DOCUMENT RESUME

ED 360 934 HE 026 672

AUTHOR Takalkar, Pradnya; And Others
TITLE A Search for TRUTH in Student Responses to Selected Survey Items. AIR 1993 Annual Forum Paper.
PUB DATE May 93
NOTE 18p.; Paper presented at the Annual Forum of the Association for Institutional Research (33rd, Chicago, IL, May 16-19, 1993).
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Admission; College Applicants; College Bound Students; Comparative Analysis; *Error of Measurement; Higher Education; Institutional Research; Research; Research Methodology; Sampling; *Statistical Analysis; *Statistical Bias; Student Characteristics; *Student Reaction; *Surveys; *True Scores
IDENTIFIERS *AIR Forum; State University System of Florida; *University of South Florida

ABSTRACT

This study compared 4,594 student responses from three different surveys of incoming students at the University of South Florida (USF) with data from Florida's State University System (SUS) admissions files to determine what proportion of error occurs in the survey responses. Specifically, the study investigated the amount of measurement error in student responses to questions about application and admission activities to universities other than the one at which they were enrolling. A literature review is included that examines the problem of measurement error in other studies that used survey or self-report measures. The study found that considerable measurement error can exist in self-report measures even when a subject is reporting simple factual information; in this case the level of unbiased error was about 4 percent, and biased error was about 20 percent. The amount of error was directly proportional to the severity of the memory demands as well as the characteristics of the survey population. It was noted that with all but the smallest sample sizes, the measurement error will likely be larger than the sampling error, and that psychological forces, such as social desirability and cognitive dissonance, can create substantial measurement bias. Recommendations for future research are provided. (Contains 15 references.)

A Search for TRUTH in Student Responses to Selected Survey Items

by:

Pradnya Takalkar,
Gordon Waugh &
Theodore Micceri

Paper Presented at the AIR Annual Forum, Chicago, IL, May 15-19, 1993

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Third Annual Forum of the Association for Institutional Research held at the Chicago Marriott Downtown, Chicago, Illinois, May 16-19, 1993. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Chair and Editor
Forum Publications
Editorial Advisory Committee

A Search for TRUTH in Student Responses to Selected Survey Items

by:

Pradnya Takalkar,
Gordon Waugh &
Theodore Micceri

Paper Presented at the AIR Annual Forum, Chicago, IL, May 15-19, 1993

Much institutional research relies upon survey data. Error estimates for surveys, when provided, at best represent a strictly theoretical approximation of sampling error. Although most researchers acknowledge the tentative nature of survey data, few conduct empirical investigations of survey items' response validity.

The current study compares data from three different surveys of incoming students at the University of South Florida (USF) with data from Florida's State University System (SUS) admissions files, to determine what proportion of error occurs in the survey responses.

Few will deny that error is present in all measurements. Social and educational researchers are primarily interested in measuring underlying psychological states. The observed score on such measures always contains error. We assume that each observation represents both a "true score" and some "error" which distorts the accuracy of an individual observation. Classical test theory further assumes that the "errors" across a large number of observations are uncorrelated with true values, are not correlated with each other and are, therefore, randomly distributed with a mean of zero. Under this assumption, given enough observations, approximately half of the observed errors will fall above and half below the "true score". Conveniently, this means that errors will cancel each other out and the mean of many observations will equal the "true score". However, if errors consistently fall either above or below the true score, the observed scores are said to be biased. The mean of several observations, many of which contain bias, will not equal the true score. Thus, total error may involve two types of error: random error and bias.

Sources of error in social research include: (a) measurement error, (b) sampling error, (c) statistical error, and (d) errors in data entry and interpretation. If we can accurately estimate the amount of error in a measurement, we will have a better estimate of the range within which a true value falls.

Survey research is the most frequently used data collection technique in social research. Survey items ask respondents to report either their own behaviors, attitudes or background information; or to report the behaviors, attitudes, or background information of others. The former type of survey items come under the rubric of *self-reports* and are the most commonly used measures in survey research. In this study, we will focus on errors in self-report measures used in surveys. For the purpose of the current study, we will use the words survey and self-report interchangeably, since they both refer to information collected through self-reports.

To provide a quantitative estimate of the frequency with which survey or self-report measures are used in the literature relevant to higher education, we reviewed a total of 258 articles in six different journals, to determine the proportion of studies that used self-report measures. We chose flagship journals from three different fields: (a) Institutional Research; (b) Management; and (c) Industrial-Organizational Psychology. Journals from Management and I-O Psychology were selected because these fields are closer to the arena of institutional research, than, for example, Clinical Psychology.

Table 1 shows that from 38% and 93% of the articles in these journals use self-report measures as a basis of their research. Clearly, a large part of our research base depends upon self-reports. Other measures included variables such as performance appraisals, economic indicators, achievement indicators and productivity indicators.

Table 1
Survey of Journal Articles

Journal	N of Studies Reviewed	Studies Using Measures	Studies Using Self-Reports	Percent of Total	Percent of Studies Using Measures
Research in Higher Education	50	45	36	72%	80%
Journal of Applied Psychology	52	50	20	38%	40%
Personnel Psychology	50	41	38	76%	93%
Journal of Management	37	22	14	38%	64%
Academy of Management Journal	19	19	12	63%	63%
Administrative Science Quarterly	50	45	18	36%	40%

The Influence of Psychological Factors

The traditional model for responses to self-reports assumes that as long as the same question is asked of all respondents and the information is available to them, errors in response are random. The underlying assumption that most respondents are responding honestly appears on the surface to be fairly reasonable. However, researchers have come to believe that this model is too simplistic, and they have advanced more complex models of survey response. These models recognized that factors like comprehension of the question, cognitive processes of deciding what information is needed, and retrieval and organization of this information add complexity to the question-answering process. Within this model, many sources of respondent errors, at every stage of the process, contribute to measurement error. Again, if this error is random, classical test theory tells us that the long range expected value of the error is zero and that our measure will accurately determine a true score.

However, research has shown that assumptions of the randomness of error are often violated. For example, there is a tendency to overreport socially desirable behaviors and underreport undesirable behaviors (e.g., Parry & Crossley, 1950; Wyner, 1980).

> "If the event is perceived as embarrassing, sensitive in nature, threatening, or divergent from one's self-image, it is likely to either not to be reported at all or to be distorted in a desirable direction. ...Conversely, behavior perceived as desirable tends to be overreported" (Oksenberg & Cannell, 1977, p. 327).

Similarly, cognitive dissonance theory (Festinger, 1957) tells us that self-reports will reflect the attempts of individuals to change their attitudes or reported behaviors, to maintain consonance between their cognitive elements. Cognitive elements refer to the pieces of information retained by human beings in their memories. Cognitive dissonance theory assumes that inconsistency between cognitive elements is unpleasant, and that this inconsistency creates psychological tension called "cognitive dissonance". Individuals are motivated to reduce this tension and establish consonance. Dissonance reduction follows the path of least resistance, that is, cognitive elements least resistant to change are changed, to make them consonant with the other elements. For example, a heavy smoker, exposed to the literature on "smoking causes cancer", is likely to distort his/her perception of how much he/she smokes, and is likely to underreport the frequency of smoking behavior.

Measurement error and its estimation

In selecting studies for review, our first goal was to find studies similar to the current study, where researchers had compared survey responses to true scores. In general, studies with any direct comparison between observed measures and true scores are rare. The obvious reason for the lack of such research is the

indeterminacy of true scores for psychological constructs. Since true scores for psychological variables can only be inferred, their true scores are rarely available for comparison. On certain occasions, organizational records do provide true score measures of some socio-demographic and other variables. It is easier to assess the error in survey responses for such 'factual' variables, than for the purely perceptual or attitudinal variables so commonly used by social and behavioral researchers.

Due to a shortage of such studies in the literature, we then expanded the search to studies which reported some estimate of measurement error. All studies used one of three methods to estimate error, which are listed here in order of their validity: (1) comparison of self-reports with a 'factual' true score; (2) comparisons of information obtained from more than one data source; or (3) test-retest comparisons.

Studies comparing obtained scores with "True Scores"

Cannell & 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% error in reporting the visits to the doctor. An underreporting bias, where patients failed to report visits to the doctor, was most frequent. Additionally, some unknown degree of interviewer errors also contributed to the total error in this study.

Wyner (1980) compared the self-reported number of arrests with actual number of arrests obtained from police records for a group of drug addicts ($N = 79$) with known criminal records, who were working in a supervised employment program. The mean number of arrests was 9.25 and the mean number of reported arrests was 8.96, giving a mean response error of -.29, or 3.2%. There was a moderate negative relationship between response error and true score (r=-.39), which violates the assumption that "true scores" and error scores are unrelated. Underreporting of arrests increased as the number of arrests increased. Additionally, linking psychological factors to the error reports shows that, with increased intensity and undesirability of the crime, there was increased underreporting of arrests. Conversely, those subjects who had rated having a criminal record as moderately desirable or neutral overreported arrests by an average of four. Wyner reports that 87.5% of all respondents (69 of 79) gave incorrect responses and 72% of the respondents gave responses that were more than 20% from their actual value (e.g., reporting 7 arrests instead of 5 = + 40%). This study also 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." Another source of error in this study was digit preference: "...there are high frequencies of self-reports at $X_0 = 5, 10, 20$, which might be the result of a response tendency to pick convenient numbers...". The author finally notes that these findings: "...compel rejection of the third random error assumption-that errors are uncorrelated with each other. Errors are likely to be correlated between people to the extent that they share the same attitudes towards being arrested and have arrest histories that are similar in terms of their distribution in time and seriousness."

Walsh (1967) asked college students ($N = 270$ males) to report demographic information using three methods: interviews, and two forms of questionnaires. All three information sources used self-reported data. Self-reports were compared with data from the personnel records. Not surprisingly, no differences occurred in the accuracy of the information obtained from the three self-report methods. In comparison with personnel records, students were least accurate in reporting their high-school GPA (52% incorrect) and college GPA (49% incorrect). Students were most accurate in reporting the number of legal violations (2% incorrect) and number of visits to Student Affairs for disciplinary reasons (3% incorrect). Because these last two almost never occurred, the most frequent response was zero, with very little variance. In this study, students were asked to give an exact number in response to each question rather than presenting categories from which to select. This question format increased the memory and judgment demands on the respondents, which added to the error. More than any social desirability bias, we conclude that much of the error in this study resulted from the demand for detail.

In addition to the Wyner (1980) study, three other researches note bias in self-reports when factors such as social desirability or cognitive dissonance were clearly present. Parry and Crossley (1950) found overreporting of donations to the Community Chest, the frequency of voting, and whether the respondent possessed a library card or a drivers license. Oksenberg and Cannell (1977), citing another independent study,

reported that respondents overreported less serious diagnoses, like stomach ulcers, and underreported more serious or image damaging diagnoses, like genital and nervous system disorders.

Based on their review of the current literature, Means, Habina, Swan, & Jack (1992) conclude that estimates suggest 1-4% underreporting of smoking among the general population and that this underreporting increases significantly to 15-20% among those smokers who are being heavily pressured to quit. The relatively "objective" measure of smoking behavior used in these estimates were chemical indicators, such as the level of cotinine in the blood.

Studies using multiple sources of information

Methods used to gather evidence for the validity of psychological measures have benefited from the multitrait-multimethod approach (Campbell & Fiske, 1959). High correlations among multiple measures of the same psychological construct combined with low correlations among similar measures of different constructs provide evidence for the validity of the measures. However, this technique cannot assess an unmeasured "true score". Also, when such measures are subject to the same biases, (e.g., where all measures are self-report) correlations supporting a measure's validity tend to be inflated.

As a subset of the multitrait multimethod approach, researchers have sometimes used more than one independent source of data to estimate measurement error. In these studies, the percent disagreement between the multiple data sources produces an estimate of measurement error.

Laing, Sawyer & Noble (1987) examined the accuracy of student-reported extra-curricular activities and special accomplishments in high school by comparing the students' responses with information obtained from the high school staff. The typical rate of incongruent responses was about 10%. Lacking electronic records, the researchers had high school staff complete questionnaires to obtain the validating information. This step adds error to the "true score" against which student responses are compared.

Haberman and Elinson (1967) examined the agreement between husbands' and wives' reporting of household income. They had 13 income categories and found that 40.4% of the 602 couples who were separately interviewed gave different answers. Although 24.6% differed by only one category, 15.6% differed by two or more categories, which was a difference of over $2,000. To estimate the percent of error that would occur with fewer response categories, the researchers collapsed the categories and reduced the number of categories from 13 to 8. They found a reduction in error from 40.4% to 30.4%. When the categories were collapsed to 6, the error reduced to 23.9%. This supports the findings from the Walsh (1967) study, which found that attempts to obtain more detailed information increased error.

Test-retest comparisons

A more commonly used substitute for true scores involves the use of test-retest reliability estimates as an estimate of measurement error. However, these estimates only answer the question "how consistent is our measurement from time 1 to time 2?"; they fail to answer the question "how well are we measuring what we purport to measure?" Limited as they are, even studies estimating test-retest reliabilities are rare, given the high costs associated with survey research.

Porst & Zeifang (1987) investigated the test-retest reliabilities of the socio-demographic variables in the German General Social Survey (ALLBUS) (see Table 1). The purpose of the study was to see how consistently interviewees respond to the same question asked across relatively short intervals of time. The time interval between the first and the second interview was 1 month. Identifiable error in this study is primarily comprised of: (a) respondents' reporting error and (b) interviewer's recording error. There should also be some small amount of true change.

- Psychological factors such as social desirability and cognitive dissonance appear to add bias to measurement error (Means et. al, 1992; Oksenberg & Cannell, 1977; Parry & Crossley, 1950; Wyner, 1980).

Overall, responses involving simple, permanent, and factual data (e.g., age) show the smallest proportions of error (1-10%), while response tasks involving either multi-category responses or clearly involving psychological factors produce the greatest numbers of errors (15-50%).

The present study investigated the amount of measurement error in student responses to questions about application and admission activities to universities other than the one at which they were enrolling. Estimates of random error and bias were developed using both rational and quantitative grounds.

Method

Survey responses from 4594 students on their admission status at each of eight Florida public universities were compared with "true scores" derived from Florida State University System (SUS) admissions files to determine the magnitude and direction of respondent errors.

Subjects

The subjects consisted of 1527 First Time in College (FTIC) students entering USF in the Fall semester, 1991; 1793 FTIC students entering USF in the Fall semester, 1992; and 1274 transfer students entering USF in the Fall semester, 1992. Thus, the total sample was 4594 students. All of these students registered at the university's main (Tampa) campus. Only students who completed the USF Freshman or Transfer Student Survey questionnaire and wrote their correct Social Security Number on the survey form were included in the sample. These samples represented about 90% of incoming freshmen for both years and about 50% of incoming transfer students.

Materials

All subjects completed a survey questionnaire, only the first eight items of which relate to this study. These questions asked the students about their admission status at each of the other eight Florida public universities:

> For each of the following Florida public universities, indicate whether you either *did not apply, applied only*, or *applied and were accepted*. Please use the response alternatives that follow.
>
> A = Did not apply
> B = Applied only
> C = Applied and were accepted

Subjects recorded their answers on optical scanforms.

Procedure

In 1991, the University of South Florida (USF) Freshman Survey was administered to FTIC students attending one of USF's freshman orientation sessions. Several of these sessions took place during June, July, and August, 1991. The surveys were handed out to groups of about 30 students by the group leaders. Each student was given about 30 minutes to complete the survey and return it to the group leader.

In 1992, a similar procedure was followed for undergraduate transfer students. However, the 1992 Freshman survey was administered by the authors to groups of 150–300 students. This mass administration of the freshman survey was preceded by a 30-minute math placement test.

Table 2
Test-retest stability of socio-demographic variables in the German General Social Survey

Variable	Percent of responses that differed from time 1 to time 2
A. Variables with little or no true change and very few interpretational demands	
Gender	0.0
Marital status	1.9
Religion	9.1
Age	2.6
Education	6.6
B. Variables with some true change or interpretational demand	
Employment status	14.3
Current occupation	25.0
First occupation	46.6
Father's occupation	37.6
Occupational training	21.9
Monthly income	64.0

Items with very little or no expectation of true change in a one month time interval, like marital status, religion, age, and education, showed the lowest discrepancy in reports from time 1 to time 2 (see Table 1). For these types of items, discrepancies were on the order of 0-9%. When some degree of interpretational demands were made of the respondent, as in items concerning occupation & occupational training, the discrepancy in responses increased to an average of about 25%.

Additionally, when the number of response categories increased, the response error also increased. Two of the variables, father's occupation and first occupation, had 37 categories, and showed the lowest levels of agreements. This result indicates that the researchers' attempt to become too precise actually increased the magnitude of measurement error. This further supports the findings of Walsh (1967) and Haberman and Elinson (1967) that increasing the precision in reporting reduces precision in measurement. The question regarding monthly income displayed the lowest response stability (64% disagreement). This is not at all surprising, given that there is a lot of social desirability bias operating for this variable and possibly an occasional intention to hide the truth.

Boruch & Craeger (1972) examined the test-retest reliabilities of items on a freshman survey. The delay between tests was two to three weeks. Responses differed by 1.4% for region of birth to 12.5% for average high school grade. Responses differed only 1% to 5% from time 1 to time 2 for variables that had little or no probability of changing (e.g., region of birth, education level of parents, and distance of college from home). The highest instability was demonstrated by the average high school grade variable, with 12.5% of the responses differing from time 1 to time 2. The next highest instability was evident for parental income, which differed by 10.5%. Both of these variables had three things in common: (a) they had eight or more response categories; (b) they needed some degree of interpretation in responding; and (c) they were evaluative in nature and therefore subject to social desirability and other cognitive biases.

Findings from these studies indicate:

- All self-report variables contain at least a small amount of error.
- Attempts to increase the precision of measures tends to increase the proportion of error (Haberman & Elinson, 1967; Porst & Zeifang, 1987; Walsh, 1967). This suggests that the quantity of error relates to the cognitive complexity of the response task.
- Both classical test theory assumptions of error independence: (a) true and error scores are uncorrelated, and (b) error scores are uncorrelated with each other, have been shown to be untrue (Wyner, 1980).
- Memory influences the quantity of error (Wyner, 1980).

The majority of the students listed their social security number on the survey. We obtained the official admissions information for these students from the State University System of Florida's admissions files. These files contain, among other things, the final admission status of all students who applied to any of Florida's nine public universities. 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". We compared these true scores with the students' survey responses to the questions about their admissions status at the other eight SUS universities to determine the magnitude of error in our admissions survey data.

An examination of the frequencies from the SUS files of admission status by university showed that there were probably two different definitions of the "Withdrew Application After Admission Decision" admission status category among the eight universities. Two universities reported no applicants in this category, but they did report a relatively large number of applicants in the "Withdrew Application Before Admission Decision" category. Therefore, we dropped all the responses (n = 530 responses) that had an SUS admission status of "Withdrew Application Before Admission Decision."

The SUS files have more than the three admission status categories used in the survey, but the several SUS categories fit neatly into the three survey categories as shown below.

Survey Category	SUS Category
- Admitted	- Admitted
	- Provisionally admitted, non-exception
	- Provisionally admitted, exception
	- Withdrew application after being admitted
- Applied only	- Denied
- Did not apply	- missing in SUS admissions file

Analyses

The analyses compared the students' reported admission status (accepted, rejected, did not apply) with their official admission status (according to the SUS files) for each university. The number of matching and conflicting responses were counted. A sample of the completed answer forms were manually checked for response errors. It was verified that the SAS computer program had correctly identified the errors and correct responses in these forms. We also verified that the numbers of errors were being counted correctly.

A survey response was considered to be "correct" if it matched the SUS admissions file data, whereas a survey response was considered to be "wrong" if it did *not* match the SUS admissions file data. The number of correct survey responses was computed by summing the number of times a survey response matched the SUS admissions file data; the number of wrong survey responses was computed by summing the number of times a survey response did not match the SUS admissions file data. Unanswered survey items were omitted from all analyses.

Results & Discussion

The overall error rate was 4.2% for all students combined. That is, of the 36,061 items that were answered, 1,522 had wrong answers. The three groups had a similar overall error rate: 3.3% for the 1991 FTIC students, 4.1% for the 1992 FTIC students, and 5.5% for the 1992 transfer students. Table 3 gives more detailed frequency information concerning the types of errors and correct responses that were made.

While manually checking the forms, we noticed that one student had a pattern to the errors: the first four responses were wrong whereas the last four responses were correct. Perhaps this student realized, after completing the fourth item, that he/she was answering the items incorrectly but decided not to go back and correct these responses. Several other students might also have displayed patterns in their erroneous responses, but this issue was not extensively examined.

Table 3
Admissions Status Frequencies: Survey Responses vs. SUS Admissions File

Survey Response	SUS Admissions File Data		
	Accepted	Denied	Not Apply
1991 FTIC Students			
Accepted	509	75	170
Denied	7	282	112
Not Apply	7	25	10,846
1992 FTIC Students			
Accepted	661	76	237
Denied	13	394	192
Not Apply	18	42	12,428
1992 Transfer Students			
Accepted	146	20	289
Denied	12	90	209
Not Apply	5	12	9,183
Totals			
Accepted	1,316	171	696
Denied	32	766	514
Not Apply	30	79	32,457

Table 4 shows that the error rate differed depending upon the students' actual admission status. Students who were denied admission made the most errors (error rate = 24.6%) whereas students who either were accepted (error rate = 4.5%) or did not apply (error rate = 3.6%) made considerably fewer errors. The high error rate for students who were denied admission is consistent with cognitive dissonance theory (Festinger, 1957). In this study, the theory would state that students who believe that they are good students but were denied admission to a university experienced dissonance between their belief and their actual admission status. They cannot change the university's admission decision, but there are three other ways to eliminate the dissonance: (a) conclude that they are not good students, (b) conclude that sometimes schools *do* deny admittance to good students, or (c) think that they had actually been accepted by the university or that they had not applied there.

It appears that many students used the third method to resolve their dissonance. Among the students denied admission who responded incorrectly, more than twice as many students responded "Accepted" ($n = 171$) than responded "Did not apply" ($n = 79$). Finally, students can *reduce* dissonance by thinking that they were accepted by a prestigious school to which they never applied. We found that students *did* use this strategy. Our analysis focused on students who did not apply to a specific university but responded that they had been accepted there. These error rates were five times higher for the University of Florida and Florida State University, the SUS's two most prestigious universities (error rate = .054) than for the other six SUS universities (error rate = .011). Thus, these students could be denied admittance to a university and still believe they are good students if they report that they were accepted by a relatively prestigious university.

Table 4
Proportion of Responses That Were Incorrect

Student Group	SUS Admissions File		
	Accepted	Denied	Not Apply
1991 FTIC	.0268	.2618	.0253
1992 FTIC	.0448	.2305	.0334
1991 Transfer	.1043	.2623	.0515
Totals	.0450	.2461	.0359

unbiased error as shown below. The example shown below (which uses information from Table 3) is for the Denied admission category.

biased response error = total response error - unbiased response error

= .[(171 + 79) / (171 + 79 + 766)] - .0450
= .2461 - .0450
= .2011

The results for all three survey groups are shown in Table 6. Notice that there appears to be effectively zero bias among the students who did not apply. As one would expect, based on the premise of cognitive dissonance reduction, the level of biased error ranges from 1.5 to 10 times the level of unbiased error among students who were denied admittance. We appreciate that biased and unbiased error are probably not independent and are almost certainly not simply additive. However, for purposes of formulation, simplifying assumptions were used.

Table 6
Proportion of Biased and Unbiased Error

Type of Error	SUS Admissions File Accepted	Denied	Not Apply
		1991 FTIC Students	
Unbiased	.0268	.0268	.0268
Biased		.2350	-.0015
Total	.0268	.2618	.0253
		1992 FTIC Students	
Unbiased	.0448	.0448	.0448
Biased		.1857	-.0114
Total	.0448	.2305	.0334
		1992 Transfer Students	
Unbiased	.1043	.1043	.1043
Biased		.1580	-.0528
Total	.1043	.2623	.0515
		Totals	
Unbiased	.0450	.0450	.0450
Biased		.2011	-.0091
Total	.0450	.2461	.0359

How Much Error Should I Expect In My Measurement?

In this section our goal is to provide general guidelines to survey researchers about how much measurement error they can expect in their measurement.

As the literature just reviewed and common sense both indicate, self-report items having greater cognitive demands exhibit greater error than those having lesser cognitive demands. Factors which clearly influence the magnitude of response error include memory, emotional factors such as social desirability and cognitive dissonance, the number of response categories, and the precision of required discriminations. Items such as gender are low on all of the preceding factors, whereas items addressing issues such as pay and employment tend to be high on these factors. The degree of interpretational demands involved in a variable, and the potential for cognitive bias were used to categorize the survey variables from both the current study and past research. Survey variables were classified into two categories:

1. Unambiguous Variables where interpretational demands made on the respondent are **low** and the cognitive tasks involved in responding are not complex. Variables such as gender, marital status and age fall into this group. These variables are less likely to be influenced by cognitive and social biases.

The error rate was higher for students who applied to several Florida SUS universities. Table 5 shows that the error rate was directly proportional to the number of universities to which the student applied. This relationship might be due to a memory interference effect: it is more difficult to remember many things than few things. That is, a student who has applied to many universities would have more difficulty remembering which universities he/she applied to and his/her admission status at each one than a student who applied to only two universities. This result is consistent with the study by Wyner (1980), which found a positive correlation between the amount of underreporting and the actual number of arrests. Other factors that can affect memory accuracy–which were not addressed in the present study–include the importance of the event (and the importance of associated events), the recency of the event, and the amount of memory rehearsal (i.e., how often has the event been recalled).

It is unlikely that a fallible memory was a significant cause of all response errors. Notice in Table 5 that some students who applied to only two Florida SUS universities (i.e., USF and one other) still made a response error. It is hard to imagine that these students could not remember their admission status at only two universities, unless, of course, they had also applied to several universities outside Florida's State University System.

Table 5
Percentage of Respondents making at Least One Error by the Number of Schools Applied to

Number of Schools Applied to	1991 FTIC	1992 FTIC	1992 Transfer
1	13.4	14.3	25.2
2	24.9	26.9	32.9
3	33.9	29.5	32.6
4	40.0	41.3	40.0
5 or more	54.5	61.6	50.0

Respondents can exhibit both unbiased errors and biased errors. In this study, several factors (e.g., imperfect memory, carelessness, misunderstanding the item, poor hand-eye coordination) could have caused *unbiased* response errors whereas students might have committed *biased* errors when they did not want to admit (either consciously or subconsciously) that they had been denied admission to a university. The relationship between these types of error is represented in the equation below.

total response error = unbiased response error + biased response error

The estimation of these three values will be demonstrated using the data for the Totals of all three surveys (see Table 3). The number of errors is the sum of the off-diagonal values in Table 3 . The total number of responses is the sum of all the values. Therefore, the total response error was 1522 / 36061 = .0422. Unbiased response error can be estimated if we assume that there is no response bias when the student is accepted at a university. This assumption is supported by the almost equal number of Denied responses (32) and Did Not Apply responses (30) among students who were accepted. The preceding rationale produced an estimate of .045 for unbiased error as the calculation below shows.

unbiased response error = number of unbiased errors / number of unbiased responses

= (32 + 30) / (1316 + 32 + 30)
= 62 / 1378
= .0450

In order to estimate the proportion of unbiased error in other situations, we then assumed that this same level of unbiased error (4.5%) was present for all three SUS admissions categories (accepted, denied, not apply).

One would expect the amount of response bias to be greater among students who were denied admittance to a university than among students who did not apply. Therefore, the amount of bias was computed separately for these two categories (denied, not apply). As formulated, the amount of bias is simply the total error minus the

2. Complex Variables where interpretational demands made on the respondent are **not low** or the cognitive tasks involved in responding may be complex. Any variables that were not clearly in the low group were grouped here. This group includes variables such as reported monthly income and reported first occupation. Variables likely to be influenced by cognitive or social biases were also included in this category.

To identify Unambiguous and Complex variables, each author independently classified each variable from every study into categories of either Low or Higher complexity. Variables on which all three raters agreed are included in this report.

Among all unambiguous variables evaluated in the several studies, error rates ranged between 2% and 16%, with a median of about 7%. Complex variables, however, consistently exhibited error rates in excess of 10% ranging up to 53%, with a median value of 30%.[1] This suggests that researchers investigating self-report questions which are complex (not simple reports of current "facts") should expect considerable (average in the current study of 25%) error in their data. This error will probably reduce the power of statistical tests to identify "true" differences and relationships. Compounding this issue is the presence of considerable bias where emotional factors such as social desirability and cognitive dissonance influence responses. Biases on the order of 35% were found for some sub populations in the current study. Factors that create such biases may invalidate both descriptive and inferential statistical findings.

Although further research is clearly needed, the costs of obtaining "true" values will continue to deter such research. We can safely say from this and preceding research, however, that even the least ambiguous self-report variables tend to exhibit between 3-5% error, while those involving any complex factors increase this error substantially. Further, variables involving emotional factors such as social desirability and cognitive dissonance tend to "bias" results, which can cause either false negative or false positive results in research. Compounding the problem, even the most assiduous researchers will find it difficult to determine where and to what extent these factors come into play for specific variables or combinations of variables.

The Gulliver Effect, or Small Errors for Big Sub populations Become Large Errors for Small Sub populations

When small errors occur among a large sub-population, they can have a very large influence on smaller sub-populations. For example, if 900 of 1,000 respondents to a survey are white and one is particularly interested in the respondents who are minorities, cases where whites erroneously mark themselves as minorities may represent such a great proportion of the "apparent" minority respondents as to completely invalidate any comparisons. If 5% random error occurs among the 900 whites, this would mean that 45 of the apparent minority respondents were actually whites who were erroneously marked as minorities. This suggests that the proportion of respondents who represent such smaller portions may be consistently overreported in surveys due to purely random response errors by the larger proportions of the populations. In the current study, among the 33,667 who did not apply to another university, 1,210 (3.6%) erroneously reported they were either accepted or denied at another university. This small error increased the total apparent number who applied to other universities from 2,394 to 3,495, or by 46%.

Even worse, think of situations in which apparently significant differences between whites and minorities result not from true differences between actual members of the population, but rather from the relatively large number of whites who sloppily fill out surveys, produce unusual response patterns and are erroneously included in the minority groups during analysis. One would expect that people who make one error are likely to make more than one error. This means that apparent differences between whites and minorities might only represent differences between whites who carefully fill out forms and those who do not. For 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 who incorrectly mark themselves as minorities. Because these "sloppy" respondents would represent such a large proportion of reported minorities, this might lead to the false conclusion that a far higher proportion of minorities than whites were born on Mars.

Conclusions and Recommendations

Consistent with all previous studies of survey response error, the present study shows that considerable measurement error can exist in self-report measures even when a subject is reporting simple factual information. One should expect that greater levels of measurement error are present in attitudinal or evaluative self-report measures. In the current study, the level of unbiased error was about 4%, whereas the level of biased error was about 20% when conditions existed that encourage bias (i.e., when the respondent was denied admission to a university).

Supporting the findings of Wyner (1980), the study found that the amount of error was directly proportional to the severity of the memory demands. This conclusion is supported by the direct relationship between the error rate and the number of universities to which the students applied. A weaker piece of evidence for a memory factor is the low error rate for items in which the student did not apply to the university. One might argue that these items required the least memory load. For these items, the student needed to recall only one thing to correctly respond: "Did I apply to this university?". However, for the Accepted and Denied items, where the "accepted" error rate was slightly higher, the student must recall two things to correctly respond: (a) "Did I apply to this university?" and then (b) "Was I accepted?".

Characteristics of the survey population can affect the amount of error. In the present study, the amount of unbiased error for the transfer students was more than double the rates for the two FTIC surveys. These students might have been more careless or perhaps they applied to more non-SUS institutions than did the FTIC students (which would increase the memory demands). Alternatively, this may result from less stable estimates that occur with smaller samples, because far fewer transfer students applied to other universities than did FTIC students.

The results of the present study are extremely important considering the preponderance of self-report measures in social science research. All self-report measures that you use will likely have significant measurement error. With all but the smallest sample sizes, the measurement error will likely be larger than the sampling error. Psychological forces, such as social desirability and cognitive dissonance, can create substantial measurement bias. When possible, research should attempt to estimate the level of measurement error so that confidence intervals be constructed.

Future research should attempt to determine ways to minimize both unbiased and biased measurement error in self-reports. Possible paths of inquiry include: (a) what testing conditions reduce the carelessness of respondents, (b) how can memory demands be minimized, (c) what factors affect the clarity of a self-report item, (d) how can psychological forces that encourage response bias be minimized, and (e) how might measurement error can be estimated and incorporated into statistical analyses.

A very important fact to remember here is that the measurement error rates found in the present study are for "factual" types of data. We must expect such complex constructs as satisfaction and motivation to contain at least as much and almost surely more error than these factual variables. As noted earlier, from 38% to 93% of the articles in the five journals reviewed were dependent on self-report data, most of which are not of a purely "factual" nature. Furthermore, among the 7% (for Personnel Psychology) to 62% (for JAP) of journal articles not based on self-report measures, many involved either performance ratings or evaluations of product quality. Even the most naive researcher would expect measures such as these to contain substantial measurement error.

If we must expect errors that range from 3% to 50% and biases that range from 3% to 30% among our factual self-report data, what magnitude of errors can we expect for other, less easily defined constructs? How does this measurement error compare to sampling and statistical error? Generally, simulated studies of statistical procedures find errors of statistical validity to rarely approach 10% (Tan, 1981; Ito, 1980). Sampling errors also are rarely estimated to be extremely large, although these estimates are imprecise because of the difficulty in separating sampling errors from the types of non-sampling errors reported for "factual" data in this study. In any case, it appears that within the research error triangle (sampling, measurement, and statistical error), measurement error cannot be said to take a back seat to either of the other error sources, and should probably be considered the major culprit in the failure to identify "truth".

For those of you who have wondered why correlations between job satisfaction and performance, or between ACT scores and College GPA are not larger, the probable culprit is measurement error. True values and measured values can only rarely be expected to match very well. Avoiding completely the issue of bias, any analyses using two or more variables contain the error of all these measures, which may or may not cancel out, but almost certainly will increase the error variance of the measures. This increased error variance will, in turn, reduce the power of statistical tests to identify true relationships or differences. Even Wyner (1980), who found a reduced range (28 to 25) when comparing reported arrests to actual arrests, found an increase in variance from actual to reported ($s^2_{true} = 39.0$, $s^2_{observed} = 41.7$, or an increase of 6.9%). Does this mean that massive efforts should be undertaken to improve the quality of our measures? It will certainly be difficult to improve much on measures of such simple characteristics as gender and ethnicity; therefore, we may consider the findings reported here for the unambiguous variables (lowest error rate for a large sample 3.35%) to set something of a lower limit for measurement error.

One may conclude that the failure of many empirical studies to support hypotheses that appear obviously true (e.g., that job satisfaction relates highly to performance) results not from erroneous hypotheses but from a large amount of measurement error. The failure of empiricism to support "true" hypotheses, or to erroneously support "false" hypotheses may be one reason for the large number of persons who base decisions on personal information sources rather than empirical research. The only reasonable suggestion for improving the validity of research results which involve psychological constructs such as motivation and satisfaction appears to be that of Campbell and Fiske (1959). Despite the costs, the investigation of multiple traits using multiple measures in multiple studies appears to be the *sine qua non* of effective social science research.

References

Boruch, R.F. and Craiger, J.A. (1972). Measurement error in social and educational survey research. Office of Research, American Council on Education, 1 Dupont Circle, Washington D.C. 20036.

Bradley, J. W. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. Bulletin of the Psychonomic Sc 16, pp. 333-336.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

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

Festinger, L. (1957). A Theory of Cognitive Dissonance. Evanston, IL: Row, Peterson.

Haberman, P. W., & Elinson, J. (1967). Family income reported in surveys: Husbands versus wives. Journal of Marketing Research, _, 191-194

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures (Vol. VI) (pp. 199- 236). In P. R. Krishnaiah (Ed.) Handbook of Statistics. Amsterdam: North-Holland.

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

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.

Parry, H. J., & Crossley, H. M. (1950). Validity of response to survey questions. Public Opinion Quarterly, 14, 61-80.

Porst, R., & Zeifang, K. (1987). A description of the German General Social Survey test-retest study and a report on the stability of the sociodemographic variables. Sociological Methods & Research, 15 (3), 177-218

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. Communications in Statistics, A11, pp. 2485-2511.

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

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

---

[1] Most, but not all of these were comparisons with true scores.

Addendum
Decomposition of Error into Biased and Unbiased Components

One might argue that the low error rate for items in which the student did not apply to the university suggest that these required the least memory load. For these items, the student needed to recall only one thing to correctly respond: "Did I apply to this university?". However, when the student was actually Accepted or Denied (in this case, the "accepted" error rate was slightly higher) the student must recall two things to correctly respond: (a) "Did I apply to this university?" and then (b) "Was I accepted?". For those 1378 students who were actually accepted to another state university, we expect only random errors to occur. Based on the two sets of preceding assumptions, students who were accepted had two memory tasks, each of which should only have shown purely random error. First they had to remember whether they had applied to a specific university. There were 30 errors for this question, or 2.2% of the 1378 students. Then they had to remember whether they were accepted at the specific university. There were 32 errors for this question, or 2.4% of the 1348 students who had correctly answered the first question in their minds. This suggests that the purely random error rate when no reasons for bias exist is approximately 2.3% for a response task which involves only one memory task. The revised version of Table 6 below uses the preceding logic and error rates from the total to estimate the proportion of biased and unbiased error in the various responses. Note that for the Not Apply category, biased error remains essentially zero, but the small amount does approximately account for the 182 more students who said they were accepted than rejected. This does not change the overall biased error estimate.

Table 6
Proportion of Biased and Unbiased Error

Type of Error	SUS Admissions File: Accepted	Denied	Not Apply
		1991 FTIC Students	
Unbiased	.0268	.045	.023
Biased		.217	.002
Total	.0268	.262	.025
		1992 FTIC Students	
Unbiased	.0448	.045	.023
Biased		.186	.010
Total	.0448	.231	.033
		1992 Transfer Students	
Unbiased	.1043	.045	.023
Biased		.217	.029
Total	.1043	.262	.052
		Totals	
Unbiased	.0450	.045	.023
Biased		.201	.013
Total	.0450	.246	.036