively large proportion of reported minorities, the researchers might falsely conclude that a significantly greater proportion of one or more minority groups were born on Mars than were whites.

A perhaps more likely situation occurs in the real world of America's War on Terror. If one tenth of one percent of America's 300 million population is, in fact, terrorist (300,000), and if the FBI has a classification measure that is 99 percent accurate, that would mean of our 300 million people, roughly three million Americans would be classified as terrorists, for a 90 percent misclassification rate. Unfortunately, this misclassification proportion appears to be close to true given the Justice Department's conviction rate in terrorist cases.

The presence of several different minority groups does not necessarily reduce this effect. For example, an analysis of 50 Midwestern Liberal Arts Colleges' 2005 enrollment distribution, using the IPEDS Peer Analysis System, showed 81.8 percent white, 4.9 percent unknown, 4.0 percent Alien, 3.7 percent African American, 2.9 percent Asian, 2.5 percent Hispanic and 0.3 percent Native American. A random measurement error of three percent among the 82 percent of whites would distribute 24 erroneous responses (.03 × 1000) plus 5.4 from the minorities equally across all groups (5 erroneous responses per group). These erroneous responses would represent a false misclassification for minority groups of 10 percent for unknowns, 20 percent for Hispanics, and 167 percent for Native Americans. Naturally, the smaller the proportion of a given Lilliputian subpopulation, the greater the impact of random error from the Gulliverian subpopulation and the greater the chance of erroneous interpretations for dependent outcome measures.

At this point, the issue of differential error rates in different subpopulations deserves attention. First, it is not likely that error rates will be identical for any two sub-samples as was suggested in the above example, and empirical evidence suggests that sometimes they may differ substantially (e.g. Abu-Sayf, 1999). Many studies report either differential error rates or different biases for various sub-samples, some examples include Haberman and Elinson, 1967; Oksenberg and Cannell, 1977; Wyner, 1980; Means,Habina, Swan, and Jack, 1992; Takalkar, Waugh and Micceri, 1993; Cash and McFadden, 1993; Abu-Sayf, 1999; Ostroff, Atwater and Feinberg, 2004; Cristia & Schwabish, 2007. Second, the U.S. Census Bureau has documented that response rates differ for different respondent groups, particularly minorities (Sweet, 1990), and in different geographic locales (Mihm, 2000). It would not be surprising if such differential response rates associate with different error rates. Supporting this, Olson (2006, p. 737) states: “We find that the relationship between nonresponse bias, measurement error bias, and response propensity is statistic-specific and specific to the type of nonresponse.” In situations where the Gulliver Effect occurs, the overall error rate will be determined primarily by the majority sample error rate. Where the error rate is greater among small subgroups, the misclassification effects will be magnified; where it is smaller, misclassification may be reduced, depending on the size of the sub-sample. Despite the preceding, for simplicity’s same, in most examples given herein, equal error rates are used for all subpopulations.

One might argue that variables such as race/ethnicity should show relatively small proportions of error. In fact, items that request current factual information do tend to show smaller error rates than more complex items involving judgment, opinion or memory processes. However, Laing, Sawyer and Noble (1987) found a seven percent error rate on their race/ethnicity variable. Takalkar *et al.* (1993) concluded: "We can safely say from this and preceding research, however, that even the least ambiguous self-report variables tend to exhibit between 3% and 5% error." Errors on variables such as race/ethnicity, which request current and factual information, may result from misreading forms, mismarking forms, errors in transferring information to an answer sheet, misunderstanding terms (e.g., many people who are born in the continental United States think of themselves as Native Americans), general sloppiness, fear of answering correctly, or merely a desire, either joking or malicious, to misinform. Several additional problems arise in a combined racial/ethnic variable, which the Census Bureau changed to a two part variable in 2000 (Hispanic or non, and race, including multi-racial), but which has remained a single variable for most researchers and institutions. Multi-racial individuals may answer one way at one time, and another at a different time or in a different situation, further complicating the process.

The examples of Gulliver Effect impacts provided above should be fairly conservative. That is, one can probably expect that random errors greater than three or four percent occur with some frequency on various classification items. Adding to this problem are data from studies such as that of Cash and McFadden (1993) in which the non-response rate among minorities was three times that for whites. This would decrease even further the proportion of correctly classified minority respondents. If, for example, we take the error rate of seven percent cited by Laing et. al. (1987) and the non-response rate of eight percent for whites and 24 percent for minorities cited by Cash and McFadden (1993), then the following scenario would occur with a two-group race/ethnicity variable for a sampling frame of 1000 with 90 percent whites (900) and 10 percent minorities (100). In this case, 828 responses (900 × .92) would be returned by whites and 76 by minorities (100 × .76). Among these responses, a seven percent measurement error would occur (Laing et. al, 1987). Thus, 58 (seven percent) of the 828 true white respondents would erroneously report themselves as minorities, and seven percent of the 76 true minority respondents (5) would report themselves as white. Therefore, the final reported numbers would be 775 whites (828 - 58 + 5) and 129 minorities (76 - 5 + 58). Of the total 904 respondents, 14.3 percent would be reported as minority (compared to the true value of 10 percent, or a 43 percent overreporting), and only 55 percent of the 129 reported minorities would truly be minority group members (71 of 129). This example shows that greater non-response rates by minority subpopulations, a not uncommon effect, can magnify the Gulliver Effect's influence. Further, this example is not a worst-case scenario. Abu-Sayf (1999) reported errors ranging from 16 percent to 42 percent in race/ethnic test-retest reporting among his subpopulations. This suggests that the proportion of respondents who represent such smaller subpopulations might be consistently over-reported in surveys due to purely random response errors by large subpopulations and greater non-response among minority groups. We might also note that even Census data involves primarily self-reports.

The Gulliver Effect is not limited to race/ethnicity. For any analyses made between or among different subpopulations, one of which is extremely large, researchers must be aware of the potential for such an effect. One example of the Gulliver Effect occurred in the Takalkar *et al.* (1993) study. Among the 33,667 items (93.3% of all items) where the respondent truly did not apply to a university other than the home institution, 1,210 of the items (3.6%) erroneously reported that the respondent was either accepted or denied at another university. This small error inflated the total apparent number who applied to other universities by 46 percent, from its true value of 2,394 to 3,495.

The Effects of Misclassification Errors

Few articles in the social and behavioral science literature deal with misclassification errors. Among those that do, most merely call attention to specific misclassification problems rather than presenting generalized methods to deal with them. For example, Feser and Pia (2002), dealt with Maximum Likelihood Estimators (MLE) in the context of logistic regression report: "…the MLE and the classical Rao's score test can be misleading in the presence of model misspecification which in the context of logistic regression means either misclassification's errors in the responses, or extreme data points in the design space." Lilienfeld, Alliger, and Mitchell (1995) indicate that "…errors in integrity testing reflect the systematic misclassification of some honest individuals as dishonest." Hilton (1995, p. 248) notes: "Failure to recognize the role of conversational assumptions in governing inference processes can lead rational responses to be misclassified as errors and their source misattributed to cognitive shortcomings in the decision maker." Ferguson, Horwood and Lynkset (1995, p. 384) suggests a method to convert continuous scales into dichotomous decision forms where he used a latent Markov model requiring at least three historic data points which took account of errors of measurement in the classification of children (see also Friedman, 1997). On this topic, Dwyer (1996, p. 360) outlines some problems that such dichotomizing creates when used for cut scores, which "…(a) always entail judgment; (b) inherently result in misclassification; (c) impose an artificial dichotomy on an essentially continuous distribution of knowledge, skill or ability; and that (d) no 'true' cut scores exist." Also discussing dichotomized cut scores, de Moor, Barnowski, Cullen and Nicklas (2003, p. 393) note: "Current methods of dietary assessment may not be reliable enough to attain acceptable levels of correct classification."

Other non-biometric disciplines also provide some examples, Kapteyn, Kapteyn, and Ypma (2007, p. 513) found "…very substantial biases…" when comparing misclassification between administrative and survey data in simple econometric models. Wiley and Martin (1999, p. 134) suggest "…incorporating misclassification errors into a latent class model for observed response states. We apply this model to survey responses dealing with government welfare programs and suggest that our approach can retrieve information where unidimensional and multidimensional models do not fit." The misclassification errors reported in these several studies can lead to incorrect or inaccurate conclusions if careful adjustments either *a priori* or *post hoc* are not made,

Perhaps the most damaging arena for the Gulliver Effect is false positive diagnosis of such as HIV/AIDS. For example, Kleinman, Busch, Hall, Thompson, Glynn, Gallahan, Wonby, and Williams (1998, p. 1082), using the gold standard Western blot test, found "The rate of false-positive Western blot results documented in this study was 0.00041% of all donations tested (95% CI, 0.00026%-0.00058%) and 4.8% of donations with results classified as Western blot positive." Placing this within context for a national sample, proportions of American populations that test positive for AIDS range from 0.008 percent for whites to 0.076 percent for African Americans (Kaiser Foundation, 2003). CDC (2003) estimates are that roughly 0.3 percent of 300 million Americans have HIV/AIDS with some 24-27 percent undiagnosed. The incidence rate reported by the Kaiser Foundation suggests that 76 of 100,000 African Americans and 8 of 100,000 whites have HIV/AIDS. However, the error rate of Kleinman, *et al* indicates that 41 false positives would occur among each 100,000 meaning that more than half of the African Americans and 41 of 49 whites would be misclassified. Other, quicker and less expensive HIV/AIDS diagnostic techniques provide far less accurate results (CDC, 2007). Given the social, emotional and health care costs of false positive HIV/AIDS diagnoses, the Gulliver Effect in this situation can prove extremely damaging.

Impacts of the Gulliver Effect on Statistics

Even when adjustments are made in an attempt to control for the Gulliver Effect (see the section on reducing the Gulliver Effect below), problems can arise in statistical analyses. Not wishing to delve too deeply into the vast statistical literature, nonetheless, measurements form the basis for statistical analyses, thus, any biases resulting from either direct measurement or from inaccurate adjustment necessarily impair statistical results. Among the several problems with commonly used statistical analyses are some fallacious assumptions which underlie much theoretical, simulation and empirical published work. Perhaps the most insidious fallacy is the concept of a symmetric error distribution around a population mean that underlies almost all measurement and statistical theory and practice. Modern perspectives on this derive from one of the original sources of modern science, which was repeated measurements of astronomical phenomena (Bessel, 1818). Assuming that no mechanical bias existed in the observation instruments (telescopes), then each observation contained both random and non-random error. When many such observations were compiled, it became apparent that the arithmetic mean was not an unreasonable estimate of center, because the observation errors in one direction tended to be offset by errors in the opposite direction. Thus developed the following misleading perspective, largely as a result of the work of Gauss, who sponsored Bessel and determined "Error distributions are symmetric in shape, with every positive error offset by a negative error of similar magnitude." This convenient assumption was originally developed partly because it fit the data, but largely because it fit the arithmetic mean, and thereby allowed for mathematically soluble formulae. It was pushed strongly into the academic mindset by Gauss and Galton (Seal, 1967). Although this idea may be true for repeated observations of the same object, it appears inherently absurd to over generalize repeated observation of a single object to single observations of many different objects. To keep this convenient assumption, measurement texts tell you that a measure's values reflect a distribution of "error" about a central true score. Thus a sample of the weights of 50 people represents the distribution of error around the one "true" human weight of the target population. However, this appears to be an unrealistic assumption for measures involving humans, or almost any other naturally occurring population. In any representative sample from a human population there must be at least two subgroups (male and female), each of which varies around a different center on almost any measure one can imagine. Additionally, other subpopulations always exist, each of which has a different center and spread in the subpopulation.

Geary (1947, p. 240) provides some history by noting that up to the end of the 19th Century, for several reasons, there existed a

> ...prejudice in favour of the hypothesis of universal normality...With the development about the beginning of the century, of the theory of moments, statisticians became almost over-conscious of universal non-normality. ...Our historian will find a significant change of attitude about a quarter-century ago following on the brilliant work of R.A. Fisher who showed that when universal normality could be assumed, inferences of the widest practical usefulness could be drawn from samples of any size. Prejudice in favour of normality returned in full-force and interest in non-normality receded to the background (though one of the finest contributions to non-normal was made during the period by R.A. Fisher himself), and the importance of the underlying assumptions was almost forgotten. Even the few workers in the field (amongst them the present writer) seemed concerned to show that 'universal non-normality doesn't matter': References (when there were any at all) in the text-books to the basic assumptions were perfunctory in the extreme. Amends might be made in the interest of the new generation of students by printing in leaded type in future editions of existing text-books and in all new text-books: *Normality is a myth; there never was, and never will be a normal distribution.*

Geary (p. 241) said the preceding sentence was an over-statement, due to a lack of evidence. The evidence has become available over the past 50 years. Following in the footsteps of numerous prior researchers, Micceri (1989, p. 161) found that among 440 large sample measures in Education and Psychology, where error is conceived in texts and theory as a symmetric and light tailed (Gaussian), that all were significantly non-normal at the alpha .01 significance level. Among achievement tests and gain score measures, 4.3 percent were relatively symmetric, smooth and unimodal with tail weights not extremely distant from those expected at the Gaussian. No psychometric measures (n=125) nor criterion/mastery tests (n=35) were in this group. Among these distributions, tail weights ranged from the uniform to the double exponential. Other non-Gaussian distribution properties that occurred included exponential level asymmetry, severe digit preferences, multimodalities and modes external to the mean/median interval. As is noted, these findings replicate those from empirical studies in every research arena ever investigated, the important factor being that measures in education and psychology are specifically designed to produce Gaussian type distributions. This was the last bastion of possible Gaussian distributions.

Another common fallacious assumption is that errors tend to cancel each other and zero out. In the hard sciences, where measurement errors can better be estimated, it is well known that they frequently are either asymmetric, or propagate rather than canceling each other across multiple measurements or measures (e.g. Taylor, 1982). Engineers who fail to properly consider this propagation effect, may find their bridges collapsing. One should expect similar, but greater effects to occur in the far less accurate and validated measures common to the social and behavioral sciences.

Regarding statistical analyses and Type I (alpha) and Type II (power) statistical errors, it is probably worthwhile to provide a short summary of what factors have long been known to impact statistical error. A variety of univariate, bivariate and multivariate distributional anomalies (non-Gaussian distributions) can sometimes severely impact obtained alpha and/or power. On this among the numerous studies that have been conducted, perhaps McClelland (2003) best elucidates the situation for Psychology by noting that statements in psychology statistics texts claiming Ordinary Least Squares (OLS) robustness in all but highly unusual situations have been disproved by more recent research. In either a two-group or multi-group situation statistically, unequal cell sizes create difficulties for OLS statistical techniques. In the two-group situation (***t***), adjustments for unequal ns such as Satterthwaite's approximation to the ***t*** (Satterthwaite, 1946) reduce power by reducing the sample n, but also serve to reduce Type I errors. Although most statistical textbooks provide examples having equal cell sizes for factorial ANOVA, because unequal ns cause problems (Blair, 1981; Ito, 1980). When heteroskedasticity is present, a common and 'massive' problem in ecological inference (King, 1997, p. 17), both Type I and II errors can differ substantially from nominal alpha or expected power (McClelland, 2003). Because the Gulliver effect, when present, may either increase or decrease the gap between two groups, with an increased gap, assuming no heteroskedasticity, the effect would falsely increase a test statistic's power to detect differences, and with a reduced gap, it would falsely reduce a test statistic's power to detect possibly real differences.

Researches involving small sub-samples deal primarily with the reliability of estimates, increasing this, or making adjustments to correct errors resulting from small sub-samples and are largely limited to Statistical and Health Sciences journals (see below). This small sample concern is well founded, because the Central Limit Theorem tells us that, as the sample size of the sampling distribution of OLS means increases to the asymptote, the distribution shrinks closer and closer around the parameter value μ. Recently published research on such topics as misclassification is largely absent from Psychology, Education and Sociology journals. Biometrics, following on the work of Karl Pearson (1895), has come closer to dealing with the way the real world works, unlike many other disciplines who have almost blindly adopted the Null Hypothesis Significance Testing paradigm (NHST). Several critiques and evaluations of the controversy exist, including, but not limited to (Rozeboom, 1960; Greenwald, 1993; Nix and Barnette, 1998, and Nickerson, 2000). Nickerson (p. 341) concludes "…that NHST is easily misunderstood and misused but that when applied with good judgment it can be an effective aid to the interpretation of experimental data." Unfortunately, many who apply NHST fall into the group that Rozeboom (1960, p. 416) describes: "…in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures learned more or less by rote from persons assumed to be more knowledgeable of statistics than he."

Unfortunately, Rozeboom's statement applies all too frequently, and it's effects on research, can be substantial. McClelland (2000, p. 393) notes: "Many psychologists, particularly those far removed from their last statistics course in graduate school, believe that the standard least-squares statistical procedures are relatively robust against all but the most serious kinds of nasty data."

Can One Reduce the Influence of the Gulliver Effect?

In this paper we focused primarily on a single type of error because it can have such large effects on estimates for rare subpopulations. This is important because frequently, rare subpopulations are of great interest to scientists, decision and policy makers, and therefore, to researchers in all disciplines. Unfortunately the Gulliver Effect can create a substantial proportion of erroneous classifications for small subpopulations, which can alter estimates of measured differences between specific subpopulations and detrimentally affect sometimes inappropriate statistical procedures. Gulliver Effects may generate either false positive or false negatives, and can thereby cause misguided judgments and policies regarding such differences. The implications surrounding the Gulliver effect are somewhat unnerving when one considers the instructive examples noted above.

At this point, the question arises: are any methods available that can reduce the detrimental influence of the Gulliver Effect on findings and conclusions? Several methods for such purposes appear in the literature, for example, the use of a gold standard cross validation using a carefully controlled small sample to test the accuracy of the primary sample and make adjustments if necessary. Several adjustment methods are considered or reported below.

One *a priori* historical method that can help reduce the impact of the Gulliver Effect is disproportional stratified sampling, where small groups are oversampled to assure reasonable numbers from Lilliputian subpopulations in the final obtained sample (Sudman, 1976). The fact that minority subpopulations tend to be less likely to respond or participate than majority populations makes this an even more important consideration in many situations (Goyder, 1987). Although well known among theoreticians, this technique does not appear to be widely used by practitioners. A systematically sampled review of 146 recently published articles (1998 through 2006) in flagship journals from Institutional Research (Research in Higher Education, N=46) and Industrial/Organizational Psychology (Personnel Psychology, Journal of Applied Psychology, N=100) found only two cases where disproportional stratified oversampling was employed. Both of those were studies employing data from a National Research

Database where oversampling was applied. This sample of articles verified the currency of Takalkar, Waugh and Micceri's 1993 study, as 75.6 percent of the articles used self-report measures, and 55.9 percent used surveys, so surveys occurred in 74% of articles using measures. By far the most common sampling technique was the convenience or availability sample (86.4% of samples), most frequently of students. Regarding the presence of Lilliputian subpopulations, in 90.3 percent of the situations where dichotomous or multichotomous values were reported, the majority segment of the sample comprised 84.6 percent or more of the total sample. Majority sample sizes ranged from a low of 40.6 percent to a high of 97 percent, with most being above 88 percent and only two near 85 percent. Clearly, situations where the Gulliver Effect may impact findings are not uncommon.

A second method to control effects, recommended by Gustafson (2004) involves using Bayesian adjustments with uncertainty through Markov Chain Monte Carlo (MCMC) sampling with probabilities varied across a range of possible values to obtain a more valid estimate of truth. Gustafson notes (2004, p. 131): "…the Bayes-MCMC analysis provides a straightforward route to statistical inference which accounts for misclassification of a binary explanatory variable." When one has an idea from historical research what a likely probability of a variable's cases will fall into one category or another, probabilities may be assigned differentially over a specific range. For example, regarding majority presence in a population, the naïve assumption is that the obtained percentages are true, however this is unlikely to be true for many reasons, including the Gulliver Effect. Assuming, for example, a three percent random error with an obtained majority presence of 90 percent one could have a best guess adjustment estimate of two point seven percent more respondents in the majority group and two point seven percent less for the minority group (as explained above). The assigned probabilities for the majority group could then be something like 92.7, 0.25, 91.7 and 93.7, 0.175, 90.7 and 94.7, 0.10, 89.7 and 95.7, 0.05, and so on to the end of a possible range. The situation becomes more complex in the multichotomous situation, although the use of multiple categorical variables is apparently not a difficult task in MCMC. Finally, Gustafson notes (2004, p. 131) that "…often something other than 'off-the-shelf' MCMC techniques may be needed for model fitting." This is certainly not a impossibly difficult method to help reduce the impact of category misclassification resulting from the Gulliver Effect among other error sources.

Another adjustment approach which Gustafson (2004, p. 104) critiques, is adjusting with false certainly, where one randomly reassigns an expected misclassified portion of the majority and minority population to the other population using MCMC. This approach, at least in theoretical simulations, is of questionable value (Gustafson, 2004, p. 104), because "…the reduction in bias achieved by adjusting for misclassification may not fully offset the increased variability incurred by admitting that misclassification is present. However, this should not be viewed as justification for using a naïve rather than an adjusted analysis."

Other approaches for making *post hoc* adjustments, some of which require sophisticated skills[ii], are explicated in: Lash, and Fink (2003), Rice, and Holmans (2003), Thurigen, Spiegelman, and Blettner (2000), Raudys and Jain (1991). Friedman (1997) proposes a method for a theoretically unbiased classification of continuous variables into groups.

Conclusions and Discussion

The Gulliver Effect is one specific type of misclassification, resulting as a straightforward mathematical consequence of extreme sub-sample disproportions. Although not limited to race/ethnicity, that situation is of importance because so many social and behavioral science and economic researches use race/ethnicity classifications as independent variables and as a proxy for affluence, and because some management decisions are based on self-reported race/ethnicity characteristics. The Gulliver Effect has its greatest impact where one or more subgroups make up a very small portion of the total population. Some extreme examples include rare minorities such as Native Americans, rare plants, rare animals, rare fungi, rare diseases, rare accidents, or any other phenomenon that makes up less than five percent of a population, like the kidnapping of juveniles which "…constitutes only one-tenth of 1 percent of all the crimes against individuals." (Finkelhor and Ormrod, 2000, p. 1).

An important question to ask is just how frequently and how influential the Gulliver Effect may be in empirical research? As was noted above in the review of recent journal articles, self-report measures are commonly used, and majority populations almost always comprise at least 85 percent of the total sample. What is perhaps most alarming about this review is the fact that so much research, and particularly so in Psychology, makes use of availability samples rather than the generally accepted random sampling techniques. Oversampling to reduce the Gulliver Effect's influence is almost non existent in the recent literature. Regarding how this may influence analyses and findings, the Gulliver Effect can bias results in a fairly wide variety of research situations, ranging from those using Logistic Regression to those applying Chi Square analyses. Admittedly, one hopes to have comparatively equal binary groups when conducting Logistic Regression or discriminant analysis which reduces the misclassification effects, however, in the Chi Square case, cell sample sizes of five (5) are usually considered adequate for a reliable estimate (Cochran, 1954). Further, in two-group or multi-group analyses, this effect will most likely exacerbate the detrimental effects of heteroskedasticity. As was noted in several examples, the Gulliver Effect can result in seriously inaccurate classifications. Supporting this position, Gustafson (2004, p. 37), regarding the relative influence of the same error rate for continuous and binary measures, states: "Having 10% of all the measurements 'entirely corrupted' in the binary case is clearly much more damaging than having all the measurements corrupted about 10% in the continuous case!"

This issue deserves more research attention. More studies directed specifically at determining the proportion of both large majority and minority groups that are typically misclassified by self-report techniques should be undertaken. Further, the impacts on outcome variables should also be addressed. Unfortunately, such studies require both extremely large samples and an available "true" value. A three percent error in a sample of 1,000 only amounts to 30 responses, and one cannot trust such a small number to reliably represent the proportion of misclassified individual. Ideally, samples of 10,000 to 100,000 would be used. Such studies should be conducted in a consistent fashion to cover every arena in where large majority populations occur in order to develop reasonable adjustment estimates such as that proposed by Gustafson (2004).

References

Abu-Sayf, F.K. (1999). *The integrity of ethnicity data*. Paper presented at the AIR Annual Forum, Seattle, WA. May 27-31, 1999.

Bessel, F. W. (1818). *Fundamenta Astronomiae pro anno 1755, Regiomonti 336*.

Blair, R. C. (1981). A reaction to “Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance.” *Review of Educational Research*, 51, 499-507.

Cannell, C., & Fowler, F. (1963). *A study of reporting of visits to doctors in the National Health Survey*. Survey Research Center, The University of Michigan, (mimeo).

Cash, R. W., & McFadden, K.S. (1993). *Who doesn't respond to applicant surveys? An analysis of respondent bias*. Paper presented at the AIR Annual Forum, Chicago IL, May 14-19, 1993.

CDC (2007). *Basic statistics*. Department of Public Health, Center for Disease Control and Prevention, Retrieved January 1, 2008 http://www.cdc.gov/hiv/topics/surveillance/basic.htm.

CDC (2003). *Using the BED HIV-1 capture EIA assay to estimate incidence using STARHS in the context of surveillance in the United States*. Center for Disease Control, Retrieved January 1, 2008 http://www.cdc.gov/hiv/topics/surveillance/resources/factsheets/BED.htm

Cochran, W. G. (1954). Some methods of strengthening the common chi-square tests. *Biometrics 10*: 417-451.

Cohen, S.B. (1997). An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data. *The American Statistician*, 51:3, 285-292.

Cristia, J., & Schwabish, J.A. (2007). *Measurement error in the SIPP: Evidence from matched sample administrative records*. Working Paper Series, Congressional Budget Office, Washington, D.D., Retrieved March 15, 2007 from http://www.cbo.gov/showdoc.cfm?index=7762&sequence=0

de Moor C.; Baranowski T.; Cullen K.W.; Nicklas T. (2003). Misclassification associated with measurement error in the assessment of dietary intake. *Public Health Nutrition*, Volume 6, Number 4, June 2003, pp. 393-399(7)

Dwyer, C.A. (1996). Cut scores and testing: statistics, judgment, truth, and error. *Psychological Assessment*, 8:4, 360-62.

Fergusson, D.M., Horwood, L.J. & Lynksey, M.T. (1995). The stability of disruptive childhood behaviors. *Journal of Abnormal Child Psychology*, 23:3, 379-396.

Feser, V. & Pia, M. (2002). Robust inference with binary data. *Psychometrika*, 67:1, 21-32.

Finkelhor, D. & Ormrod, R. (2000). Kidnaping of juveniles: Patterns From NIBRS. *Juvenile Justice Bulletin*. The Bulletin was prepared under grant number 98-JN-FX-0012 from the Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice. Retrieved March 10, 2007 from http://www.ncjrs.gov/html/ojjdp/2000_6_2/page2.html

Friedman, J.H. (1997). On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*. 1:1, pp. 55-77.

Fuller, W.A. (1987). *Measurement error models*. John Wiley & Sons: New York.

Geary, R. C. (1947). Testing for normality. *Biometrika*, 34, 209-242.

Goyder John C. (1987) *The silent minority*. Westview Press, Boulder, CO.

Greenwald, A.G. (1993). Consequences of prejudice against the null hypothesis. In Keren, G. & Lewis, C. (Eds) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Erlbaum, Hillsdale, NJ, 419-448.

Groves, R.M. (1989). *Survey errors and survey costs*. John Wiley & Sons, N.Y.

Gustafson, P. (2004). *Interdisciplinary statistics: Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman & Hall/CRC, Boca Raton, FL.

Hilton, D.J. (1995). The social context of reasoning: conversational inference and rational judgment. Psychological Bulletin, 118:22, 248-271.

Ito, P.K. (1980). Robustness of ANOVA and MANOVA procedures. In P.R. Krishnaiah (ed.), *Handbook of Statistics*, Vol. I, 199-236. New York: North Holland Publishing Company.

Kapteyn, A., Kapteyn, Arie; Ypma, Jelmer Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, 25:3, 513-551.

Kleinman, S., Busch, M.P., Hall, L., Thompson, R., Glynn, S., Gallahan, D., Wonby, H.E. & Williams, A.E. (1998). False-positive HIV-1 test results in a low-risk screening setting of voluntary blood donation. *Journal of the American Medical Association*, 280:12, 1080-1085.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, N.J.

Laing, J., Sawyer, R., & Noble, J. (1987). *Accuracy of self reported activities and accomplishments of college-bound students*. ACT Research Report Series 87-6. American College Testing Research Report Series, P.O. Box 168, Iowa City, IA 52243

Lash, T.L. & Fink, A.K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 14:4, pp. 451-458.

Lilienfeld, S.O., Alliger, G. & Mitchell, K. (1995). A commentary on W.J. Camara and D.L. Schneider's "Integrity tests: Facts and unresolved issues," which appeared in the American Psychologist, vol. 49, 1994, pp. 112--119. *American Psychologist*, 7: 457~458.

McClelland, G.H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. Reis, H.T. & Judd, C.M. *Handbook of Research Methods in Social and Personality Psychology*, Cambridge University Press, Cambridge, Eng., 393-411.

Means, B., Habina, K., Swan, G. E., & Jack, L. (1992). *Cognitive research on response error in survey questions on smoking*. Vital Health & Statistics, Series 6, Number 5. Hyattsville, MD: National Center for Health Statistics.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, 105:1, 156-166.

Micceri, T. (1987). *Testing for normality and evaluating the relative robustness of location estimators for empirical distributions derived from achievement tests and psychometric measures*. Unpublished doctoral dissertation, University of South Florida

Mihm, J.C. (2000). *Short- and long-form return and response rates*. U.S. General Accounting Office. Washington, D.C.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5:2, 241-301.

Nix, T. & Barnette, J.J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5:2, 3-14.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, 70:5, pp. 737-758.

Ostroff, C., Atwater, L.E. & Feinberg. B.J. (2004). Understanding self-other agreement: a look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology*. 57:2, pp. 333-375.

Pearson, K. (1895). Contributions to the mathematical theory of evolution - II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society*, A186, 343-414.

Raudys, S.J. & Jain, A.K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 13:3, pp. 252-264.

Rice, K.M. & Holmans, P. (2003) □*Allowing for genotyping error in analysis of unmatched case-control studies* □*Annals of Human Genetics*. 67:2, pp. 165–174.

Rozeboom, W.W. (1960) The fallacy of the null-hypothesis significance test. *Psychological Bulletin,* 57:5, 416-428.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2,:6, pp. 110-114

Seal, H.S. (1967). Studies in the history of probability and statistics. XV: The historical development of the gauss linear model. *Biometrika*, 54:1/2, 1-24.

Student [W. S. Gosset] (1908). The probable error of a mean. *Biometrika*, 6, 1-25.

Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

Sweet, J.A. (1990). Differential response rates of tertiary respondents. *NSFH Working Paper #25*. National Survey of Families and Households. Center for Demography and Ecology. University of Wisconsin-Madison.

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 14-19, 1993. (ERIC Document #ED360934)

Tapia, R. A. and Thompson, J. R. (1978). *Nonparametric Probability Density Estimation*, Baltimore: Johns Hopkins University Press.

Thurigen, D., Spiegelman, D. & Blettner, M. (2000). Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research*, 9:5, pp. 447-474.

Walsh, W. B. (1967). Validity of self-report. *Journal of Counseling Psychology, 14*, 18-23.

Wentland, E. J., and Smith., K. W. (1993). *Survey responses: An evaluation of their validity* . NY: Academic Press.

Wyner, G. A. (1980). Response errors in self-reported number of arrests. *Sociological Methods and Research, 9*, 161-177.

Wiley, J.A. & Martin, J.L. (1999) Algebraic representations of beliefs and attitudes: Partial order models for item responses. *Sociological Methodology*, 29:1, 113-.145.

[i] In this document, the term theoretical is used any time estimates are based on theoretical distributions such as the Gaussian, the Uniform, the Binomial or the Double Exponential Tapia and Thompson (1978). Micceri (1987, pp. 9) notes on the topic: “The complexity and lack of availability of real world data compels many researchers to simplify questions by retreating into either asymptotic theory or Monte Carlo investigations of ‘interesting’ mathematical functions.”

[ii] Cohen, 1997, reviews computer programs for sophisticated sampling and adjustments.