A COMPARISON OF SOCIAL DESIRABILITY BIAS AMONG
FOUR WIDELY USED METHODS OF DATA COLLECTION
AS MEASURED BY THE IMPRESSION MANAGEMENT SUBSCALE
OF THE BALANCE INVENTORY OF DESIRABLE RESPONDING

A dissertation submitted to the
Kent State University College and Graduate School
of Education, Health, and Human Services
in fulfillment of the requirements
for the degree of Doctor of Philosophy

by

John C. Rossiter

May, 2009
A dissertation written by

John C. Rossiter

B.B.A., Kent State University, 1986

M.Ed., Kent State University, 1993

Ph.D., Kent State University, 2009

Approved by

_________________________, Co-director, Doctoral Dissertation Committee
Rafa Kasim, Ph.D.

_________________________, Co-director, Doctoral Dissertation Committee
Shawn Fitzgerald, PH.D.

_________________________, Member, Doctoral Dissertation Committee
Milton Harvey, Ph.D.

Accepted by

_________________________, Interim Chairperson, Educational Foundations and
Melody J. Tankersley, Ph.D. Special Services

_________________________, Dean, College and Graduate School of
Daniel F. Mahony, Ph.D. Education, Health, and Human Services
A COMPARISON OF SOCIAL DESIRABILITY BIAS AMONG FOUR WIDELY USED METHODS OF DATA COLLECTION AS MEASURED BY THE IMPRESSION MANAGEMENT SUBSCALE OF THE BALANCE INVENTORY OF DESIRABLE RESPONDING (194 pp.)

Co-Directors of dissertation: Rafa Kasim
Shawn Fitzgerald

Four different data collection methods (face to face interviews, telephone interviews, mail surveys, and Web-based surveys) compared Social Desirability Bias (SDB). This study used Multiple Comparison tests and a randomized post-test only control-group design. No SDB differences were found among methods. For SDB no gender (2-factor), ethnicity (2-factor), nor was a gender and ethnicity (3-factor) interaction found. This study was more rigorous than other studies because 13 known extraneous influences were controlled for and two more were avoided. Effect sizes ranged from .002 to .029. Implications are: SDB need not trouble researchers when comparing the 4 methods and SDB may be decreasing over time. Suggestions for future research include (1) studies between non-published and published studies; (2) meta analytic method comparisons over well-established constructs; (3) meta analytic studies on SDB over time; (4) measurement invariance of the 4 methods on SDB; and other suggestions.
DEDICATION

To honor my Mom: who instilled in me by example the virtues of tenacity, being gutsy, and “can do it” versatility.
ACKNOWLEDGMENTS

Successfully completing this dissertation was truly a “team” win. I clearly could not have done it alone. Sure, I feel blessed to earn my Ph.D. However, I feel even more blessed to have had the opportunity to successfully work with so many talented, dedicated, accommodating, and supportive individuals.

To all my financial contributors I thank you wholeheartedly. Among the contributors were: HallmarkChannel.com, especially Father Bob Bonnot; the Portage County Board of MR/DD, especially Jim Rock and Gina Brown; the Coleman Foundation, especially Camille Fauser; Miller’s Rental & Sales, especially Dan Lipka and Dan Craig; Denise “Nic” Norman; Laura McFarland; Suzie and Mark Dominick; Ken and Mara Byers; and Jennifer Stackhouse. However, my primary financial contributor was my Mother, Su Sa Rossiter. She made certain I had good food, tons of toner, and reams of paper for my printer. She also made certain we took care of always paying my bills ahead of time and under-budget, caregiver costs, employer taxes and reporting, my medical needs, and the other “mundane”, but necessary activities of daily living. I can only hope I can “live up” to her finely honed skills in “managing the mundane” in the years ahead without her.

I must thank all my data collectors who took their valuable time to help me collect the data I needed. They were Kara Buell, Nan Garrison, Christine Hudacek, Mary Katherine Rattai, Dr. Ryan Wooley, Ph.D., Rodney Adolphson, Laura Moore, Rosie Glidewell, Denise Wright, Claudia Phipps, Cindy Zronek, Dr. Steve Denney, Ph.D., Dr.
Rachel M. Queen, Ph.D., Elizabeth Tussey, Sue Brown, and Megan Riley. Although they
gave their valuable time, their diligence, patience, goodwill, and friendship were and
always shall be invaluable to me.

I also had awesome office assistants assisting me. Two people deserve extra
praise for their excellent work: Nan Garrison, my typist and confidant; and Denise (Dee)
Wright, my Editor-in-Chief, software expert, and my Miss Reliable. I must also thank all
my other office assistants: Jennifer Powell, Erin Denney, Elizabeth Doerr, Amanda
Surace, Jennifer Stackhouse, Margie Kierstead, Miranda Monroe, Alecia McNeely,
Claudia Phipps, Dianna Duff, Elizabeth “Liz” Tussey, Heather Wilson, Jacquie
Blanchard, Jillian Liff, Kara Wilson, Laura Moore (my Perky Office Assistant),
Kimberly Moody, Lindsey Murphy, Nicole Davis, Stephanie Berenthal, Walter Lang,
Rachel Nordhoff, Olivia Colony, Fred Minkowski, Heidi Rolf, and Danielle Eckert. A
special thanks must go to my brother, James Rossiter, and my dear friend Jessica Jones
who kept me motivated, grounded, and my office well-stocked. Of course my Mother, Su
Rossiter, also helped with some office work, but was especially adept at keeping me well-
grounded as only a great Mom can.

I must praise the goodwill and professionalism of some Kent State professional
staffers for their assistance, patience, and willingness to go the extra “ten miles” to help
my research project and subsequent logistical efforts succeed. They were Jacquie
Parsons, William Bulger, Wayne Schneider, Jill Church, Dr. Ray Leone, D.O., Dr. Scott
Olds, Ph.D., and Dr. Nancy Barbour, Ph.D. Furthermore, high praise indeed deservedly
goes to all the library personnel who photocopied and sometimes located the over 1,500
sources (We wish this number was an exaggeration!). Specific praises go to Sue Lipiec, Raida Gatten, Jessica Gardner, Colleen Chalfant, and Leigh Mauer.

I deeply appreciate the moral support and cajoling of Dr. Kim Sebaly, Ph.D., Dr. Bruce Dockstader, Ph.D., Christina Ughrin, Dr. Karen Flemming, Ph.D., Dr. Paul Schoenberg, Ph.D, Dr. Rachel M. Queen, Ph.D., Emily Parker, Nan Garrison, Dr. Deborah Webster, Ph.D., Sue Lipiec, Carrie Giovannone, Katherine Morgan, Dr. James Webb, Ph.D., Alexa Wanchick, Dr. Babette Cameron, Ph.D., Dr. Nancy Barbour, Ph.D., Dr. Mary Hennessey, Ph.D., Dr. Phillip Rumrill, Ph.D., Dr. Doreen Larson, Ph.D., Dr. Emily Kraft Ray, M.D., John Kindelberger, Thelma Willis, Jessica Jones, Sharon Brown, Dale Liff, Jillian Liff, Laurel Myers Hurst, and Sharon Myers. When I needed their willing ear, sound advice, encouragement, and sometimes even a strong, swift “kick in the pants,” they eagerly provided them.

When I had questions about Kent State, survey research, or a minor statistical issue, I consulted with Kent State officials, local and national survey researchers, and statistical consultants. When I needed to know something, they either pointed me toward the answer or answered my question directly. They were John Kindelberger, John Bushery, and Thelma Willis of the U.S, Census Bureau, Sharon Brown of Cleveland State, Michelle Henry of the University of Akron, Eileen Doherty and Dr. Doreen Larson, Ph.D. of Lakeland Community College, Dr. Michael Humphrey, Ph.D., Brian McDonald, MBA, and Christina Ughrin. When I needed general information about Kent State’s students, William Bulger, Wayne Schneider, Linnea Carlson, Deb Shelstak, and Brian McDonald were extremely helpful and timely with my information requests. Dr.
Delroy Paulhus, Ph.D., from the University of British Columbia in Vancouver and Dr. Don Dillman, Ph.D. of the University of Washington were helpful in sending some articles, instrument manual, and providing bits of their input into my study.

Finally, I must thank my Dissertation Committee members. Dr. Rafa Kasim’s suggestions, input, expertise, and good humor were a real blessing. He helped me make my study and my manuscript much stronger. Dr. Shawn Fitzgerald’s patience, excellent writing skills, and writing advice were invaluable. Dr. Milton Harvey was always encouraging and pushing me toward my limits, as he has a reputation for doing with all his students. His “let’s get it done” attitude toward me and the whole Committee were a pivotal catalyst toward finishing this “project”. His sense of humor, eagerness to impart and share his knowledge and his “unwillingness to be torn away from his keyboard” has always inspired me that I can, should, and would do better.

As stated earlier, this research project succeeded because I had the honor of working with a great team. It is my dissertation only in name. It is actually a “team win” in spirit. Because of everyone’s help, including the Lord’s, it is done.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xii</td>
</tr>
<tr>
<td>CHAPTER I: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A Brief History of Survey Research</td>
<td>1</td>
</tr>
<tr>
<td>Social Desirability</td>
<td>2</td>
</tr>
<tr>
<td>Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>Rationales for the Study</td>
<td>6</td>
</tr>
<tr>
<td>Research Questions</td>
<td>8</td>
</tr>
<tr>
<td>Research Design</td>
<td>12</td>
</tr>
<tr>
<td>Research Hypotheses</td>
<td>13</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER II: REVIEW OF THE LITERATURE</td>
<td>19</td>
</tr>
<tr>
<td>A History of Survey Research Methods</td>
<td>19</td>
</tr>
<tr>
<td>The Advantages and Disadvantages of Different Survey Methods</td>
<td>27</td>
</tr>
<tr>
<td>Threats to the Validity of Surveys</td>
<td>39</td>
</tr>
<tr>
<td>Social Desirability Bias Defined and How it is Related to Survey</td>
<td>46</td>
</tr>
<tr>
<td>Specific Threats to the Validity of this Study as Determined by a Review of the Literature</td>
<td>63</td>
</tr>
<tr>
<td>CHAPTER III: METHOD</td>
<td>69</td>
</tr>
<tr>
<td>Research Design</td>
<td>69</td>
</tr>
<tr>
<td>Participants</td>
<td>69</td>
</tr>
<tr>
<td>Sampling</td>
<td>70</td>
</tr>
<tr>
<td>Procedures</td>
<td>73</td>
</tr>
<tr>
<td>Instrument</td>
<td>73</td>
</tr>
<tr>
<td>Reliability and Validity of the IM subscale of the BIDR</td>
<td>74</td>
</tr>
<tr>
<td>Additional Questions Other Than the IM Subscale</td>
<td>76</td>
</tr>
<tr>
<td>Scoring</td>
<td>78</td>
</tr>
<tr>
<td>Settings of the Study</td>
<td>78</td>
</tr>
<tr>
<td>Web Site Considerations</td>
<td>80</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>80</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>81</td>
</tr>
<tr>
<td>How Non Response Was Addressed</td>
<td>81</td>
</tr>
<tr>
<td>Procedure</td>
<td>82</td>
</tr>
<tr>
<td>Data Analyses</td>
<td>88</td>
</tr>
<tr>
<td>IV RESULTS</td>
<td>96</td>
</tr>
<tr>
<td>Sampling Data Characteristics</td>
<td>96</td>
</tr>
<tr>
<td>Sample Characteristics</td>
<td>97</td>
</tr>
<tr>
<td>Results of Research Questions</td>
<td>98</td>
</tr>
<tr>
<td>Possible Threats to Validity—Results of Controlling Analyses</td>
<td>101</td>
</tr>
<tr>
<td>Falsification</td>
<td>101</td>
</tr>
<tr>
<td>Summary of Falsification Results</td>
<td>103</td>
</tr>
<tr>
<td>Extreme Response Style</td>
<td>105</td>
</tr>
<tr>
<td>Summary of ERS Results</td>
<td>109</td>
</tr>
<tr>
<td>Supplemental Analyses and Results</td>
<td>112</td>
</tr>
<tr>
<td>A Summary of the Results</td>
<td>113</td>
</tr>
<tr>
<td>Answers to Research Questions</td>
<td>113</td>
</tr>
<tr>
<td>V DISCUSSION</td>
<td>115</td>
</tr>
<tr>
<td>Research Hypotheses</td>
<td>115</td>
</tr>
<tr>
<td>Controlling Hypotheses</td>
<td>117</td>
</tr>
<tr>
<td>Integration of this Study Results with the Literature on Method Effects</td>
<td>119</td>
</tr>
<tr>
<td>Convergent Findings</td>
<td>119</td>
</tr>
<tr>
<td>Divergent Findings</td>
<td>120</td>
</tr>
<tr>
<td>Implications of Findings</td>
<td>120</td>
</tr>
<tr>
<td>Delimitations</td>
<td>123</td>
</tr>
<tr>
<td>Limitations</td>
<td>124</td>
</tr>
<tr>
<td>Suggestions for Future Research</td>
<td>126</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>131</td>
</tr>
<tr>
<td>APPENDIX A DEMOGRAPHICS</td>
<td>132</td>
</tr>
<tr>
<td>APPENDIX B SURVEY AND INTERVIEW QUESTIONS</td>
<td>135</td>
</tr>
<tr>
<td>APPENDIX C RESULTS OF DATA ANALYSES</td>
<td>139</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>150</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Depiction of social desirability effect size differences among different methods</td>
<td>71</td>
</tr>
<tr>
<td>2.</td>
<td>Ethnicity Percentage Breakdown</td>
<td>133</td>
</tr>
<tr>
<td>3.</td>
<td>Age Range of Analyzed Participants during the Fall 2006 Semester</td>
<td>134</td>
</tr>
<tr>
<td>4.</td>
<td>Survey and Interview Questions</td>
<td>136</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The Mean Difference Between Methods</td>
<td>140</td>
</tr>
<tr>
<td>2. Means and Standard Deviations of Caucasian and Non-Caucasian Participants</td>
<td>141</td>
</tr>
<tr>
<td>3. Duration of the Face to Face Interview Across Interviewers</td>
<td>142</td>
</tr>
<tr>
<td>4. Duration of Telephone Interviews Across Interviewers</td>
<td>143</td>
</tr>
<tr>
<td>5. Eliminating or Controlling Extreme Response Style (ERS) Bias</td>
<td>144</td>
</tr>
<tr>
<td>6. Response Rate Information by Method</td>
<td>146</td>
</tr>
<tr>
<td>7. IM Summary Statistics Across Methods and Questions</td>
<td>147</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

A Brief History of Survey Research

The word “survey” has Latin and Old French origins (Converse, 1987).

It meant seeing over [italics copied], in the sense of a broad view or scan . . . ‘to stand on a hill and survey the surrounding country’ as Webster put it in his 1838 edition [dictionary] for Americans . . . . There was another connotation, that of counting or measuring, establishing value or extent (Converse, 1987, p. 18).

However, the word “research,” which also originates from Old French, means “to look again . . .” (Aiken, 1997, p. 13).

Survey research methods have improved throughout history, and the pace of this evolution has risen dramatically in the twentieth and now the twenty-first century—much like the pace of modern technology. If a census can be called a survey, the Babylonians were conducting surveys since about 2300 B.C.E. The ancient Jews had a “shekel” census in which each person aged 20 or above was to donate a half shekel to a sanctuary and half a shekel to the Lord (Exodus, 30:11-14, [Today’s English Version]). In this way they could keep track of adults and revenue (Halacy, 1980). Sir Francis Galton administered a lengthy survey to eminent scientists of the 19th century to study mental giftedness (Aiken, 1997). Political polling and market research have contributed heavily to refinements in and developments to survey research methods since the 1890s. World War II accelerated the collaboration between disciplines and industries in producing the Strategic Bombing Surveys and the various morale surveys.
The invention and improvements of computers further accelerated the progress in survey research. As Aiken (1997) noticed, the synergy between the escalating speed and versatility of computers, statistical theory, and improved questionnaire design made the decision processes of organizations and agencies easier and more accurate. Survey research has come a long way from counting shekels. While face to face interviews and written surveys are still used, the more recent inventions such as the telephone, computer, and the World Wide Web are being used with greater frequency because of their efficiency (e.g., time for administering, collecting, analyzing, etc.), and cost savings.

Social Desirability

Survey researchers could get more bang for their research bucks if they could rely on an accurate ordering of methods of data collection. A problem with survey accuracy is that it is multifaceted. A major, specific threat to the accuracy of surveys is Social Desirability Bias (SDB) (Aiken, 1997; Bradburn and Sudman, 1983; Belson, 1986; Dillman, 2000; Dillman and Tarnai, 1991; Fowler, 1985; Groves, 1989; Turner, 1984; Mirowsky and Ross, 1984; Smith, 2004). All of the preceding authors, without exception, listed SDB as a major source of bias in survey research. Yet, these same authors could only sporadically point out other different sources of bias that might occur in conducting survey research, in addition to SDB bias. Such sources of bias might include acquiescence, question ordering, telescoping, respondents’ failure to understand the question as the researcher intended, lack of knowledge, difficulties with recalling, participants’ over-cautiousness, extreme response style, primacy effects, recency effects, top of the head effects and, other minor biases. Nagao and Martin (as cited in Rosenfeld,
Booth-Kewley, Edwards & Thomas, 1996, p. 264) defined social desirability as “the tendency to stretch the truth in an effort to make a good impression,” (p. 72). Vogt (1999) in his “Dictionary of Statistics and Methodology” defined SDB as “Bias in the results of interviews or surveys when the respondents are trying to answer as ‘good’ people ‘should’ rather than in a way that reveals what they actually believe or feel” (p. 268). These definitions coincide with each other, yet Martin and Nagao’s definition is simpler, captures more of the flavor of what is meant by the literature, and is less subject to ambiguity.

**Problem Statement**

There has been a plethora of studies of method comparisons concerning SDB with most of them comparing only two methods of data collection at one time. There have only been five studies comparing three methods of collecting data including self-administration (mail-out/mail back administration), interviewer administered (face to face), and telephone interviewing (De Leeuw, 1992; Hochstim, 1967; Thornberry, 1976; Verrips et al., 2001; Wiseman, 1972). However, Richman, Kiesler, Weisband and Drasgow (1999) conducted a Meta Analysis comparing self-administered, computerized questionnaires, face to face interviews, and mail questionnaires. More importantly, to date there has been no comparison of (e.g. face to face interview, telephone interview, mail survey, and Web-based survey). Even the three method comparisons mentioned above did not use multiple-comparison procedures, which are appropriate to the problem, as this study did. The MCP’s were appropriate because they controlled for family-wise Type I error. This study explored the differences in data collection methods in terms of
SDB. However, it was not possible for this study to rank order survey methods in terms of SDB.

Researchers have already documented that self-administered (including mail and computer) surveys elicit less SDB than a telephone or face to face interview (Acree, Ekstrand, Coates & Stall, 1999; Aquilino, 1992; De Leeuw, 1992; Fournier & Kovess, 1993; Groves, 1989; Knudsen, Pope, & Irish, 1967; Pless & Miller, 1979; Schwarz, 1991; Smith, Adler, Tschann 1999; Wiseman, 1972). To try to order survey methods in terms of SDB, this study investigated what method of surveying elicits less SDB—telephone or face to face interviewing. The literature was not conclusive. Kornendi (1988), Sykes and Collins (1988), Weeks (1983), Pless and Miller (1979), Rogers (1976), Thornberry (1976), and Colombotos (1965), indicated the telephone elicits less SDB than face to face interviewing. Others claimed face to face interviewing elicits less SDB than telephone interviewing (De Leeuw & Van der Zouwen, 1988 and Henson, Roth, & Cannell, 1977). Furthermore, De Leeuw’s Meta Analysis (1992), and Colombotos’ (1969) study also claimed there is no difference between telephone and face to face in SDB.

Closely related, if not identical, to SDB is the overreporting of “desirable” behaviors and underreporting of “undesirable” behaviors. Two behaviors prone to underreporting are the use/abuse of drugs/alcohol and having abortions. Johnson (1989) and Fendrich and Vaughn (1994) indicated less underreporting of substance use in a face to face method compared to telephone interviewing. Johnson and Fendrich and Vaughn have asserted that the face to face method leads to more underreporting than the telephone method, which supports De Leeuw and Van der Zouwen’s (1988) and Henson
et al.’s (1977) view that face to face interviewing elicits more SDB than telephone interviewing. Further support that the face to face method leads to more underreporting can be found in Smith, Adler, and Tschann’s (1999) study of underreporting abortion.

Contrary to the previously discussed studies that found face to face interviews elicit more SDB than telephone interviews, Luepker, Pallonen, Murray and Pirie (1989) indicated less underreporting of smoking in an Australian sample occurred in their telephone interviews than their face to face interviews. Their results were validated by physiological-chemical methods. Furthermore, Aquilino and LoSciuto (1989) indicate more underreporting in African Americans’ use of alcohol or marijuana in the face to face method than the telephone method, even after controlling for demographics.

To add to the conflicting findings above, two studies stated there is no difference between face to face and telephone interviewing in terms of SDB. Greenfield, Midanik, and Rogers (2000) and Aquilino and LoSciuto (1990) indicated underreporting of substance use in the face to face method is comparable to the telephone method except when dealing with minority or low income populations.

Thus, the question of whether telephone interviewing elicits more SDB than face to face interviewing is unsettled. To rank order the methods of collecting data in terms of SDB, this study weighed in on the above incongruity.

SDB occurs when a respondent distorts the truth to be perceived positively by the interviewer and/or the researcher (Martin & Nagao, 1989; Rosenfeld, Booth-Kewley, Edwards & Thomas, 1996; Vogt, 1999). This definition is in agreement mostly with Martin and Nagao (1989), and partially with Vogt (1999).
Rationales for the Study

Dillman (2000) stated, “It is particularly important to be aware of . . . potential sources of response differences. They include: “social desirability . . .” (p. 226). SDB has been a concern to survey researchers for over 75 years (Bernreuter, 1933; Meston et al., 1998; Paulhus, 1991; Smith, 1997). The Joint Committee on Standards for Educational Evaluation (1994) stated in their book, “The Program Evaluation Standards”

When collecting opinions, consider whether the respondents are motivated to tell the truth. Word questions to maximize understanding and minimize bias in responses . . . . Often it is desirable to employ nonreactive procedures . . . (p. 147).

Janda (1998) echoed the Committee’s concern about SDB indicating in “Psychological Testing: Theory and Applications” many researchers have found most people say socially desirable things about themselves when responding to tests of personality.

Computer administration of surveys—not on the Web—have been shown to be capable of reducing SDB (Evan & Miller, 1969; Kiesler & Sproull, 1986; King & Miles, 1995). Moreover, Mitchell (1993) found computer administration of surveys decreased Impression Management (IM), a specific factor of SDB. Paulhus (1991) declared,

The term ‘impression management’ was chosen to represent one traditional view of SDR (social desirable responding): that some subjects are purposefully tailoring their answers to create the most positive self-image. Of the impressions they try to present, this factor represents only one: a socially conventional, dependable persona (p. 21).

King and Miles (1995) also found that computerized administration of a survey decreased IM. Edwards (1996) found computers elicited less SDB than did the mail-in survey condition as measured by the Marlowe-Crowne Social Desirability Scale (MCSDS). The MCSDS scale was one of the first instruments to measure SDB in non-clinical
populations. Edwards reasoned an increase in candidness should translate into less SDB. Studies by Allen (1987), and O’Brien and Dugdale (1978) also support this “computer candidness” effect. However, some researchers have qualified this as a “computerized effect.” Edwards also reasoned there could be an anonymity effect. He found higher IM scores on the Balanced Inventory of Desirable Responding (BIDR) in identified rather than anonymous conditions. He also found that IM scores were significantly higher in the computer-linked, “big brother” condition than in his other conditions. Rosenfeld et al., (1996), Rosenfeld, Giacalone, and Riordan (1995), Dahl (1992), Paulhus (1984), and Dunnette and Heneman (1956), concur with this anonymity effect. For example, employees were more inclined to give an honest opinion of their supervisor when they were answering anonymously than when they were answering non-anonymously, when they were less likely to fear supervisory reprisals.

Decision Support Systems (1999) indicated Web-based surveys gave participants more of a sense of anonymity that leads to more honest and more critical answers. They also found participants seem to put more thought into their decisions as compared to mail surveys. Furthermore, the asynchronous nature of the Web, the fact that participants could have responded to a survey whenever they desired, further encouraged participants to put even more thought into their responses.

By comparing the four methods of data collection (face to face, telephone, mail, and Web-based) simultaneously the exploration of ordering the plethora of the literature can be taken. The susceptibility of this ordering, if it exists, of methods to SDB can help
researchers in planning, implementing, and reducing the costs of many future studies, especially in using mixed method designs as Dillman (2000, 2007) recommends.

**Research Questions:**

**Question 1:** Does the method of data collection (face to face, telephone, mail, Web-based) elicit Impression Management (IM) (the deceiving others factor of social desirability) in respondents?

For the purposes of this investigation, which deals with a non-clinical population, a concept such as Self-Deception Enhancement (SDE) is too introspective a variable to be studied. Furthermore, when researchers such as Marlowe and Crowne first studied SDB, their primary purpose was to study the IM aspect, factor, of SDB. Marlowe and Crowne developed their scale to measure a single factor, the need for social approval (Crowne & Marlowe, 1960). Yet, factor analyses revealed it was not a single factor (Loo & Thorpe, 2000; Paulhus, 1991; Smith, 1997). Paulhus (1991) found the Marlowe-Crowne Scale loads on both the IM and the SDE factors. The separation of the two factors resolves some of the discrepancies in the psychological personality “adjustment” literature. SDE is correlated strongly and positively with healthy, psychological adjustment but IM is not correlated with adjustment. Since Paulhus’ BIDR divides inventory, and the MCSDS does not, Paulhus’ inventory was used.

As Chapter III details Cronbach alphas in the SDB literature ranged from .65 to .86, which is an acceptable range for most social science measures. Test-retest reliabilities, which are generally lower than Cronbach alphas for any instrument (Crocker & Algina, 1986; Kerlinger & Lee, 1999; Kerlinger, 1986), ranged from .41 to .93 with
typical test-retest reliabilities ranging from .59 to .76. Since all the reliabilities addressed
a wide range of contexts, the BIDR can be said to be generally reliable. Paulhus’ (1991) BIDR achieved higher reliability coefficients than the MCSDS, which was the most widely used measure of SDB until the BIDR outperformed it.

There is evidence of the validity of the BIDR. Lajunen et al. (1997) found two factors in SDB in a traffic context. Also, Smith (1997) concluded that SDB is composed of two latent and highly correlated “.53” factors (p. 22).

The concurrent validity of the entire BIDR with the MCSDS was .71, and the concurrent validity with the Multi-Dimensional Social Desirability Inventory was .80 (Paulhus, 1991). Paulhus asserted convergent validity by stating, “The IM scale correlates highly with a cluster of measures traditionally known as lie scales (e.g., Eysenck’s Lie scale, MMPI Lie Scale) and role-playing measures (e.g., Wiggins’ Sd, Gough’s Gi).” Additionally, Davies, French, and Keogh (1998) reported, “The Lie scale was correlated . . . even more significantly with Impression Management (r = .61)” (p. 16). Paulhus (1991) established discriminant validity showing his pair of BIDR subscales, IM and SDE, were independent of each other. He stated the subscales were intercorrelated “from .05 to .40” (Paulhus, 1991, p. 39), which was lower than his prior versions when the factors were more highly intercorrelated. Moreover, in another study, Holden Starzyk, McLeod, and Edwards (2000) found that their congruence coefficient between the observed and theoretical target matrix was .91 for IM, which used Confirmatory Factor Analyses, supporting the construct validity of the BIDR-6. Furthermore, Stober, Dette, and Musch (2002) found concurrent validity between the BIDR-6 and the Social
Desirability Scale 17 (SDS-17) of .46. Therefore, the reasons for selecting the BIDR-6 were, first, the evidence for the BIDR was more reliable and valid than any non-clinical measure of SDB (Paulhus,; 1991, 1998, 2002; 2003; Paulhus & Reid, 1991; Kroner & Weekes, 1996; Meston & Heiman, 1998; Mitchell, 1993; Kwong-Liem, 1999; Vitelli, 1999; Davies, French, & Keogh, 1998). Second, of all the available instruments, the IM subscale more closely approaches what Crowne and Marlowe (1960) intended to measure when they developed their MCSDS. Third, besides the MCSDS, most of the established instruments measure SDB for clinical populations (Paulhus, 1991; Paulhus & Reid, 1991; Kroner & Weekes, 1996; Kwong-Liem, 1999; Vitelli, 1999). Fourth, the SDE subscale is too “introspective” to interpret and most social scientists desire dealing with phenomena, attitudes and/or behaviors that are more observable or that can be reasonably inferred from what is observable (Bartley, 1982; Bechterev, 1928; Boring, 1963; Bower & Hilgard, 1981; Hothersall, 2004; Leahy, 1997; Lowry, 1971; Philips, 1987; Popper, 1983; Sahakian, 1968; Schwitzgebel, 2004; Smith, 1983; Stagner, 1968; Thomson, 1968; Watson, 1948; Watson 1913). If people are deceiving themselves, concepts can become too idiosyncratic and unpredictable; because, if respondents are, in fact, deceiving themselves, they may have difficulties being able to accurately express their thoughts or attitudes to an interviewer or researcher through surveys. Therefore, researchers cannot as easily interpret a respondent’s level of SDE as with Impression Management (IM) without raising reasonable doubts (Davies, French, & Keogh, 1998; Lindeman & Marjaana, 1995; Paulhus, 2002, 1991, 1986, 1988).
Analysis: Ryan’s (REGWQ) Multiple Comparison Procedure: Since the BIDR measures IM on a ratio scale the Ryan test was used. Furthermore, “The bottom-line choice of MCP (Multiple Comparison Procedures), if you do all possible pairwise comparisons, have equal sample sizes . . . and want good power, would be Ryan” (Toothaker, 1993, p. 57). Moreover, “The MCPs presented thus far [including the Ryan] do not need a significant overall F test before they are computed.” (Toothaker, 1993, p. 41; Toothaker, 1991, p. 47-55). A priori MCPs are also generally more powerful than post-hoc MCPs (Toothaker, 1993, 1991; Glass & Hopkins, 1996, 1986). Furthermore, if the variances are unequal the Games-Howell (GH) MCP was used instead of the Ryan as recommended by Toothaker (1993, p. 62-63; 1991, p. 98-99, 105).

Question 2: Does the particular order of the IM subscale means for the four methods of data collection, as suggested by the literature cited on pages 3-7 and pages 12-15 (found in Question 1), remain consistent for both genders?

Glass and Hopkins (1996, p. 485) stated, “The absence of interaction is the statistical justification for generalizability . . . . If there is no interaction between a treatment factor and a relevant characteristic of the subjects (e.g., gender), the overall effect can be generalized to both sexes without qualification.” By overall effect, Glass and Hopkins, (1996) meant a single treatment factor or single demographic factor such as gender and not the synergistic effects of treatment and gender.

Question 3: Does the particular order of the IM subscale means for the four methods for data collection (found in Question 1) remain consistent across the two ethnic groups, Caucasian or non-Caucasian?
Based on the same rationale discussed in Question two “the absence of interaction is the statistical justification for generalizability” (Glass and Hopkins, 1996, p. 485) of the effect of data collection methods on IM subscale across the two ethnicities—Caucasians and non-Caucasians. A 4 x 2 unbalanced ANOVA was used.

Question 4: Does the particular order of the IM subscale means for the four methods of data collection depend upon both the participant’s ethnicity and gender? That is, does an interaction exist among methods of data collection, gender, and ethnicity in terms of SDB?

The same rationale for a two-way ANOVA, “the absence of an interaction is the statistical justification for generalizability,” (Glass and Hopkins, 1996, p.485) can be extended to a three-factor ANOVA. Therefore, a 4 x 2 x 2 unbalanced ANOVA was used.

Research Design

A randomized post-test only control group design was used. Students were randomly selected from students attending the main campus of Kent State University. Those undergraduates were then randomly assigned to a face to face interview, telephone interview, mail survey, or Web-based survey. There were several reasons this design was selected. First, as Kerlinger (1986) stated in regard to comparing two or more experimental groups, “As long as there is an attempt to make two groups systematically different on a dependent variable, a comparison is possible” (p. 307). Random selection, random assignment, and specific procedures having been strictly adhered to, ensured the samples were unbiased and systematically different. Second, a post-test only design circumvents reactivity issues (Rudestam & Newton, 2000). For example, it is impossible
to pre-sensitize participants to a treatment if a pre-treatment did not exist. Furthermore, “A pre-test can have a sensitizing effect on subjects” (Kerlinger, 1986, p. 310). Third, for the Web-based and mail survey groups, the researcher can assure anonymity (Isaac & Michael, 1995), which has been shown to affect SDB in computer surveys (Evan & Miller, 1969; King & Miles, 1995) and email surveys (Kiesler & Sproull, 1986). Fourth, this design controls for history, selection (if random assignment is also used), and pre-testing (Kerlinger, 1986; Kerlinger & Lee, 1999; Isaac & Michael, 1995). Fifth, the addition of a pre-test would have been more costly and inconvenient (Isaac & Michael, 1995). Since the four treatments in this study were defined as the four different methods of data collection, all four of the first five reasons for using a randomized post-test only control group design mentioned above become relevant for comparing the four different treatments (methods) on IM, which was the purpose of this study, because the above five reasons relate to either differential treatments or differing methods of data collection.

Research Hypotheses

Hypothesis 1: On the average the IM means will follow this order (from greatest to least): face to face, telephone, mail, and Web-based (refer to page 7 for research question 1). The higher means are more biased than the lower means. In the face to face method the interviewer is less of a stranger than a disembodied voice over the telephone. The respondent has eye-contact with the interviewer, thus giving the face to face respondent more of a reason to “look good” to the interviewer.

Evidence that face to face interviewing tends to generate more SDB than telephone interviewing: Kornendi (1988), Colombotos (1965), Rogers (1976) Weeks
(1983), Sykes and Collins (1988), Pless and Miller (1979), and Thornberry (1976) indicate telephone interviewing elicits less SDB than face to face interviewing. Thornberry showed the least amount of underreporting was in the mail method, followed by the telephone method, and the most underreporting in the face to face method. Only the face to face strategy was significantly different in underreporting, and Thornberry interpreted this difference as a SDB. Verrips et al. also found less underreporting in their mail surveys than the telephone or face to face interview. Nederhof and Anton (1985) indicated less SDB in mail surveys than in either face to face interviews or telephone interviews. From the above rationale and evidence this researcher will weigh in by stating the face to face method should elicit more SDB than the telephone method.

Evidence that face to face interviewing tends to generate more SDB than a mail survey: De Leeuw (1992) found less SDB in mail surveys than her face to face interviews. In their Meta Analysis, the researchers Richman, Kiesler, Weisband and Drasgow (1999) concluded computerized questionnaires elicit less SDB than face to face interviews. King and Miles (1995) also found computerized administration of a survey decreased IM. Potosky and Bobko (1997), Edwards (1996), Weisband and Kiesler’s (1996) Meta Analysis, Lapham (1991), and Kiesler and Sproull (1986), indicated computerized surveys decrease SDB relative to paper and pencil surveys.

Evidence and theory that a mail survey tends to generate more SDB than a survey administered through the Web: In using the Web, one must also use a computer. Therefore, all the benefits of computers in terms of SDB would have extended to Web-based surveys with the added protection of a secure server. Davis (1999) found Web-
based surveys promoted increased frankness and self-disclosure. Finally, Joinson (1999) indicated participants using the World Wide Web scored significantly lower on measures of SDB than participants responding with mail. Abilgaard (1999) agrees with this Web effect. The logic of this particular hypothesis is since computerized surveys elicit less SDB than paper and pencil surveys, Web-based surveys should have elicited comparable SDB as do computerized-only surveys. In addition, Web-based surveys can utilize a secure socket. This increased “sense” of privacy should have even further decreased SDB.

Hypothesis 2: There is no interaction between method of data collection and gender. While Meston and Heiman (1998), who were studying sexuality and SDB issues found only anecdotal evidence of an interaction between IM and gender without a method comparison, Miles and King (1998) and Dahl (1992) found no interaction between SDB and their method comparisons. Moreover, Edwards (1957) in a no-method comparison like Meston and Heiman’s study found no differences in SDB between males and females. Furthermore, Dahl’s dissertation was the only study to randomly assign participants to groups. She also used random selection, which the other studies did not use. In her study, both multivariate and univariate tests for the interaction between method and gender were not significant. So, the preponderance of the evidence points toward no interaction between method of data collection and gender.

Hypothesis 3: There is an interaction between method of data collection and ethnicity. A 4 x 2 unbalanced ANOVA was used with the four levels for method of data
collection being face to face, telephone, mail, and Web-based, and the two ethnic groups being Caucasian and non Caucasian.

Aquillino (1992, 1994), Rogers (1976), and Zhang and Gerstein (1999) all found an interaction between race/ethnicity and method of data collection, but Fowler (1997) and Greenfield et al. (2000) found no interaction. However, the Aquillino (1994), Rogers (1976), and Zhang and Gerstein (1999) studies used random assignment while the other studies did not. All three studies that used random assignment support an interaction effect. Moreover, except for the Rogers study, the context of the study was the use or abuse of drugs or alcohol. Furthermore, with the exception of the Aquillino (1994) study, the methods compared were telephone and face to face. Therefore, because of the similar substantive characteristics of the studies, an apples to oranges argument is less valid regarding the six studies above.

Hypothesis 4: There is no three-way interaction between method of data collection, gender, and ethnicity. As indicated in Askenasy’s (1978) abstract of his conference paper there was no evidence of a three factor interaction as measured by Edward’s Social Desirability Scale of SDB, gender and ethnicity. Every reasonable attempt was made to locate this conference paper including writing to the author, but the study could not be found. Askenasy did not mention a three factor interaction between gender, ethnicity and social class on SDB, but did mention some main effects of gender, ethnicity and social class on SDB. While this is not proof there was no three factor interaction between method of data collection, gender and ethnicity, there was no evidence of an interaction between gender and ethnicity on SDB. Additionally, there is
no reason to presuppose a three factor interaction between method of data collection, gender, and ethnicity.

Significance of the Study

To date, there has not even been one study of the simultaneous comparison between face to face, telephone, mail, and Web-based methods on any psychological measure that deals with response patterns of individuals, let alone social desirability. However, there have been five studies comparing three data collection methods simultaneously (face to face, telephone, and mail) (De Leeuw, 1992; Hochstim, 1967; Thornberry, 1976; Verrips et al., 2001; Wiseman, 1972). There have also been studies comparing mail versus Web-based surveys (Abilgaard, 1999; Hancock & Flowers, 2001; Joinson, 1999). There have also been studies comparing solely computerized versus mail surveys (King & Miles, 1995; Lautenschlager & Flaherty, 1990; Mitchell, 1993; Potosky, 1997; Rosenfeld, 1989; Whitener & Klein, 1995; Wilkerson, Nagao, & Martin, 2002). Moreover, although the cost order is known or informally known by researchers, the SDB order is not known. The order of method costs will be briefly discussed in Chapter II. Thus, to give survey researchers even more bang for their research bucks, an SDB ordering would be extremely useful to researchers, sponsors, stakeholders, legislators, and so on because it could enlighten these groups to recognize the trade-offs between SDB, costs and other aspects of different survey methods. All the interested parties potentially have to do is analytically juxtapose, or compare, the results of this study with the discussion of the relative costs of each survey method as described in Chapter II. Furthermore, none of the studies that compared only face to face, telephone, and mail
used MCPs. The studies that did compare the three methods of data collection used only comparisons of conventional, arithmetic means and not MCPs. This study used MCPs because they are appropriate to the nature of the problem. Thus, the probabilities of Type I errors could be controlled.
CHAPTER II
REVIEW OF THE LITERATURE
A History of Survey Research Methods

Throughout history, surveys have been taken for primarily economic, but sometimes military reasons. As previously mentioned, if a census can be called a survey, and Babbie (1973), Parten (1950), and Converse (1987) argued that a census is a survey of a population, the Babylonians were the first to conduct surveys. The clay tablets they left behind were the first written surveys. They are dated as early as 3800 B.C.E. (Halacy, 1980). Not until 2300 B.C.E. did each region take their own censuses, which were taken for taxation and revenue management reasons. However, there is some evidence the Chinese had taken a type of a census as early as 3000 B.C.E. Yet, this evidence is far from definitive. What is known is Confucius conducted a census of population, agriculture, industry, and commerce around 500 B.C.E. (Halacy, 1980).

The earliest Egyptian “population count” was taken about 2500 B.C.E. It was not until the reign of Ramses II, which began about 1250 B.C.E., that detailed and accurate censuses were conducted. The ancient Jews conducted a “shekel” census in 1491 B.C.E. King David conducted a census in 1017 B.C.E. against the Lord’s will (Halacy, 1980). Since that time, censuses, surveys, and statistics have been associated with evil until modern times. Nevertheless, David’s son Solomon, also conducted a census.

Servius Tullius (578-534 B.C.E.) is considered the founder of the Roman census. The ancient Roman censuses asked for name and age, class, family position, and personal possessions (including slaves, real estate, and personal property). The Romans kept a
tally of how many men, women, and children they had by requiring the men to deposit certain coins, the women to deposit a different coin, and the children to deposit yet a different type of coin. The Romans’ need for censuses and surveys were very practical. They needed to know how many soldiers they had, how much tax to collect, and how much food they had (Bradburn & Sudman, 1988; Halacy, 1980). Besides censuses, survey research can be traced as far back as the ancient Egyptians, making survey research at least 3,000 years old (Glock, 1967).

“The Dark Ages were lit by few statistics . . .” (Halacy, 1980). After the Dark Ages, Europeans began conducting surveys again. As trade grew in Europe there was a need for censuses. In gathering records for the “Domes Day Book” in what was the first recorded face to face survey interviews, William the Conqueror had his men canvassing England in 1085 A.D. obtaining information from every hundredth man—a primitive sampling technique (Bradburn & Sudman, 1988; Halacy, 1980). Not until 1377 did Edward III institute a census in England. He had a practical reason for instituting a census—he wanted to determine the scope of the damage caused by the Black Plague that occurred around the middle of that century. Around the same time the French were progressing with their census efforts, but they were not as successful as the English, in part because of the corruption in French census efforts. After 1377, wars and threats of wars preoccupied European nations, and no nation took a survey or a census for several centuries (Halacy, 1980).

As the Middle Ages were dominated by theology, the 1600s were dominated by the discovery of the “Natural World” (Clark, 1972). “In the century which began with
Bacon and ended with Newton, the most tangible product of this connection was the rise of statistics, the quantitative study of social facts,” (Clark, 1972, 16). Modern statistical studies can trace their origin back to the founding of the Royal Society (now the Royal Statistical Society) and John Graunt’s publication, “Observations of the Bills of Mortality.” Both events occurred in 1662. From this it can be seen surveys and statistics did not arise from nothing, rather, this new “political arithmetic” as they called it, arose out of the background of what the times were capable of accepting (Clark, 1972). This new political arithmetic also rose in America as early as 1824 in the form of straw-vote journalism for a presidential election (Converse, 1987; Gallup, 1940; Moon, 1999; Northrop, 1971; Parten, 1950).

Building on the advances by Laplace and Gauss in probability, Adolphe Quetelet introduced the idea of the “average man” to social science (Aiken, 1997; Lazarsfeld, 1961). Instead of basing his theories on unsystematic observation and insight (unlike Marx and Comte), he based his theories on systematic observations and his experience in collecting similar data. In 1834, Quetelet assisted in establishing the Statistical Society of London, later known as the Royal Statistical Society. This Society and the Manchester Statistical Society, established a few years earlier, represented the pinnacle of survey research until the works of Charles Booth established a new summit (Glazer, 1959). They were concerned about the effects of industrialization and urbanization on poverty. It has been argued the Manchester Statistical Society sent the first mail survey, although some sources claim Karl Marx sent the first mail survey. What can be said is the Manchester Statistical Society sent the first mail survey in 1834, but Karl Marx became the first
individual to send a survey in 1880. (Babbie, 1973; De Leeuw & Cole, 1972; Collins, 1997; Elesh, 1972). The French and Germans contributed to survey research as well, but the British were the nation who embraced surveys and statistics. They publicized both of them! It was no historical accident. As they led the Industrial Revolution, they also led the way in the social scientific approach to problems (Glazer, 1959).

Charles Booth, a wealthy steamship owner, was probably influenced by the Manchester Statistical Society’s works. He is considered the father of modern survey research because he surveyed London, the largest city in the world—a monumental task (Abrams, 1951; Bradburn & Sudman, 1988; Converse, 1987; Parten, 1950). In the 1880s and 1890s he used school inspectors as informants in a systematic sampling of London (Cole, 1972; Converse, 1987; Glazer, 1959). Just as individuals in the Manchester Statistical Society wanted to change society for the benefit of the poor, he also wanted to help the poor. Because he was a businessman, he held a deep respect for numbers, but he also had a novelist’s eye for character and details. When he conducted his poverty surveys, he wanted them to be scientific—unbiased and objective. There was great controversy in London in the 1880s over what the actual poverty rate was (Abrams, 1951; Converse, 1987). Booth concluded the poverty rate was “a whopping 30.7 percent of the population—in the richest city in the world” (Converse, 1987, 14). Furthermore, he explored the causes of poverty. Because of his detailed and systematic observations and procedures, he was more correct about what caused the miseries of poverty than his contemporary, Karl Marx (Glazer, 1959).
The Reformers, people like Charles Booth, dominated survey research until World War I. After World War I, the United States government showed some interest in survey research and after both the Great Depression and World War II the government showed a great responsiveness to developments in survey research (Converse, 1987).

As the Reformers and the government were increasingly using surveys, survey research was “evolving” from complete enumeration at one end of the sampling continuum to full probability sampling at the other end. By the time of Franklin D. Roosevelt’s New Deal plan, survey researchers were either engaged in “intuitive” sampling—not a complete enumeration but not quite probability sampling—and “structured” sampling, which incorporated intuitive sampling and some aspects of probability sampling (Converse, 1987). In England in 1912, Arthur Bowley, an economist and a statistician, and his colleagues “were the first to apply the principles of scientific sampling in selecting the population to be surveyed” (Parten, 1950, 7) while in the United States, the first scientific poll was conducted by Fortune magazine in 1935 (Northrop, 1971).

Since much of Survey Research is about attitudes, a brief summary of the history of studying attitudes seems appropriate. In Social Psychology the study of attitudes thrived starting in the 1920s. Psychologists studied attitudes in their experimental and controlled laboratories. Moreover, the study of attitudes was a natural extension of mental testing. Psychologists like Sigmund Freud, Carl Jung, and Alfred Adler resurrected research about attitudes. The sociologists studied attitudes using case histories, fieldwork, and analyzing aggregate statistical data. This schism between psychology and sociology
was not clear cut, as researchers in both disciplines “borrowed” principles and techniques from the other discipline to study attitudes. There was also a schism within sociology. Emory Bogardus espoused a more holistic view of sociology and, therefore, his studies of attitudes were more qualitative in nature. Floyd Allport was more focused on the individual. Influenced by psychologists, he created a scale to measure attitudes by using answers from 60 of his students and six judges to structure his scale more “objectively” than Bogardus’ Social Distance scale. Influenced by Allport’s scaling, the psychologist L.L. Thurstone created a scale based on judges’ favorability. The judges placed each attitude onto 11 piles that represented equal distances of favorability. The problem with this scale was it was too costly to implement because of its intensive labor costs. Rensis Likert’s scale bypassed paying judges through administering the questions directly to subjects and afterward ascertaining if they were internally consistent. Likert’s scale correlated highly with Thurstone’s scale and had the advantage of obtaining higher reliabilities with fewer questions. Because of the above reasons and because it was much cheaper, the Likert scale was widely adopted. In 1934, Likert also “retested” his students. This was the first panel measurement of attitude change that had a number of years between measurements. Furthermore, during World War II, Luis Guttman developed his own method of scaling, and after the war, many researchers and social scientists used his scale (Converse, 1987).

While the psychologists were improving scaling, two political scientists under the direction of Charles E. Merriam did pioneering survey work at the University of Chicago during the 1920s. Leonard D. White studied the prestige of government versus private
industry jobs. Initially, he used a mixture of methods like Bogardus. He had to rely on written surveys rather than personal interviews because he had to survey 4,700 people with only four graduate students and himself. He found private sector jobs were perceived as more prestigious than government jobs, individuals in lower socioeconomic groups perceived government jobs more favorably than higher socioeconomic groups, and when individuals were asked about their own dealings with the government only 24 percent expressed clear dissatisfaction. Furthermore, Harold F. Gosnell also used personal interviewing and written questionnaires. He conducted the first field experiment outside of psychology when he tested the differential effect of mailed reminders on voting registration behavior. He found the overall increase in voting registration influenced by the mail reminders was 10 percent and reminders were not as successful among more educated voters. Gosnell was not the only one interested in voting. Survey researchers owe much of their field’s progress to the pollsters of the 1930s, 1920s, and before (Converse, 1987).

The pollsters had what the academicians did not—funding for survey research and access to national data. Archibald Crossley, George Gallup, and Elmo Roper constructed public opinion polls by merging the two fields of market research and straw-vote journalism. The market research funded their polls while their straw-votes publicized them. In the 1920s and 1930s, interviewers for market researchers took the “middle ground” in interviewing. On one pole personal interviewing was more of a conversation with its ebb and flows, but on the other pole they used only a few standardized questions—the same questions given to each participant. Applied psychologists who
brought with them the principles of laboratory experimentation, stressed standardized
interviews. They knew from their testing experiences that slight variations in the way a
question was worded could result in different answers. Moreover, in the early 1920s A.T.
Poffenberger found that if you asked housewives meaningless or ridiculous questions,
they would still answer these questions. In 1926, John Jenkins found a truncated list of
alternative answers greatly differed from a list of non-truncated answers, even though the
participants in the truncated answer group were greatly encouraged to supply their own
responses. Jenkins then recommended survey researchers produce complete checklists
from their pre-tests to circumvent this truncated list bias (Converse, 1987). These two
experiments illustrate the Garbage-In-Garbage-Out (GIGO) principal. In the 1940s,
Henry C. Link saw the advantage of having many interviewers before Gallup—it reduced
interviewer bias. Link and his colleagues advocated reading questions precisely as written
(Converse, 1987).

Poffenberger was also opposed to surveys that encouraged respondents to guess
what their answers were. Many market researchers urged reducing the difficulty of
survey questions. Experienced market researchers of this era also realized the weaknesses
of the mail survey, which typically had response rates of 10 to 20 percent, and the
returned surveys were often atypical of the non-respondents (Converse, 1987).

Gallup had predicted the “Literary Digest” debacle—the “Literary Digest”
forecasted an Alf Landon victory over Franklin Roosevelt in the election of 1936—even
before the Digest had collected any data. Gallup and the Digest collected data from the
same sources—lists of telephone subscribers, automobile owners, and registered voters.
The problem the Digest encountered was that these lists were overly represented by the upper classes and the upper classes tended to vote for republicans rather than democrats (Converse, 1987).

Archibald Crossley conducted the first recorded telephone interview surveys in the 1930s (Converse, 1987). The first Computer Assisted Telephone Interviewing (CATI) systems were developed by Chilton Research Services and the University of California, Los Angeles (for academic usage) in 1972 (Frey, 1983; Groves et al., 1988; Shanks, 1983). In 1994, the first Web-based survey was completed by the Graphics, Visualization, and Usability (GVU) Center at the Georgia Institute of Technology to study the demographics of Web surfers (Hewson, Yule, Laurent, & Vogel, 2003). Furthermore, in 1997, the first Web-based experiment published in a scientific journal was conducted by Krantz, Ballard, and Scher who studied the determinants of female attractiveness (Musch & Reips, 2000; Riva, Teruzzi, & Luigi, 2003). The use of computer technology has revolutionized the survey industry (Tourangeau, Rips, & Rasinski, 2000).

The Advantages and Disadvantages of Different Survey Methods

When comparing costs associated with the various survey methods, face to face interviews appear the most costly and the Web-based surveys the least costly. Face to face interviews are expensive because the researcher has to pay for wages, travel expenses, and the “unproductive” time between households for a face to face interview. When considering telephone interviews the researcher has to pay some labor costs, but not as much per interview as a face to face. For telephone interviews long distance toll
charges do exist, but these are quite small compared to the labor costs of a face to face interview. With mail surveys the primary variable costs are from handling the survey (folding and envelope stuffing), the postage, the data coding, and the data entry.

While an individual researcher’s costs may vary, the cost order can be useful as a rough guide to beginning researchers. In terms of variable costs, the cost order from most expensive to least expensive is: face to face, telephone, mail, and Web-based. For example, Dillman (1978) in comparing the costs of face to face, telephone, and mail surveys stated, “it is clear that in most cases in-person [face to face] interviews are the most expensive and mail questionnaires the least” (p. 72). Survey researchers should realize by deduction the telephone interview, in general, costs more than a mail survey but less than a face to face interview. This position has been supported by various research studies that have investigated the cost-effectiveness of different methods of survey research (Hochstim, 1967; McHorney, Kosinski & Ware, 1994; O’Toole, Battistutta & Couch, 1986; Siematycki, 1979; Sudman, 1967; Thornberry, 1976; Weeks, Kulka, Lessler, & Whitmore, 1983). Finally, Dillman (2000) indicated after the programming of the Web site has been accounted for, the Web-based method is cheaper than mail or telephone. When one compares the variable costs with Web-based surveys to other survey methods, they are far less expensive. For example, Schleyer and Forrest (2000) indicated a Web-based survey is 38% cheaper than a mail survey, and the Web-based cost effectiveness tends to increase as sample size increases. Dillman (2000; 2007) also suggested Web-based surveying is cheaper than mail or telephone. Dillman (2000), Ramit (1999), Schleyer and Forrest (2000), Sheehan and Hoy (1999), and Weible and
Wallace (1999) also suggested the Web-based method is quicker and can reduce item non-responses if programmed appropriately.

At the local level, Kent State’s Department of Sociology charged an average $23.99 per telephone interview for variable costs with a range of $23.75 to $24.17, depending upon sample size. Their Web-based survey costs were about $7 per participant (B.L. McDonald, personal communication, November, 11, 2004). Kent State’s Bureau of Research Training and Services of the College of Education charged $2 per survey for a mail-in/mail-back survey (D.A. Shelstak, personal communication, October 5, 2004). At the University of Akron’s Center for Policy Studies, the variable costs for telephone and mail surveys were approximately $5 and $3, respectively (S.M. Henry, personal communication, November 9, 2004). So, the cost order appears to hold, but there is variability between organizations.

The next attribute that shall be evaluated is non-response bias. Of the following methods of data collection: face to face, telephone, and mail, mail is most susceptible to non-response bias, with telephone showing an intermediate level of non-response bias, and face to face showing the least susceptibility to non-response bias (Bourque & Fielder, 1995; de Leeuw, 1992; Golden, 1982; Hacket, 1981; Hox & de Leeuw, 1994; Kennedy & Vargus, 2001; Shermis & Lombard, 1999; Zelen, 1975). It has also been shown that mail is more susceptible to non-response bias than telephone, telephone is less susceptible to non-response bias than face to face, and mail is more susceptible to non-response bias than face to face alone (Adams, 1989; Bailey, 1994; Bradburn & Sudman, 1988; Dillman Gallegos, & Frey, 2001; Fowler, 1993; Goyder, 1987; Hagan, 2003; Zelen, 1975). Fowler
(1993) however, qualified the findings that face to face interviews are less susceptible to non-response bias than telephone. He found non-response bias is less of a problem in urban areas over the telephone than a face to face, and that a face to face is even more advantageous in non-response terms than a telephone call when considering suburban and rural samples. Moreover, it has been shown mail is more susceptible to non-response bias than Web-based surveys (Cobanolu, Ward, & Moreo, 2001; Higgins, Dimnik, & Greenwood, 1987; McCabe, Boyd, Cooper, & Crawford, 2003). However, for a college sample, Sax, Gilmartin, and Bryant (2003) stated that Web-based surveys proved more susceptible to non-response bias than “mail” administrations. Sax, Gilmartin, and Bryant (2003) also found, contrary to conventional wisdom, their male students engaged in more non-response behavior using the Web-based survey than for mail surveys. The female students, however, showed more non-response in the mail than in the Web-based method.

Therefore, in the mail, telephone, and face to face comparison it could be “safely” concluded from the previous paragraph that mail was the most susceptible to non-response bias, with telephone showing an intermediate non-response bias, and face to face showing the least amount of non-response bias. Because of the contradictory and scarce number of studies involving Web-based surveys, no firm conclusions could have been made.

The attribute of speed will be considered. Face to face interviewing is notorious for being a slow method of collecting data when compared to others (Hagan, 2003). Of the three “traditional” methods of collecting data—face to face, telephone, and mail—the telephone is faster than mail, which in turn is faster than face to face (Bourque & Fielder,
1995; Golden, 1982; Salant & Dillman, 1994). Bailey (1994), Mandall, Eaden, Mayberry, and Mayberry (2003), and Vasu, Moriarty, and Pelfry (1995) supported the previous statement by indicating telephone is faster than mail or face to face. Moreover, the telephone has been shown to be faster than mail alone (Dillman, Gallegos, & Frey, 2001). Furthermore, the telephone has been shown to be faster than a face to face alone (Day & Campbell, 2003; Holbrook, Green, & Krosnick, 2003).

Web-based surveys have been shown to be faster than face to face, telephone, and mail, because the data is already in a machine-readable format rather than having to be keyed in by an interviewer (Braithwaite, Emery, de Lusignan, & Sutton, 2003; Dillman & Bowker, 2003; Hagan, 2003; Pealer & Weller, 2000; Saxon, Garrat, Gilroy, & Cairns, 2003). Web-based surveys have also been shown to be faster than mail (Cobanolu, Warde, & Moreo, 2001; McCabe et al., 2002; Stewart, 2003). Therefore, the speed ordering would be from fastest to slowest: Web-based, telephone, mail, and face to face.

The next attribute of a survey to consider is data quality. For the purposes of this discussion data quality refers to the amount of item non-response—which is different from survey non-response (not responding to any questions in a survey)—and the meticulousness of respondents’ answers to open questions, but does not refer to respondent error which will be discussed later. It has been shown item non-responses are higher in Self-Administered Questionnaires (SAQs) than Interviewer-Administered Questionnaires (IAQs) except when the SAQ has been computerized, and that data collection errors are higher in SAQs than IAQs (Gribble et al., 1999; Salant & Dillman, 1994; Touramgeau, Rips, & Rasinski, 2000; Wright, Aquillino, & Supple, 1998). Their
rationale was a face to face interview can motivate the respondent to give a more complete interview except in the case of a computerized survey which can branch to the unanswered questions before survey completion. The computer essentially obliges the respondent to answer all unanswered questions. Moreover, IAQs allow interviewers to probe respondents for answers whereas the SAQs tend not to allow as much probing (Frey & Oishi; 1995; Gribble, Miller, Rogers, & Turner, 1999). Corroborating the findings above, Jackson and Rothney (1964) reported more complete interviews in a face to face interview than a mail survey; Herzog and Kulka (1983) and Groves and Kahn (1979) found face to face interviews are more complete than telephone interviews, and Herzog and Kulka (1983) also found their telephone interviews were more complete than their mail surveys; and Bailey (1994) recounted face to face interviews were more complete than mail surveys. However, it should be noted IAQs are better than SAQs with populations having illiterate participants (Bajos, et al., 1992; Bourque & Fiedler, 1995; Fowler, 1993; Fowler, 1984; Francis, Frey, & Hartly, 1979; Gribble et al., 1999; Harrison & Hughes, 1987; Salant & Dillman, 1994). Furthermore, IAQs are better at eliciting responses to open-ended questions than SAQs, further improving data quality (Dillman, 2000; Fowler, 1993; Golden, 1982).

Respondents are more prone in SAQ methods to answer questions out of the order intended by the researcher than they are in IAQ methods (Bourque & Fiedler, 1995; Frey & Oishi, 1995). Moreover, Groves et al. (1988) indicated a response “order” effect was more prevalent through the telephone method than a SAQ method, which is congruent with Bourque and Fiedler’s (1995), and Frey and Oishi’s (1995) position. When using the
Web-based method, the surveyor has a chance to ensure the respondent answers the questions in the intended order through “branching” the questions (Braithwaite, Emery, de Lusignan, & Sutton, 2003). In the mail method, the researcher relinquishes control over the “order” in which the respondent answers questions over to the respondent when compared to a face to face interview (Bailey, 1994). When comparing the two interviewer-administered methods, face to face seems to increase data quality over telephone interviews. For example, Day and Campbell (2003) indicated that face to face interviews are more complete than telephone interviews and, as previously mentioned, Herzog and Kulka (1983) also found face to face interviews were more complete than telephone interviews. However, it must be noted Farnworth, Bennet, and West (1996) found their mail surveys were more complete than their telephone surveys. Nevertheless, Hagan (2003) indicated researchers can get more in-depth answers to open-ended questions when using face to face interviews than telephone.

There are two kinds of control—control over who responds to the survey, and control over the environment and pace of the survey in which the respondent answers the question. Regarding control over who responds to the survey, the researcher in a face to face has the most control over who responds to the survey, followed by the telephone, and has the least control over mail respondents (Bailey, 1994; Bourque & Fiedler, 1995; Day & Campbell, 2003; de Leeuw, 1992; Fowler, 1993; Francis, Fray, & Hartly, 1979; Salant & Dillman, 1994). Regarding control over the environment and pace of the survey to which the respondent takes the survey, the researcher has more control in the telephone than the face to face interview (Hagan, 2003; Mitchell & Rogers, 1958). For example, the
respondent in the face to face interview may insist the interviewer have a cup of coffee or a piece of cake before proceeding with the interview whereas in the telephone interview, the interviewer can more quickly focus on the task at hand—getting a completed interview. Moreover, Dillman, Sangster, Tarnai, and Rockwood (1996) sensibly found the researcher has more control in a telephone interview versus a mail survey. Likewise, in a face to face interview the researcher has more control than in a mail survey (Bailey, 1994). Furthermore, Riva, Terruzzi, and Luigi (2003) commented the researcher has little control over where and under what conditions a Web-based survey is completed, but that no definitive conclusions with regard to the Web-based method could be made because there were no comparisons with other methods with the respondent control variable.

Mail and Web-based are both SAQ methods. Both are more vulnerable to a “volunteer” bias than IAQs (Bailey, 1994; Frey & Oishi, 1995; Riva, Terruzzi, & Luigi, 2003).

Different survey methods differ in the manner in which they can handle complex questions. According to Hagan (2003) and Bailey (1994), a face to face interview is more amenable to changes in the structure of a survey question than are telephone interviews or mail surveys. Day and Campbell (2003) have shown the face to face is more adaptable than telephone interviewing. Fowler (1993), however, argued a telephone interview can be as flexible as a face to face interview if the unfolding technique—the complex question is broken into two or more simpler questions—is used. However, Fowler did not argue with the relative rigidity of a mail survey. On the other hand, Kennedy and Vargus (2001) contended telephone interviews must contain shorter questions because of
respondents’ memory limitations. Notice that Fowler’s and Kennedy and Vargus’s positions are reconcilable, but Hagan and Bailey’s position that the face to face “can” be more adaptable than a telephone interview or a mail survey is indisputable. Furthermore, logically, because of the Web-based surveys’ capacity for sophisticated branching of questions, Web-based surveys can be more amenable to complex question structures than a mail survey (Fowler, 1993; Higgins, Dimnik, & Greenwood, 1987).

Calder (1998) was only partially correct in enumerating the factors that constrain the researcher in a “practical” sense. She specified the four factors which will force the researcher to choose one method of data collection over another are time, funding, trained or experienced staff, and research sponsorship. For the time factor, if a survey has to be completed in one or two days, the researcher may have little choice except to survey by telephone (Borque & Fiedler, 1995; Calder, 1998; Dillman, 1978). However, Hagan (2003) indicated Web-based surveying is faster than telephone surveying. Moreover, non-response bias now may become a problem because the people away from their homes for one or two days will be excluded from the survey (Salant & Dillman, 1994). For the funding factor, the face to face interview should obviously be used when a researcher has an adequate (large) budget (Calder, 1998; Salant & Dillman, 1994). Moreover, logically, for a fixed dollar amount per survey, the cost order mentioned previously should be considered. For the trained or experienced staff factor, a face to face or telephone interview cannot be executed well unless the organization has a trained and sometimes experienced supervisors or interviewers (Calder, 1998; Salant & Dillman, 1994). For the
research sponsorship factor, the sponsor of the research can influence all aspects of the research process including data collection (Calder, 1998).

Calder (1998) did not mention some factors that have an influence on the method of data collection. For example, Salant and Dillman (1994) indicated face to face interviews are to be preferred for populations that have no list. They also suggested a face to face interview can be used when respondents will not respond willingly or accurately by telephone or mail. However, this “get in the face” approach to a respondent comes with a high expense of time and money (Salant & Dillman, 1994). Moreover, Calder (1998) omitted geographic dispersion from her list of aspects to consider when choosing a data collection method. If the population to be surveyed is greatly dispersed geographically, a face to face method becomes enormously expensive (Bourque & Fiedler, 1995; Francis, Frey & Hartly, 1979). Consequently, the mail, telephone, and Web-based methods become more attractive. The researcher must also consider as an aspect the literacy and language of the population (Bourque & Fielder, 1995; Pless & Miller, 1979). If a substantial portion of the population cannot read, or cannot read the language in which the survey is written, then sending a mail or Web-based survey may be out of the question. Furthermore, if a researcher is studying a highly specialized population, such as illicit, recreational drug users, the Web-based method may be the only workable option to gain access to this special population (Duncan, White, & Nicholson, 2003). This paragraph was not written to criticize Calder’s (1998) work because she did, in fact, succinctly summarize the “main” factors a researcher should consider when deciding what data collection method to use. Rather, it attempted to
sensitize the future researcher to all the aspects of using different methods to consider in
deciding which method to use. As such, the last two paragraphs were not an exhaustive
list of aspects to consider when choosing a method of data collection, but it attempted to
illustrate the main aspects and to stimulate thinking of some prospective ones.

Mail surveys require the least amount of interviewer training, supervision, and
monitoring (Fowler, 1993; Mitchell & Rogers, 1958; Salant & Dillman, 1994). According to Salant and Dillman (1994) a survey researcher can spend months writing a
mail survey, but once the mailing of the surveys has begun there is little the survey
researcher can do except process the returned surveys, prepare additional mailings, or
cross out participants’ names on a list. In contrast, a face to face or telephone interview
would require the survey researcher to answer questions in a fast-paced and high
pressured manner (Salant & Dillman, 1994). The researcher would have to answer the
interviewers’ questions or describe how to solve the interviewers’ problems as they came
up. However, if a researcher can use an already professionally trained interviewing staff,
then interviewer control concerns lessen (Fowler, 1993; Mitchell & Rogers, 1958).
Furthermore, with a mail-in, mail-back method, the interviewer control problem
completely disappears because there is no interviewer interface (Salant & Dillman,
1994).

To the extent that an address list is inaccurate, inaccessible, not up-to-date, or
incomplete, the results of the survey are not able to be generalized (Bourque & Fielder,
1995; Fowler, 1994). The coverage error of a mail survey is generally lower than the
telephone or Web-based because while the vast majority of people have a postal address,
they are less likely to have a telephone number or email/Web address (Cobanolu, Warde, & Moreo, 2001; Dillman, 2000; Salant & Dillman, 1994). However, what seems to be an accurate, up-to-date listing of people will have a considerable proportion of incorrect or vacant addresses (Bradburn & Sudman, 1988). Moreover, differentiating a non-response from an invalid address can be problematic (Bailey, 1994). Furthermore, it is impossible to administer a mail survey without names for a rural sample (Fowler, 1993).

Survey methods differ in their tendencies to yield respondent (or response) errors. Respondent errors occur because of (a) respondents’ faulty memories, (b) respondents’ ignorance on certain topics, (c) misunderstandings by the respondent, (d) coincidental aspects that affect the respondent and interviewer in an interview (i.e., a baby crying, a dinner burning, or a tornado warning), and/or (e) the respondent wanting to be perceived by the interviewer or researcher in a positive light—the definition of SDB (Belson, 1986; Foddy, 2001; Hansen, Hurwitz, & Madow, 1953; Moser & Kalton, 1971; Semon, 2000; Wentland & Smith, 1993). Respondent error is defined as occurring when the answer by the respondent differs from what his “true” answer would be (Hansen, Hurwitz, Marks, & Mauldin, 1951; Marquis et al., 1981; O’Muircheartaigh, 1982). According to Forsman and Schreiner (1991) another type of respondent error is planned fabrication by interviewers of interview results. A “planned fabrication” is not regarded as a respondent error because the respondent did not err—the interviewer did. Well written and executed surveys or interviews and, if applicable, well trained interviewers and quality assurance, can greatly reduce the respondent errors to reasons one and two and to a lesser extent three through appropriate probes, but there is no known method to effectively reduce four
except giving the respondent a self-administered survey. A researcher can check for respondent error due to reason four by re-interviewing a subsample or the whole sample.

A face to face is less susceptible to respondent errors (assuming the interviewers are well trained) than Web-based, mail, or telephone surveys because the interviewer can clarify the meaning of questions and can notice the verbal and non-verbal cues of respondents’ confusion or frustration (Hagan, 2003). It seems intuitive the telephone would be better at avoiding respondent errors than Web-based or mail because the interviewer could listen to the respondent and clarify anything unclear to the respondent. In fact, it can only be said the telephone is better than only mail at avoiding respondent errors. Web-based surveys can be programmed to eliminate logical inconsistencies, thereby reducing respondent errors (Hagan, 2003; Higgins, Dimnik, & Greenwood, 1987; Pasveer & Ellard, 1998; Pettit, 2002; Pettit, 1999; Saxon, Garrat, Gilroy, & Cairns, 2003; Weible & Wallace, 1998; Wright, Aquilino, & Supple, 1998).

Threats to the Validity of Surveys

From the above definition and descriptions of respondent errors it is easily recognizable how they could be considered a threat to the validity of surveys. Specifically, the threats to the validity of surveys due to respondent errors are as follows:

1. Often, a respondent does not understand a question’s meaning or its implications.
2. Sometimes the respondents may not provide an accurate reply because they gave little attention to the question, they are uncomfortable about sharing embarrassing information, or they think it is no one’s business except their own.
3. Often, respondents reply carelessly after having grasped the meaning of the question accurately.

4. Sometimes respondents’ memories are faulty.

5. Sometimes respondents will change the “meaning” of a question to answer what they thought a question meant, or so they can answer a question in such a way that it allows them to answer the question in their idiosyncratic way.

6. Respondents may be reacting to a question when its context is not fully circumscribed.

7. Sometimes the number of alternatives respondents are offered is incomplete,

8. Sometimes the number of alternatives may be too long for respondents’ memories.

9. Sometimes some questions get “scripted” but not exactly true answers.

10. Respondents may misinterpret the question because it has unfamiliar or technical terms.

11. Sometimes respondents may simply misunderstand the meaning of a question (Belson, 1986; Campbell, 1946; Cantril & Hadley, 1944; Foddy, 2001; Martin, 1984). Moreover, Cantril and Hadley (1944) commented that questions may be meaningless to a portion of the participants because the survey was designed for minority groups or specific, targeted subsamples of a population. Furthermore, identical answers could mean different connotations or denotations for different respondents (Belson, 1986; Campbell, 1946).

Closely related to respondent error is the respondents’ appropriate comprehension of a survey question. Belson (1986) defined miscomprehension as “the respondent fails to
interpret the question in the way intended by its designer” (p. 4). Belson (1986) indicated respondents may, consciously or unconsciously, misconstrue a question so that they can answer it more easily or in a way that changes the meaning of the question. He discovered this by administering a survey first, then following it up with an intensive cognitive interview. Belson’s results are congruent with Nuckols’s (1953) results, in which he gave his participants an opportunity to have a question clarified and yet no one asked for clarification, even though the results indicated that some participants misunderstood a question. Moreover, Foddy (2001), and Bradburn and Sudman (1991) reported participants will answer questions on a subject they know especially little.

The way a questionnaire or interview is worded can affect the validity of the survey. For example, the intonation of a question can affect its validity, especially in IAQs and most specifically face to face interviews (Campbell, 1946; Cantril & Hadley, 1944; Groves, 1989; Hunt, Sparkman, & Wilcox, 1982; Sudman & Bradburn, 1982). For example, voice inflections, gestures, body language, and facial expressions can influence how a question is asked and subsequently answered (Groves, 1989; Hunt, Sparkman, & Wilcox, 1982). Furthermore, if the vocabulary requirements of the survey are too high, some respondents may not appropriately comprehend the question (Belson, 1981; Cantril & Hadley, 1944; Terris, 1949).

The ordering of the questions in a survey and the ordering of the options on a rating scale can affect the validity of a survey. For example, the responses to earlier questions can influence answers to questions asked later (Belson, 1986; Foddy, 2001; Martin, 1984; Sudman & Bradburn, 1982). Moreover, the ordering of the options on the
rating scale can affect how questions are answered (Belson, 1986, 1966; Dillman, 2000; Foddy, 2001; Schwarz & Hippler, 1991).

Both Foddy (2001) and Belson (1986) acknowledged respondents’ faulty memories are a threat to the validity of a survey. Belson (1986), and Forsyth and Lessler (1991) emphasized researchers’ are obligated to prod respondents’ memories by using appropriate descriptions or visual aids because as Belson (1986) stressed, “human beings are not recording machines to be turned on by an interviewer’s doorstep questioning” (p. 19).

Interviewer error can threaten the validity of surveys. Interviewer error can range from planned fabrication of interviews to the subtle influences interviewers’ socioeconomic and demographic characteristics like race, sex, and clothing worn by interviewers may have on a respondent (Forsman & Schreiner, 1991; Groves, 1989). More frequently, deviations from the “script” of the interview can threaten its validity. For example, DeMaio (1984) found the interviewers’ roles as persuaders and impartial communicators may conflict. To illustrate this, interviewers must “persuade” the respondents to participate in their interviews but must assume the role of “neutral and objective” observers in eliciting and recording participants’ responses. Moreover, Belson (1986) warned surveyors to be aware interviewers are under constant pressure to depart from their scripts due to situational factors and expediency. Foddy (2001) specifically warned of interviewers changing the wording of questions, not giving the same tones to their questions from participant to participant, and the adoption of flawed recording techniques. Groves (1989) corroborated this, but also indicated the same interviewer
when faced with different situations will use different probes—thus differentiating the question from participant to participant. Thus, interviewers can have the effect of making the interview not “standardized.”

Since the late 1940s surveyors have known of another threat to the validity of surveys—acquiescence bias (Cronbach, 1946, 1950). Acquiescence bias has been defined as “a tendency for some persons to agree with all statements of another person, apparently disregarding the content of those statements” (Couch & Kenniston, 1960; Groves, 1989). This is the yea-saying bias. Couch and Kenniston (1960) also found the opposite of the yea-saying bias, which is the nay-saying bias. That is, the respondent will tend to disagree with all statements regardless of content (Couch & Kenniston, 1960). In supporting “balanced” scales like the Balanced Inventory of Desirable Responding (BIDR), Couch and Kenniston (1960) have in essence proposed a statistical solution to both yea-saying and nea-saying by having these effects cancel each other out. The Marlowe-Crowne Social Desirability Scale (MCSDS) is not balanced, but the BIDR scale is balanced. Hence, this is another reason for favoring the BIDR over the MCSDS, which has been the traditional method of gauging SDB.

Extreme Response Style (ERS) is another threat to the validity of surveys. Bachman and O’Malley (1984) defined ERS as, “The tendency for some individuals to use the extremes, the end points, of response scales (e.g., strongly agree, strongly disagree) while others are more likely to use the middle values (e.g., mostly agree, mostly disagree) (Hamilton, 1968).” Woodworth (1938) indicated it is psychologically easier for respondents to answer a question using the extremes of a scale than it is for them to use
points in between the extremes. Interestingly, Bachman and O’Malley (1984) found that when responding to Likert-style scales African Americans used ERS more so than Caucasians. Bachman and O’Malley interpreted this to mean that Caucasians qualified their answers more than African Americans did, but they did not interpret this to imply African Americans were simply responding in an easier fashion. There is no evidence of the latter position but there is evidence for the former position. Clarke, Stanton, and Rao (1996) essentially replicated Bachman and O’Malley’s findings. Clarke (2000) then extended his inquiry of ERS to different cultural groups. He found Hispanics and African Americans were more likely to use ERS than Caucasians, but he also found that the French showed more ERS than the Australians. Marin, Gambia, and Marin (1992) also found that Hispanics demonstrate more ERS than Caucasians of non Hispanic ancestry. They also demonstrated that less educated Hispanics displayed more ERS than educated Hispanics. Support for the position that people’s culture influences their ERS is a study by Chen, Lee, and Stevenson (1995). They found their U.S. and Canadian samples demonstrated more ERS than their Japanese or Chinese samples. Considering all the above evidence, ERS apparently is a stable response style although one’s culture may influence it (Chen, Lee, & Stevenson, 1995; Cronbach, 1946; Gambia & Marin, 1992; Hamilton, 1968; Merrens, 1970).

A survey respondent may choose the same alternative (e.g., all strongly disagree, all agree) regardless of the content of each question (Cronbach, 1950; Singer & Young, 1941). For example, in using a five-point scale certain individuals essentially use position one and five, while others will use one, three, and five, and numerous individuals will use
the whole scale (Cronbach, 1946; Osgood, 1941). This rigidity of response was later hypothesized by Hershey and Hill (1976), but because of limitations to their study, they were unable to demonstrate it. Through appropriate data analysis, a survey researcher could detect such a pattern of responses. More troublesome, however, is the phenomenon that a respondent may engage in random responding.

As Beach stated (2001), “Random responding . . . may become a more serious problem as we increase our use of computer administered questionnaires.” He reported more random responses in his computer surveys than paper and pencil surveys. Studies have been done on versions of the Minnesota Multiphasic Personality Inventory (MMPI) to detect random responding or partial random responding (Archer, Handel, Lynch, & Elkins, 2002; Berry et al., 1992; Berry et al., 1991; Clark, Gironda, & Young, 2003; Pinsoneault, 1997). These authors noted the MMPI is especially susceptible to some random responding because of its extreme length. The Pinsoneault study was the only study to measure random responding without using the MMPI. Pinsoneault used the Jesness Inventory to detect random responding.

Aiken (1997) listed two minor threats to the validity of surveys: overcautiousness and oppositionalism. Aiken (1997) defined overcautiousness as “the tendency to be excessively careful in responding” and oppositionalism as “the tendency to respond in a direction opposite to that in which one actually believes” (p. 77). Aiken is the only author to list these specific threats to the validity of surveys. Other threats exist but only the major threats were discussed or will be discussed.
Social Desirability Bias Defined and How It Is Related to Survey Research Methods

The last and more often cited major threat to the validity of surveys to be discussed is SDB. Martin and Nagao defined SDB as “the tendency to stretch the truth in an effort to make a good impression.” (p. 264) Dillman (2000), Groves (1989), Fowler (1993) Belson (1986), Foddy (2001), Aiken (1997), and Hagan (2003) each considered SDB as a possible threat to the validity of surveys. As indicated above, it is not the only threat but it is a major one. However, in studying SDB, Paulhus (1991) revealed SDB possesses two distinct factors: IM, which is a conscious shaping of answers to project a positive appearance and Self-Deception Enhancement (SDE), which is an unconscious but natural tendency for “normal” people to think highly of themselves. However, Crowne and Marlowe (1960) defined SDB as the need for social approval. Marlowe and Crowne (1960), the developers of the Marlowe-Crowne scale, never said SDB is anything but one factor, the need for social approval. Therefore, SDB has historically been defined in a way that it meant essentially IM (Crowne & Marlowe, 1960; Paulhus, 1991). After all, to maximize social approval one must “manage” the impressions of others. Because SDE is not conscious, is too introspective to experimentally manipulate, and because SDB has traditionally meant the IM factor of SDB (Paulhus, 1991), for the practical purposes of this study, SDB is defined as IM.

For the purposes of possibly taking the first step toward ordering the main data collection methods, the effects of each of the main survey methods must be made in terms of SDB. Therefore, survey methods are compared using SDB as a criterion variable. The next four sections deal with hypothesis one. For this hypothesis there is
consistent but not indisputable evidence that on average face to face interviews elicit more socially desirable answers than telephone interviews.

Evidence a face to face interview elicits more socially desirable answers than a telephone interview (this section addresses the first part of hypothesis one):

First, Sykes and Collins (1988) found an inclination for face to face interviews to encourage more socially desirable answers than a telephone interview. Their rationale for this phenomenon was sound. They thought the telephone interview conveyed weaker social pressures than the face to face interview. Instead of using an instrument like the MCSDS or Paulhus’ BIDR, they operationalized SDB with questions about criminal and sexual activity. To be fair, questions about criminal and sexual activity are sensitive, but this may have magnified the social desirability effect. What is unknown is whether or not this magnification was equal in the face to face and the telephone interviews.

Second, Kormendi (1988) found more SDB in face to face than the telephone interviews. She asked questions about income to Danish citizens which could be verified by the tax authorities. Thus, while she too did not use the MCSDS or BIDR, she had access to the Danish tax authority’s data. She interpreted overreporting of income as SDB. She found no general tendency to underreport income although some subgroups, like women and the unemployed, did underreport their incomes. Nevertheless, she found more overreporting of income in the face to face than telephone interviews, which she interpreted as a social desirability influence. Because she used tax records, randomly selected respondents, and randomly assigned participants to groups, this study can be considered to be more rigorous than the above study by Sykes and Collins. She attempted
to do 1,000 telephone interviews and 2,000 face to face interviews. The reason for more face to face than telephone interviews is Kormendi expected more people not to be at home for the face to face, whereas in using the telephone interview, an interviewer could simply call later. Her response rates for telephone interviews and face to face interviews were 77.9% and 76.7%, respectively. This difference was not statistically significant even though the sample sizes were large.

Third, Colombotos (1965) also found more socially desirable answers by American physicians through face to face rather than telephone interviews, although his results were not statistically significant. He operationalized SDB, again not through the MCSDS nor BIDR, but by asking for the number of postgraduate classes they had taken in the last three years, how many medical journals they read consistently, and how many articles they had published. He presumed that the more doctors who were concerned with social desirability, the more they would report that they had done the above actions. He randomly assigned physicians to a face to face or a telephone interview, but did not report his sample size nor his effect size. He theorized the “social involvement” characteristic of the face to face interview would encourage more socially desirable answers than the telephone interview.

Fourth, the investigators Weeks, Kulka, Lessler, and Whitmore (1983) reported more accurate responses by telephone than for face to face interviews from former patients of a hospital. Like the Kormendi study, this meant they had a “gold standard” because they could check medical records from hospitals or other medical facilities for a validity check of their data. More accurate data by telephone than face to face interview
would imply more SDB in the face to face than the telephone interview. They found more disagreements in the reasons for hospital stays and in the durations of the hospital stay in the face to face than the telephone interviews. In fact, on average, their hospital stays showed an underreporting effect. They could not possibly randomly select nor randomly assign, because they had to obtain permission slips from their patients at the time of hospital stay or lab visit. However, because of these limitations they did analyses on subgroups. None of their results on different socio-demographic variables seemed to qualify their discovery that the face to face interviews elicited less accurate responses than their telephone interviews. There were 172 face to face interviews and 77 telephone interviews. Thus, while this study did have limitations, it also possessed data confirmed by hospital records.

Fifth, Pless and Miller (1979) reported more SDB as measured by their patients’ parents’ reports of usually using safety belts on their children. They assumed the biased—SDB—answer would be the parents had used safety belts on their children. They reasoned a biased answer would not underreport safety belt use when the parents had, in fact, used the safety belts. Conversely, falsely reporting they had used safety belts when they had not, would have been overreporting, a SDB. They compared three methods of data collection: face to face, telephone, and mail. In this section the only comparison at issue is the face to face versus the telephone interview. 17.5 percent answered “rarely” or “never” when answering whether they had used safety belts by face to face while 27.8 percent answered “rarely” or “never” by telephone. In their wishes to be conservative, they stated their three factor ANCOVA was not significant, but this does not necessarily
mean a correlated t-test between face to face and telephone interviews would not have been significant. They said their small sample size decreased their power. Additionally, the difference between the two most extreme means, which were for face to face versus telephone interview was .55 on a scale ranging from zero to six. The sample sizes for the face to face, mail, and telephone were 194, 184, and 137, respectively. They tried to randomly assign, but due to an overly industrious clerk who acted to reduce travel costs, the face to face interviews were not accomplished randomly since the clerk wanted to decrease interviewers’ travel times. Nevertheless, statistical adjustments were introduced to control for socioeconomic status variables and age. They used ANCOVA and found insignificant differences between the three methods. Because of the facts presented in this paragraph, one cannot say whether there would have been a difference between face to face interviews and telephone interviews. Still, one could say this study provides anecdotal evidence of a difference between face to face and telephone interviews.

Sixth, Rogers (1976) found, from a New York City sample, that people interviewed with a face to face method tended to give more socially desirable answers than people interviewed by telephone. Moreover, she indicated African Americans gave more socially desirable answers by face to face than telephone interview, but the method of interview did not matter to Hispanics. She randomly assigned participants to conditions. Furthermore, she validated her results by examining voting records from the board of elections. Her sample sizes were 90 for face to face and 81 for telephone interviews. Her study was exemplary except she operationalized SDB in terms of overreporting votes, not using the MCSDS or BIDR.
Seventh, Henson, Roth, and Cannell (1977) had participants answer two psychiatric inventories—the Midtown Symptom Scale and the Lubin Depression Inventory. The Midtown Scale was designed to measure the presence of psychiatric symptoms. A high score on the Midtown was undesirable for a non-psychiatric population. The Lubin Depression Inventory was designed to measure depression. Hansen, Roth, and Cannell administered these inventories to a general (non-psychiatric) sample in Kansas City, Missouri in 1972. They found no differences for the face to face and telephone interview groups on the Midtown Scale but did find the Lubin Depression scores were significantly lower by telephone than by face to face. Lower depression was expected on the Lubin scores if SDB effects were stronger in the face to face case. They could not explain their failure to find a difference in the Midtown Scale. However, they did randomly assign participants. The sample sizes for the face to face and telephone interviews were 476 and 409, respectively. They also used standardized instruments in measuring SDB, while the other studies idiosyncratically operationalized SDB. It is puzzling, given their population, why the MCSDS was not used.

Eighth, a dissertation by Thornberry (1976) using the general population of Rhode Island provides some evidence face to face elicits more SDB than a telephone interview. More specifically, Thornberry indicated that face to face interviews were greater in SDB than telephone interviews, which in turn were greater in SDB than his mail surveys. Again, the face to face has been shown to be greater in SDB than telephone interviews. However, there are three major weaknesses of his study. Firstly, he could not randomly assign to groups. Secondly, he did not preserve his method effects. That is, if he could not
get a response from a face to face interview, he would then use a telephone interview; and if he could not get a response from the telephone interview, he would use a face to face interview. Further, if he still could not get a response using both a face to face and a telephone interview, he resorted to a mail survey. Thirdly, he measured SDB by asking health related questions, which may or may not be socially desirable to different subpopulations. Nevertheless, he operationalized it, not as how participants answered the MCSDS or BIDR, but in his own idiosyncratic way.

**Summary of Evidence that SDB is greater in face to face interviews than telephone interviews (this addresses the first part of the first hypothesis):** The preponderance of the evidence stated in the summaries of the studies described above point toward the face to face being more prone to SDB than the telephone interview. Only Pless and Miller (1979) ever came close to contradicting this position, and they cited inadequate sample size as a reason they did not find a statistically significant result. Nevertheless, seven out of eight studies idiosyncratically defined and operationalized SDB. They defined it in their own way rather than using a standardized instrument such as the MCSDS or the BIDR. Because of this general weakness reigning throughout the studies, on average one cannot definitively declare face to face interviews elicit more socially desirable responding than telephone interviews. Besides, in this era of identity theft and telemarketing, one would assume consumers would be more “guarded” over the telephone than a face to face interview in answering surveyors’ questions. Therefore, why would consumers be more honest and candid (show less SDB) over telephone than a face to face interview, when in a face to face interview they could identify the interviewer or
at least have the interviewer show the respondents their identification? Moreover, SDB is context specific, meaning SDB depends on the purpose of the survey. Furthermore, a publication bias may exist because only statistically significant results tend to get published whereas non significant results do not tend to get published (Wolf, 1986). In short, there is “some” reasonable doubt, because of the reasons mentioned above, that face to face are more susceptible to SDB than telephone interviews. However, the “preponderance” of the accumulated evidence, despite its weaknesses, points toward a tendency SDB is more prominent in face to face than telephone interviews. Besides, a stranger is less of a stranger when one can “look him in the eye” rather than replying to a disembodied voice over the telephone.

Evidence that a face to face interview elicits more SDB than a mail survey (this addresses the second part of hypothesis one): The weight of the evidence indicates face to face interviews elicit more SDB than mail surveys (Cannell & Fowler, 1963; Cantrill, 1944; Ellis, 1947; Hinkle & King, 1978; Hochstim, 1967; Knudsen, Pope, & Irish, 1967; Krohn, Waldo, & Chiricos, 1975; Nederhof, 1984; Siemiatycki, 1979; Sudman, Greely, & Pinto, 1965; Wiseman, 1972; Zeiner-Henriksen, 1972). Only two studies indicated SDB was greater in the mail than face to face method (Assael & Keon, 1982; O’Toole, Battistutta, & Crouch, 1986). There was only one study that indicated there is no difference in SDB (McDonagh & Rosenblum, 1965). Before surveyors unequivocally conclude mail surveys elicit less SDB than face to face interviews, it should be emphasized the studies by Assael and Keon (1982), and O’Toole, Battistutta, and Crouch (1986) used administrative records to verify the accuracy of their responses. Yet, the
Siemiatycki (1979) study was also verified by administrative records, and concluded the opposite of the Assael and Keon, and O’Toole and Battistutta and Crouch studies, which stated mail surveys elicit less SDB than face to face interviews. Moreover, the studies by Cannell and Fowler (1963), Kronin, Waldo, and Chiricos (1975), Nederhof (1984), and Wiseman (1972) all used random assignment to either a face to face or mail group, which supports mail eliciting less SDB than a face to face interview. Conversely, no study supported face to face interviews eliciting less SDB, while simultaneously using random assignment to data collection methods. Even the study by O’Dell (1962), which indicated there was no significant difference in SDB between face to face interviews and mail surveys, did not use random assignment. Nevertheless, “all” the studies mentioned in this paragraph idiosyncratically defined SDB regardless of the direction of the bias. Again, no one used the MCSDS nor the BIDR.

During the 1990s only two studies were found that investigated the issue of what elicits more SDB—face to face interviews or mail surveys. Both studies were reported in de Leeuw’s dissertation. Her Meta Analysis showed slight but not trivial differences in SDB (average $d = .18$ with a 95% confidence interval between .14 and .22) and this bias was more prevalent in the face to face interviews than mail surveys (de Leeuw, 1992). De Leeuw based her Meta Analysis on 13 studies, including the studies described in the previous paragraph, but heterogeneity of studies was found. On the other hand, if she deleted the studies responsible for the heterogeneity, she could be shrinking an already small number of studies. Nevertheless, she replicated this result in a “field” study in the same dissertation. Again, she found face to face interviews were more prone to SDB than
mail surveys. This difference was also small but not trivial. For each data collection method a stratified random sample was selected (De Leeuw, 1992). Moreover, she randomly assigned participants and counterbalanced interviewers between face to face interviews and telephone interviewers, which controlled for interviewer effects. Surprisingly, her response rate for mail was 68 percent (N=400) and for face to face 51 percent (N=530), but she attributed this difference to respondents who could not otherwise be accessed. She stated despite the difference in response rate, for each method, the differences in her socio-demographic variables followed the same pattern.

De Leeuw (1992) did not mention what type of effect sizes she used. However, this study used partial eta squared ($\eta_p^2$). It was calculated this way:

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

(Equation 1)

Summary of evidence that a face to face interview elicits more SDB than a mail survey (this addresses the second part of hypothesis one): The preponderance of the evidence in the previous two paragraphs points toward a difference between the face to face and mail methods. Face to face interviews elicit more SDB than mail surveys. De Leeuw’s (1992) Meta Analysis, supporting this position, included all the studies mentioned in the previous two paragraphs except her own field study. She cited anonymity as a reason mail surveys elicit less SDB than a face to face interview. The number of studies favoring more SDB in face to face interviews rather than the mail surveys, and the methodological rigor of the face to face interviews being greater than the mail survey study set versus the opposite study set, all point toward the face to face eliciting more SDB than the mail method. A publication bias could be argued for the
studies mentioned in the previous two paragraphs, but many of the studies hypothesized there would be no difference in SDB elicited by both methods. Nevertheless, studies were found that dispute de Leeuw’s (1992) conclusion, but they had methodological weaknesses and were fewer in number when compared to the studies that did not dispute her conclusion.

_Evidence that telephone interviews generate more SDB than mail surveys (this addresses the third part of hypothesis one):_ The evidence overwhelmingly indicates a telephone interview elicits more SDB than a mail survey. Not one study suggested the mail method elicits more SDB than the telephone method. Only two studies declared the mail and telephone methods were equal in terms of SDB. They were Thornberry’s (1976) thesis and Locander, Sudman, and Bradburn’s (1976) study. However, several studies indicated telephone interviews elicit more SDB than mail surveys (Acree, Ekstrand, Coates, & Stall, 1999; Farnworth, Bennett, & West, 1996; Hennigan, Maxson, Sloane, & Ranney, 2002; Morrisey, 1995; Siemiatycki, 1979; Wiseman, 1972). Moreover, Frendrich and Vaughn (1994) concluded telephone interviews generate more SDB than SAQs, which were not mailed. Each of the previously mentioned studies except the Locander, Sudman, and Bradburn study did not use random assignment. Locander, Sudman, and Bradburn used random assignment. Nevertheless, there are additional studies that did, in fact, use random assignment while simultaneously indicating that telephone interviews elicit more SDB than mail surveys (Fournier & Kovess, 1993; Hall, 1995; McHorney, Kosinski, & Ware, 1994; O’Toole, Battistutta, Long, & Crouch, 1986). Moreover, in their reviews of the literature Dillman, Sangster, Tarnai, and Rockwood
(1996), Dillman and Tarnai (1991), Schwarz, Strack, Hippler, and Bishop (1991) and Tyebjee (1992) all indicated the same. Incidentally, the Dillman and Tarnai literature review included the Locander, Sudman, and Bradburn study—the one study that could have best persuaded Dillman and Tarnai that SDB is equal when comparing telephone and mail methods—but Dillman and Tarnai were not convinced. Instead, they stated in their literature review telephone interviews elicit more SDB than mail surveys. Finally, de Leeuw’s (1992) Meta Analysis indicated more SDB in telephone interviews than mail surveys. She observed a mean effect size of .12 with a 95% confidence interval of .06 to .19. Her field study in the same dissertation corroborated the telephone interviews elicited more SDB than mail surveys.

All but two studies that compared telephone interviews with mail surveys were about physical or mental health issues, weakening an apples to oranges argument. The two studies, not about health, were about criminal justice and individual’s relations with the police.

Finally, there is one caveat to mention. Like many previous sections, all the studies except the Fournier and Kovess study idiosyncratically operationalized SDB. Fournier and Kovess used the Diagnostic Interview Schedule Self-Administered (DISSA), an instrument used to assess mental disorders. While it is not the BIDR nor the MCSDS, it is a standardized instrument.

_Evidence that a mail survey should generate more SDB than a survey administered through a computer [this will help address the fourth part of hypothesis one (appearing on page 54)]:_ Because the mail method can share a myriad of characteristics
with the paper and pencil method, a discussion of the differences between computerized and paper and pencil methods would be almost equivalent to comparing computerized and mail methods. However, there are two caveats. First, a mail survey must be mailed, but a paper and pencil survey must not necessarily be mailed. Second, most of the time, adults can fill out a mail survey anytime and anywhere they want. Yet, some paper and pencil surveys must be completed in a classroom or a lab. Nevertheless, in most classroom and lab settings, the participants can privately complete their surveys. Therefore, mail and paper and pencil methods are principally equivalent.

What follows are comparisons of computerized and paper and pencil surveys. Kiesler and Sproull (1986) and King and Miles (1995) studies both concluded that a paper and pencil survey, which can be used in mail surveys, elicits more SDB than a computer survey. These two studies share the weakness of not having been randomly assigned, but they also share a strength all the previous studies did not possess, they operationalized SDB using a standard instrument, rather than in their own idiosyncratic manner, as most of the previous studies had done. Kiesler and Sproull (1986) used MCSDS, while King and Miles used Paulhus’s BIDR. A study by Lapham, Kring, and Skipper (1991) used neither a standard instrument nor random assignment, but they did have the advantage of urine analyses of pregnant women tested for the smoking of cigarettes, drug abuse, or alcohol abuse—all socially undesirable acts. They too determined the paper and pencil method elicited more SDB than a computer method.

Four studies found the mail method was not significantly higher in SDB than a computerized survey. First, Edwards, Rosenfeld, Booth-Kewley, and Thomas did use the
BIDR, but did not use random assignment on their U.S. Navy participants. They stated the two methods were equal in SDB. Second, Miles and King (1998) found no significant difference between computerized and mail questionnaires. They used the BIDR but did not use random assignment. Third, Booth-Kewley, Edwards, and Rosenfeld (1992) found no significant differences in IM or SDB for male naval recruits. They did use the BIDR and random assignment. Fourth, Potosky and Bobko (1997) were the only ones who found computer surveys—not connected to the Web—elicited more SDB than mail surveys. They too used the BIDR but did not use random assignment.

Dahl’s (1992) dissertation further supports the hypothesis that mail surveys elicit more SDB than computer surveys. She used the MCSDS and did use random assignment. This was a strong study. Furthermore, a Meta Analysis by Weisband and Kiesler (1996) demonstrated mail elicits more SDB than computer surveys. Using 32 studies, they found an adjusted (the outliers were deleted) effect size of .21 favoring mail surveys as having more SDB than computer surveys. This gave a z-score of 6.85 (p < .001) with a 95% confidence interval of .15 to .27. The unadjusted results were even more divergent. Based on 39 studies they obtained a z-score of 11.76 (p < .001). The mean effect size was .33 with a 95% confidence interval .28 to .39. However, they found heterogeneity of variance and had to reanalyze the data. The adjusted results are reported above and still point toward mail surveys eliciting more SDB than computer surveys.

*Theory and evidence that a mail survey tends to elicit more SDB than a survey administered through the Web (this addresses the fourth part of hypothesis one):* Davis (1999) reported more “self-focused rumination” (p. 575) using a Web versus a mail
method. This implies mail surveys elicit more SDB than a Web-based survey. Davis used the Ruminative Response Scale, but not the MCSDS, BIDR, nor random assignment to methods. Nevertheless, Joinson (1999) also found mail surveys elicited more SDB than Web-based surveys. He did use the BIDR but not random assignment to methods. Furthermore, as Abilgaard (1999) found in her dissertation, mail surveys elicit more SDB than Web-based surveys. She, too, used the BIDR but not random assignment of methods. Therefore, based upon this limited evidence and the evidence about computerized surveys, it can be assumed that mail surveys elicit more SDB than Web-based surveys. The logic of this particular hypothesis is since computerized surveys elicit less SDB than mail surveys, Web-based surveys should elicit comparable, if not less, SDB as solely computerized surveys. In addition, however, Web-based surveys can utilize a secure socket. This increased “sense” of privacy should even further decrease SDB.

This next section addresses hypothesis two (an interaction between method and gender on IM): There will be no interaction between method of data collection and gender. A two factor unbalanced ANOVA was used with the four levels being face to face, telephone, mail, and Web-based while the two levels were male and female. The evidence which supported this hypothesis was as follows:

Miles and King (1998) found no interaction between SDB and gender in their method comparisons. Miles and King (1998) compared data collected from computers in a lab with data from a paper and pencil classroom administration. Moreover, Edwards (1957) in a no method comparison study found no differences in SDB between males and
females. Furthermore, Abilgaard’s (1999) dissertation showed females engaged in more IM in a Web-based survey than an in-class survey. Abilgaard used the BIDR, but did not use random assignment. She also reported a much greater number of males participated in her study. So, based on all of the available evidence, it was expected that no interaction between methods of data collection and gender would be found.

The next section addresses hypothesis three (no interaction between method and ethnicity on IM): There will be a significant interaction between ethnicity and method of data collection. A two factor unbalanced ANOVA was used with the four levels being face to face, telephone, mail, and Web-based, and the two levels being Caucasian and non-Caucasian. The rationale was as follows:

Aquillino (1992, 1994), Rogers (1976), and Zhang and Gerstein (1999) all found an interaction between race/ethnicity and method of data collection, but Fowler, Gallagher, and Fowler (1997) and Greenfield, Midanik, and Rogers (2000) found no interaction. However, the Aquillino (1994), Rogers, and Zhang and Gerstein studies used random assignment while the other studies, including those studies by Fowler, Gallagher, and Fowler (1997) and Greenfield, Midanik, and Rogers (2000) did not. Furthermore, Zhang and Gerstein used urine analyses as a validity check. All three studies that used random assignment support an interaction effect. Additionally, with the exception of the Rogers study, the context of the study was the use or abuse of drugs or alcohol. Moreover, with the exception of the Aquillino (1994) study, the methods compared were telephone and face to face interviews. Furthermore, with the exception of the Zhang and Gerstein study, all of the studies in this paragraph were sampling from the general public
unlike the majority of studies in this literature review, which used undergraduates. The
Zhang and Gerstein study sampled arrested drug users. Therefore, because of the similar
substantive characteristics of the studies, an apples to oranges argument is not strong
regarding the six studies mentioned above. So, because of the literature cited above, an
interaction between method of data collection and ethnicity was expected in this study,
especially between telephone and face to face interviews.

The next section will address hypothesis four (no method by gender by ethnicity
interaction on IM): As indicated in Askenasy’s (1978) abstract of his conference paper,
there was no mention of a three factor interaction, as measured by Edward’s Social
Desirability Scale, of SDB, gender and ethnicity. Every reasonable attempt was made to
locate this conference paper, including writing to the author, but the actual study could
not be located. In his abstract Askenasy did not mention a three factor interaction
between gender, ethnicity, and social class on SDB, but did mention main effects of
gender and ethnicity on SDB. While this is not proof there was no three factor interaction
between method of data collection, gender, and ethnicity, there is no evidence an
interaction between method, gender, and ethnicity on SDB existed. Likewise, there was
no a priori reason to suppose a three factor interaction existed.

The hypothesis of a three factor interaction was the weakest of the four research
hypotheses. However, there is no evidence for a three factor interaction. Therefore, for
this study a three factor interaction was not expected, but one could not unquestionably
assert this due to the lack of literature.
Specific Threats to the Validity of this Study as Determined by a Review of the Literature

There are five major, identified threats to the internal validity of this particular study. To organize these five major threats their associated paragraphs are labeled A, B, C, or D.

A. Even if a survey is truly anonymous, the respondent may not perceive or “feel” that the survey is actually anonymous (Hyman, Cobb, Feldman, & Stember, 1954; Koson, Kitchen, & Stolodosky, 1970; Krohn, Waldo, & Chiracos, 1973; Nass, Moon, Fogg, Reeves, & Dryer, 1995; Supple, Aquillino, & Wright, 1999; Synodinos et al., 1994). Specifically regarding Interviewer Administered Questionnaires (IAQ), the actual presence or conversely the absence of an interviewer can affect how participants answer a question (Cannell & Fowler, 1963; Krohn, Waldo, & Chiracos, 1973). Moreover, if respondents feel their confidentiality has been compromised, they respond to protect their self-images (Koson, Kitchen, & Stodolosky, 1970). Because the participants of this current study were randomly selected, it is not possible to have them respond truly anonymously. To have the participants participate one has to randomly select them, identify them, and obtain their postal address, email address, or their telephone number. This would not preclude at least some participants from feeling “psychological” anonymity, especially through a secure socket. So, they might feel a sense of anonymity even though, because of random selection, their responses cannot be truly anonymous.

B. If, as cited in Dillman (1978) and Dijkstra (1987), the respondents were properly motivated to focus solely on the task at hand, they would be less likely to yield to SDB influences. The interviewers can persuade the participants to adopt the above
orientation (Assael & Keon, 1982; Holbrook, Green, & Krosnick, 2003). It is also possible for participants to feel a lack of motivation despite repetitive attempts to motivate them, because they still may respond mechanically (Singer & Young, 1941). Because of the threat to internal validity mentioned above, the interviewers were instructed not to excessively persuade their telephone and face to face respondents to participate, because there would have been no excessive persuasion in the Web-based and mail methods.

C. This group of threats is associated with three possible problems found in survey research. The first problem is coverage error. That is, telephone interviews are less likely than the face to face interviews to include drug users, street dwellers, transitional people, low-income people, and people from rural communities, because many of these people do not own phones or have a phone line (Gribble, Miller, Rogers, & Turner, 1999; Telser, 1976). Undergraduates are not as likely to be drug users or street dwellers. However, they are transitional and have little, if any, income. Nevertheless, at least 97.3% of undergraduates attending Kent State’s main campus in Fall 2006 had access to a telephone (L.A. Carlson, personal communication, February 6, 2009). Thus, there should be little coverage bias for the face to face and telephone comparisons because a telephone call may have been required to schedule interviews. The second problem is inappropriate verbal and nonverbal feedback by the interviewer, when the interviewer is supposed to be neutral (Schober, 1999). This problem was dealt with through appropriate interviewer training. Third, falsification of interviews by the interviewers is also a potential threat (Cannell, Miller, & Oskenberg, 1981; Forsman and Schreiner, 1991).
Because of this falsification threat, numerous interviewers were used to decrease interviewer workload. Workloads being too high have been associated with falsification by both the Cannell, Miller, and Oskenberg (1981) and Forsman and Schreiner (1991) studies and a quality check was conducted before proceeding with any data analyses to detect probable falsification.

D. The surveys could be measuring something other than SDB. There are eight attributes that could confound the measurement of SDB. First, participants may excessively use response alternatives in comprehending the meaning of a question (Gaskell, O’Muircheartaigh, & Wright, 1994; Gruder, 1977; Rockwood, Sangster, & Dillman, 1997; Schwarz, Bless, Bohner, Harlacher, & Kellenbenz, 1991; Schwarz & Hippler, 1987; Schwarz, Strack, Muller, & Chassein, 1988; Sudman, Bradburn, & Schwarz, 1996; Wheeler et al., 1969). Therefore, three “rigged” questions were asked to detect if participants excessively used the response alternatives in skimming the questions instead of fully comprehending what was asked. For example, one set of “leaf” alternatives for a question, “a stem,” was: 1) Please 2) choose, 3) this, 4) particular 5) option. Second, if all IM scores are “assumed” to be disingenuous, the ones who strive for honesty and morality are being unfairly penalized (Lindeman & Verkasalo, 1995; Sudman, Greely, & Pinto, 1965). However, it would not especially affect the means. Furthermore, random assignment should equalize this “morality” effect among data collection methods. Third, if all high IM scorers are classified as being dishonest, the ones with high ego strength could be unfairly penalized (Edwards & Heathers, 1962). Nevertheless, it also should not affect the means. Fourth, some of the questions may not
apply to certain people with the result being a confused answer(s)—even to the
respondent (Fowler & Gallagher, 1997; Marquis, Marquis, & Pollich, 1983). Therefore,
at the end of each question a “not applicable” option was available. Questions answered
“not applicable” to a sizable proportion of the participants could be deleted if they proved
problematic. Fifth, some individuals predominantly use an Extreme Response Style
(ERS) as indicated by Cronbach (1946) and Hamilton (1968). For example, in using a
five-point scale, a respondent may use a “one or five” or a “one and five.” Therefore, the
data was analyzed for both types of ERS effects and if it had been deemed appropriate,
the extreme values would have been adjusted statistically. Specifically, the variances of
the different methods will be compared to detect ERS. Sixth, participants may slant their
responses to appear like an “average man” (Aiken, 1997; Maynes, 1965). Therefore,
analyses revealing central tendencies were performed. The questions will also be
examined by judges in terms of content whether there were extreme points the “average”
person would be predisposed to select. Seventh, some participants may use SDB
hesitantly. That is, they only use it when the direction of the interview has becomes
readily apparent (Back & Gergen, 1963; Sudman, Greely, & Pinto, 1965, Sutherland &
Spilka; 1964). Therefore, for surveys and interviews, the later halves of the surveys were
compared with the first halves of the surveys to determine if hesitant ingratiation did
likely occur. Eighth, there may be no “substantive” difference between a mail-in/mail-
back survey and a structured interview (McDonagh & Rosenbloom, 1965). However, the
results of the Ryan test, which will have compared all four different methods of
collecting data, will test this issue because both types of interviewers—face to face and telephone—were instructed and trained to give a structured interview.

Of the studies mentioned thus far in this section, which identify specific threats to the internal validity of this study, only the Lindeman and Verkasalo (1995) study used the BIDR as this study has. The remaining studies defined SDB in their own idiosyncratic ways. Consequently, much more confidence can be placed on the existence toward an SDB ordering, if one had existed as hypothesized in Chapter I, than those few studies mentioned previously in section D. Furthermore, even if those few studies mentioned above were without any methodological flaws, the ordering of the methods, or parts of it, could still exist.

There are also three specific threats to the external validity of this study. First, Cotter, Cohen, and Coutler (2001) found, regarding some of their race related questions, the race of the interviewer affected how the questions were answered. Specifically, when the interviewers and participants were of the same race, the respondents tended to disclose more than if the two were of different races. However, they also found, regarding questions of a non racial nature, the race of the interviewer had no effect on respondents’ answers. Since the BIDR has questions of a non racial nature, the specific threat identified by Cotter, Cohen, and Coutler does not appear applicable to this study.

Second, Latinos answer questions differently than European Americans. Latinos have a concept called “simpacia,” which means concern for the group as a whole. Therefore, they might answer surveys differently (Shultz & Chavez, 1994). If any
identifiable Latinos answer the survey, their pattern of answers will be scrutinized carefully and any actions deemed appropriate will be taken.

Third, African Americans have been shown to answer questions using the extremes of a scale more so than Caucasians (Clarke, Stanton, and Rao, 1996). This will be carefully scrutinized and, if deemed necessary, statistically adjusted for.

It is possible there could be more threats to the internal and external validity of this study, but a review of the literature consisting of 388 sources revealed only the preceding specific threats. As Kish (1965) commented regarding specification errors,

First, even after controlling every possible variable, all the potentially confounding variables are not ordinarily removed. Second, neither the advance of science nor the application of statistical tests can wait for the control of all relevant, but still uncontrolled and confounded, factors.
CHAPTER III
METHOD

Research Design

A randomized post-test only control group design was used including the following groups: face to face, telephone, mail, and Web-based. This design was recommended by Kerlinger (1986), Rudestam and Newton (2000), and Isaac and Michael (1995) to compare treatments, because “randomization techniques permit the experimenter to declare at the time of assignment the groups were equal” (Isaac & Michael, 1995, 75).

Participants

Undergraduates enrolled at the main campus of Kent State University were randomly selected. The Office of Research, Planning and Institutional Effectiveness supplied students’ postal addresses, phone numbers and email addresses. For the telephone, mail, and Web-based methods all participants had a chance to win three $250 prizes, one for each method. For the face to face interview group only, because of the difficulties in getting enough responses, participants were compensated in two different ways: Firstly, 28 participants had a chance to win one $250 prize, and secondly, 146 others were given a five dollar bill immediately after participating. All participants were randomly assigned to a data collection method mentioned above. For research question one, a total sample size of 580 (121 for each of the four data collection methods [580 = 121 x √2]) because for a targeted effect size of .2 power equals about .92 at α = .05 for a directional t-test (Cohen, 1988). For research questions two and three, a sample size of
819 could have yielded a power of .82 at $\alpha = .05$ for a two factor ANOVA for a targeted effect size of .14 (Faul & Erdfelder, 1992). According to the Office of Research, Planning and Institutional Effectiveness, about 81 percent of Kent State’s population was Caucasian and 19 percent was non-Caucasian (W.G. Schneider, personal communication, August 6, 2004). Therefore, for research question four, a sample size of 819 could have yielded a power of .82 at $\alpha = .05$ for a three factor unbalanced ANOVA for a targeted effect size of .14 as recommended by Faul and Erdfelder (1992).

Sampling

The target population was undergraduate students attending the main campus of Kent State University. An empty cell in ANOVA can make interpretation more difficult because researchers are forced to use cell means model(s) as derived by Searle (1993) (Godbout et al., 1975; Jennings & Green, 1984; Kirk, 1995). The particular pattern(s) of empty cells in the cell means model(s) may not address the research questions posed by the researcher, and, therefore, empty cell models should be avoided by researchers if possible (Halpin, et al., 1991; Keren & Gideon, 1993; Jennings & Green, 1984; Kirk, 1995; Searle, 1993; Spinner & Gabriel, 1981). A random selection of students along with their postal addresses, email addresses, and phone numbers was provided by Kent State’s Office of Research, Planning and Institutional Effectiveness. Random assignment to the four groups was applied. In addition, more participants from all four groups were selected than was actually needed—an oversampling. One hundred thirty-eight participants per group, which requires at least 552 (138 x 4) total participants, were required to obtain at least a power of .79 for research question one at $\alpha = 0.05$; for research questions two and
three at least 164 participants were required for a power of 0.80 at $\alpha = 0.05$. For research question four at least 186 participants were required for a power of .80 at $\alpha = 0.05$. As will be discussed shortly, all power and sample size calculations were based on a projected effect size of .2, which was .12 effect sizes below the average effect size of the two different sets of studies: Web-based compared to mail, and computer-only compared to mail. The mean effect size of the Web-based as compared to mail methods is based upon the effect sizes reported in or calculated from the following studies: Abilgaard (1999); Hancock and Flowers (2001), and Joinson (1999). The average effect size difference for the Web-based compared to mail methods is .5. The average effect size difference, favoring the computerized-only surveys—revealing less SDB—as compared to mail, is .13. A graphic depiction of the ordering of methods along with their average effect sizes is illustrated in Figure 1 below.

```
<table>
<thead>
<tr>
<th>Personal Visit</th>
<th>Telephone</th>
<th>Mail</th>
<th>Computerized Only</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td>unknown</td>
<td>.13</td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 1. Depiction of social desirability effect size differences among different methods.

The average effect size for computerized-only (not connected to the Web), as compared to mail, is based upon the effect sizes reported or calculated from the following studies: Lautenschlager and Flaherty (1990), Mitchell (1993), Potosky and Bobko (1997), Rosenfeld (1989), and Wilkerson (2002). The average of .5 and .13 is .32. Consequently, the average effect size of the two groups of studies is .32. Therefore, a targeted effect size
of .2 seems adequate, especially considering that the face to face and telephone
interviews, according to the research cited on pages 3-7, are not accounted for in the
average effect sizes. Moreover, according to Kerlinger and Lee (2000), and Mendenhall
& Beaver (1994), it is permitted for researchers to take the range of effect sizes and
divide by four. The absolute value of this range of the previously cited studies is
1.71. Therefore, if 1.71 is divided by four, the effect size becomes .43 (1.71/4=.43).
However, .2 was chosen to be more conservative.

Bausell and Li (2002) and Lipsey and Warren’s Meta Analysis of Meta Analyses
(2001) of behavioral, psychological and educational effectiveness studies indicate a mean
effect size of .5 and a median of .47 with 99% confidence intervals between .42 and .52.
Lipsey’s (1990) Meta Analysis on the same subjects indicated a 99% confidence interval
between .38 and .52. Kraemer and Thiemann (1987) concur with the mean effect size of
.45 with a sample size of 70. Lipsey and Warren (1993, 2001) indicated, in the absence of
any information on effect size, a researcher may assume an effect size of .5. Therefore,
from the above studies the mean effect size that can be assumed with a lack of data can
be justifiably argued to be between .45 and .50. However, in this instance, as can be seen
in this paragraph and Figure 1 on the previous page, the effect size can be as low as .38 or
as high as .77.

The .77 effect size may be conservative, because according to the literature, the
two highest methods in SDB are not even represented in Figure 1 on the previous page.
Aquilino (1992), de Leeuw, (1992), Fournier & Kovess (1993), Groves (1989), and
Schwarz (1991) indicated face to face and telephone interviews are higher in SDB than
mail surveys. In addition, mail surveys have shown greater SDB than Web-based methods (Acree, Ekstrand, Coates, & Stall, 1999; Smith et al., 1992). How much more SDB the face to face and mail methods elicit than the Web-based method is not known, because either the studies did not use the BIDR or they contained methodological flaws.

**Procedures**

**Instrument**

Paulhus’ Impression Management (IM) subscale of the Balanced Inventory of Desirable Responding, version 6 (BIDR-6), was used in this study. A specific factor of SDB is IM. For the purposes of this investigation, which dealt with a non-clinical population, a variable such as Self-Deception Enhancement (SDE) is too introspective to be investigated. Furthermore, when researchers such as Marlowe and Crowne first studied SDB, their primary purpose was to study the IM factor. Marlowe and Crowne developed their scale to measure a single factor, the need for social approval (Crowne & Marlowe, 1960). Yet, factor analyses revealed it was not a single factor (Loo & Thorpe, 2000; Paulhus, 1991; Smith, 1997). Paulhus (1991) found the Marlowe-Crowne Scale loads on both the IM and the SDE factors. The separation of the two factors resolves some of the discrepancies in the psychological, personality “adjustment” literature, verifying the MCSDS measures two, not one, factors. SDE is correlated strongly and positively with adjustment but IM is not correlated with adjustment. Since Paulhus’ BIDR divides these two factors but the Marlowe-Crowne does not, Paulhus’ inventory was chosen. The following section comments on the reliability and validity of IM.
Reliability and Validity of the IM subscale of the BIDR

Paulhus’ Balanced Inventory of Desirable Responding (BIDR) is a 40-item, five-point scale scored from 1 to 5 with 1 meaning “not true” and 5 meaning “very true.” In 11 studies the BIDR’s IM subscale’s reliability ranged from .65 to .86 (Kroner & Weekes, 1996; Holden, Starzyk, McLeod, & Edwards, 2000; Kwong-Liem, 1999; Meston & Heiman, 1998; Paulhus, 2003; 1991, 1988; Paulhus & Reid, 1991; Lajunen, Corry, Summala, & Hartley, 1997; Mitchell, 1993; Vispoel & Forte, 2000; Vitelli, 1999).

A study of undergraduate alcohol consumption, which compared mail surveys with Internet responses reported, “The test-retest reliabilities ranged from .59 to .93” (Miller, et al., 2002, p. 59) depending upon the conditions of their study. They used intraclass correlation coefficients. The Web group ranged from .41 to .93 and the mail survey group ranged from .55 to .89. Finally, the test-retest reliabilities of the BIDR over a five-week period were .65 in 1991 and .76 in 2003 for a four-month time period for solely the IM subscale (Paulhus, 1991; 2003). Considering the different contexts in which the IM subscale of the BIDR was successfully applied, it seemed reliable at least in a general sense. Furthermore, the internal consistency of this particular study was computed using coefficient alpha of the BIDR-6.

The evidence of the validity of the BIDR-6 is as follows. Lajunen et al. (1997) found two factors in SDB in a traffic context. However, Kroner and Weekes (1996), using a confirmatory procedure, indicated no statistical difference between a two and a three factor solution. Nevertheless, Smith (1997), in her Structural Equation Modeling
study, concluded SDB is composed of two latent but highly correlated “.53” factors (p. 22).

The concurrent validity of the entire BIDR with the Marlowe-Crowne Social Desirability Scale (MCSDS) was .71, and the concurrent validity with the Multi-Dimensional Social Desirability Inventory was .80 (Paulhus, 1991). Paulhus asserted convergent validity by stating, “The IM scale correlates highly with a cluster of measures traditionally known as lie scales (e.g., Eysenck’s Lie scale, MMPI Lie Scale) and role-playing measures (e.g., Wiggins’ Sd, Gough’s Gi).” Additionally, Davies, French and Keogh (1998) reported, “The Lie scale was correlated . . . even more significantly with Impression Management (r = .61)” (p. 16). Paulhus (1991) asserted discriminant validity by noting his two scales, IM and SDE, were independent of each other. The Marlowe-Crowne possesses both factors, but both are highly intercorrelated, Paulhus (1991) stated, “The version presented here, however, exhibits much lower correlations, ranging from .05 to .40” (p. 39). Moreover, in a study of undergraduates faking mental illnesses, Holden, Starzyk, McLeod, and Edwards (2000) found that the congruence coefficient between the observed and theoretical target matrix was .91 for IM. Using confirmatory factor analyses, their findings support the construct validity of the Balanced Inventory of Desirable Responding.

Paulhus (1991) argued for the discriminant validity of his BIDR by stating “males score higher on SDE, but lower on IM” than females do (p. 39). Paulhus also found the BIDR measures two distinct factors: SDE and IM. Furthermore, Paulhus argued for the BIDR’s discriminant validity by stating, “IM, but not SDE was sensitive to test
administration conditions . . . public vs. private” (p. 39). For example, since IM concerns impressions, one would expect, and Paulhus found differences like IM is higher in public than private conditions. Considering all of the above evidence, the IM subscale of the BIDR can be said to be an, essentially valid instrument, for measuring IM.

Additional Questions Other Than the IM Subscale

Along with the BIDR’s IM subscale the surveys and interviews asked additional specific questions. First, to gauge whether nearby bystanders influenced the participants’ answers, the surveys or interviews asked if there were any bystanders within a specific distance when the participants completed the interview or survey. Second, a not applicable (N/A) option followed each of the alternatives in all of the questions. This option’s availability lowered Cronbach’s alpha and may have lowered validity as well. Third, an option entitled “I prefer not to answer” because of Human Subject concerns was placed on each question. Fourth, the surveys or interviewers asked (or the face to face interviewer observed) how many individuals, other than the participants, were within earshot or within a short distance—as in the case of mail or Web-based surveys—so bystanders could see or hear the questions asked and the participants’ answers. Fifth, if the participants answered yes to the fourth specific question, they were asked if the presence of another individual affected how they answered some of the questions. Sixth, three questions were “rigged” to detect if, in fact, participants predominantly focused on the answer alternatives instead of looking at the question as a whole. For example, one question had this stem, “Which word is a verb? It had a set of “leaf” alternatives ”
1) Please 2) choose, 3) this, 4) particular and 5) option. The next two “rigged” questions were:

1. Which of the following numbers has the greatest absolute value?
   -400   -300   -200   -100   0   333 ½

2. When was the Declaration of Independence for the United States signed?
   1576   1676   1776   1876   1976   1787

Seventh, demographic questions were placed at the end of the survey or interview, following Dillman’s (2000, 2007) recommendations. A participant’s gender was asked even in the face to face interviews because, at least for a few participants, their gender may not have been readily apparent from the interviewers’ direct observations. Moreover, participants were asked their age in years. Participants were also asked their ethnicity using the following classifications: Caucasian, African American, Hispanic, Asian, foreign, Native American, and unknown—because of the few numbers of minorities at Kent State. These ethnic classifications corresponded to the classifications used by Kent State University’s Office of Research, Planning and Institutional Effectiveness. Additionally, for the face to face and telephone interviews, the first question asked the time the interview started and the last question asked the time the interview ended. Toward the end of each survey participants were asked two questions for investigating if bystanders had influenced the Web-based or mail participants’ answers to the surveys. The two questions asked were:

1. Were there people within five feet of your computer monitor who could have seen the questions and how you answered them?
2. If there were people within five feet of the monitor, did the presence of these other individuals influence the way you answered the questions?

Scoring

One may score the BIDR either continuously or dichotomously (Paulhus, 1998; Endler & Parker, 1998). Paulhus (1998) and Kwan and Vitelli (1998) indicated if a 7-point scale is used, the BIDR can be scored dichotomously. However, they also stated the BIDR can be scored continuously if a 5-point scale is used. According to the literature on scaling, should scales be changed from a seven to a five point scale, the resulting reliability will decrease or even remain the same, but is unlikely to increase (Clarke, 2000, 2001; Lubke & Muthen, 2004; Bernstein & Teng, 1989; Green & Goldman, 1993; Daamen, 1992; Stober, Dette, & Musch, 2002; Hayes, 1998; Ashkanasy, 1985). Because of the literature above, and a consequence of using dichotomous rather than continuous scoring will result in an information loss, a continuous 5-point scoring scale was used.

Settings of the Study

Since one independent variable of this study was method of data collection, the settings for the participants varied with the method. For face to face interviews, the settings were the participants’ dorms, on-campus apartments, off-campus apartments, study lounges, residence hall lounges, kitchen areas, or other places. Moreover, as indicated in Chapter II, face to face interviewers were instructed to give participants a choice in where they wanted to continue the interview: the original location or another area offering more privacy. Furthermore, the interviewers observed and recorded how many people were within earshot of the interview.
For the telephone interviews, the setting was most likely the students’ dorms or off-campus apartments. However, it is likely at least some participants answered their telephones in a place other than their dorms or apartments because of cellular phones, call forwarding, or because of personal circumstances. As previously indicated, the telephone interviewer asked the participants how many people were within earshot of the telephone interview. However, the effects of other people “listening in” on the telephone interview should have been less than the face to face interview, as indicated in Chapters I and II, because other people could not have heard the questions raised.

For the mail survey, the settings were most likely the students’ dorms or apartments because that is where the surveys were mailed to. However, because the surveys were self-administered, the response setting may not have been where the students lived. After all, sending the surveys to the students’ apartments or dorms would have not prevented them from filling out the survey in classrooms, on buses, in libraries, in a room at the Student Union, at a friend’s house, or any other place. These locations could have been more or less private than having the students fill out their surveys in their dorms or apartments. Therefore, a question was asked about how many bystanders there were when participants gave interviews or completed their surveys. Furthermore, questions were asked if there were any other bystanders present when they filled out the survey, and whether these bystanders could have influenced how the students answered their questions.

For the Web-based method the settings were most likely students’ dorms or apartments because that is where they were emailed. However, because the surveys were
self-administered, the setting in which participants responded to the survey may not have been where the students lived. After all, sending the surveys to the students’ apartments or dorms did not prevent them from filling out the survey in a classroom, in a library, in a room at the Student Union, at a friend’s house, or any other place. Again, because the questions and subsequent answers would have appeared on a computer monitor, the following questions were asked similar to the mail questions previously mentioned:

1. Were there people within five feet of your computer monitor who could see the questions and how you answered them?

2. If there were people within five feet of the monitor, did the presence of these other individuals influence the way you answered the questions?

These were among the last questions asked.

Web Site Considerations

All the Web-based data was collected on a secure server. All data was stored behind a firewall. Each participant was only allowed to submit one survey, and the participant did this anonymously. The researcher did not have access to names, social security numbers, or other identifying information because of anonymity and confidentiality concerns.

Independent Variables

One independent variable of this study was the method of data collection, with the levels having been face to face interview, telephone, mail, and Web-based survey. The other independent variables were gender and ethnicity. Gender was either male or female.
Because of practical data analysis concerns, ethnicity was transformed to Caucasian or non-Caucasian.

**Dependent Variable**

The dependent variable was the IM subscale scores of the BIDR-6. It was scored continuously on a five-point scale.

**How Non Response Was Addressed**

The Web site was programmed to decrease item non-response by reminding participants to answer all unanswered questions before submitting the entire survey to the Web site for data analysis. The Web site also required a participant to attempt to answer all questions before they were entered in a drawing. The face to face interviewers were instructed not to enter people in the drawing or give $5 cash to participants until the interviewer had asked each question and to have had participants try to answer all questions. Participants, however, were allowed to refuse to answer any question because of human subjects concerns. The mail survey was designed to minimize item non-response as much as possible following Dillman’s (2000, 2007) recommendations. For the telephone interviews, the interviewers were instructed to “impartially” motivate the participants to answer all questions. Moreover, because of the safeguards mentioned above, an oversampling was executed.

Furthermore, a large amount of entire questionnaire non-response was anticipated. To avoid this sort of non-response a pre-notification email and two follow-up surveys were sent to all four message groups.
Procedure

There were many steps in completing this study. They were as follows:

1. Judges, who all have Ph.D.s in Education or Psychology, assessed if there were any questions the “average” individual would have answered extremely to counteract any effects as found by Maynes (1965) and Aiken (1997) that people desire to appear like an average individual. If there were, consideration was taken to modify or delete the questions from the study.

2. A “viability” study with the Web-based survey was implemented. This viability study was not a pilot study because this particular sub-study only assessed certain attributes of the Web site such as:
   A. Was the Web site able to prompt potential participants to enter their data through a wide variety of computers?
   B. Did the Web site avoid multiple submissions from a particular computer within a short time period—thus avoiding “ballot stuffing”?

3. Interviewers were trained to conduct structured interviews because Foddy (2001) and Groves (1989) indicated that interviewers can affect the standardization of the interview. Five steps were taken to facilitate standardization of all interviews:
   A. All interviewers were properly instructed to give structured interviews.
   B. Interviewers were specifically trained not to excessively persuade the participants to be interviewed because there was not excessive persuasion in the Web-based or mail methods.
c. For the face to face interview, interviewers were trained to avoid inappropriate verbal and nonverbal feedback.

d. For the telephone method, the interviewers were trained to avoid giving inappropriate verbal feedback.

e. The interviewers were instructed to persuade the participants to attempt to answer each of the questions, and were instructed to make a second attempt to persuade the participants to answer all unanswered questions. However, participants were allowed to refuse to answer any questions they did not want to answer because of Human Subject concerns.

4. A firewall was constructed. This enhanced the security of the data for the Web-based method.

5. A random selection of students attending Kent State’s Office of Research, Planning and Institutional Effectiveness was obtained. All potential participants were subsequently randomly assigned to one of four methods: face to face interview, telephone interview, mail survey, or Web-based survey.

6. A pre-notification email was sent to all participants in the face to face interview, telephone interview, and mail surveys to the addresses Kent State University gave them. All participants in the Web-based method were also sent a pre-notification email to the email addresses Kent State provided to them. Sending pre-notifications provided a ballpark estimate of how many illegitimate addresses there were by having the Internet service return to the
researchers at least some of the invalid or out-of-date email addresses (Dillman, 2000; 2007). As Dillman (1978; 2000; 2007) points out, sending pre-notifications can increase a surveyor’s response rate.

7. The survey was sent following Dillman’s “Tailored Design Method” (2000, 2007) recommendations for gaining a higher response rate. In the Web-based method an email was sent with a hyperlink to the survey Web site with instructions to click on the hyperlink to begin participation in the survey. Once the Web-based participants reached the Web-based survey their responses were entered on a secure socket server.

8. To attempt to minimize costs and interviewers’ time, efforts were directed toward completing all interviews, both face to face and telephone in six weeks. Efforts were also directed toward getting more interviews than necessary.

9. This step had four activities—one for each method of surveying.

A. For the first activity of the ninth step, the face to face interviewers initially attempted to conduct 1,470 face to face interviews—the first wave. There was no data on local norms regarding response rates for face to face interviews, but it was assumed that the method effect of a face to face, which would have tended to increase response rates relative to the other methods, would have been cancelled out by the students not being at their residences. Since the initial wave of face to face interviews yielded less than 138 completed interviews, a second wave of face to face
interviews was conducted two weeks later to obtain a minimum of 138 total face to face interviews from an additional 1,470 face to face interviews.

B. Telephone interviewers called the students at the contact number they gave Kent State University. According to the Office of Research, Planning and Institutional Effectiveness at Kent State University, their telephone surveys of freshman yielded a response rate of 17.8 percent using only one call (W. G. Schneider, November 9, 2004). The Department of Sociology at Kent State University typically obtains 12.5 usable telephone interviews per 100 calls made to college students with no follow-up calls (B.L. McDonald, personal communication, November 11, 2004). The University of Akron’s Center for Policy Studies typically obtained 30-35 usable telephone interviews per 100 calls in surveying the general population, but they used typically as many as seven call backs (S.M. Henry, personal communication, November 9, 2004). Therefore, 610 calls were initially made, corresponding to a 6.25 percent response rate. The researcher used the information gained from the first wave, especially response rates, to economically plan for the second wave. Since the response rate to the first wave was lower than expected, an additional 610 participants were called in the second wave.

C. In the first wave 1,215 mail surveys were sent. This assumed a response rate of 11 percent. The rationale for assuming this response rate was as
follows: According to Kent State’s Department of Sociology the response rate among college students is typically just above 10 percent for the mail method but they mailed no follow-up surveys (B.L. McDonald, personal communication, November 11, 2004). Similarly, the University of Akron’s Center for Policy Studies typically obtained a 23 percent response rate for the general population for the mail method (S.M. Henry, personal communication, November 9, 2004). Likewise, Lakeland Community College, located near Cleveland, Ohio, obtained response rates ranging from 11-25 percent using two mailings in their community surveys (E.M. Doherty, personal communication, November 19, 2004). Moreover, the University of Michigan, another Midwestern college, obtained a response rate of 40 percent of their undergraduates for the mail method (McCabe, Boyd, Couper, Crawford, & D’arcy, 2002). Indiana University, another Midwestern college, obtained a 21 percent response rate of their undergraduates (Dvorak, 1988). Gay and Airasian (2003) stated, “Research suggests that first mailings will result in a 30 to 50 percent return rate” (p. 289), but made no mention of inaccurate addresses. Assuming Kent State’s Department of Sociology is more accurate than the other organizations, the second mailing (waves) would have further increased the response rate. Since the response rate to the first wave was lower than expected, an additional 1,215 students were sent a mail survey in the second wave.
D. A first wave of emails to 742 students was sent electronically with a hyperlink to the Web-based survey. This assumed a response rate of 15 percent. This initial response rate was chosen due to the following facts: The Department of Sociology at Kent State typically obtains a 20 percent response rate to their Web-based surveys, which included incomplete surveys in their calculations (B.L. McDonald, personal communication November 11, 2004). The Bureau of Research Training and Services in Kent State’s Education College achieved a response rate of 33.6 percent with incompletes in the response rate calculations (P.L. Yannamsetty, personal communication, August 9, 2004). Kent State’s Office of Research, Planning, and Institutional Effectiveness obtained a response rate of 34.9 percent of freshman to their Web-based surveys (W.G. Schneider, personal communication, November 9, 2004). A University of Michigan Web-based survey obtained a response rate of 52 percent (McCabe, Boyd, Couper, & D’arcy, 2002). However, the target of 138 Web-based surveys was not received after the first wave. Because the response rate to the first wave was lower than expected, an additional 742 students were sent an email with a hyperlink to the survey in a second wave of emails, further increasing total Web-based responses.

10. For the mail survey method reminder emails were sent three days after their initial, postal mailing date to increase response rates as recommended by
Dillman (2000). For the Web-based survey, reminder emails with hyperlinks to the survey Web site were sent two days after the initial e-mailings.

11. The interviewer workloads were minimized in two ways:
   A. An adequate number of interviewers were recruited.
   B. Interviewer workloads were equalized as much as possible.

12. The Web site was designed so all questions had to be displayed on the monitor to be eligible for the drawing. However, participants could refuse to answer any of the questions and were still entered into the drawing because of Human Subject concerns.

13. All telephone interviewers were instructed to offer a chance to win $250 to students before the start of the interview. This served as an incentive to participate.

14. All data was stripped of any identifying numbers or letters before the researcher analyzed any data. However, the data was coded for face to face interviews and telephone interviews as to which interviewer conducted the survey to adjust for possible falsification, if needed.

Data Analyses

Unacceptable surveys and interviews in terms of item non-responses were deleted. Unacceptable surveys or interviews were determined by this study’s power needs and the necessity of keeping the integrity of the data. Listwise deletion was employed because Cohen, Cohen, West, and Aiken (2003), Witta (2000), Raymond (1986), Roth
(1994), Pigott (2001), and Pastor and Blas (2003) had indicated listwise deletion is not problematic when the questions missing are relatively small (e.g., 5 percent).

For research question one, after listwise deletion was executed due to answers marked “does not apply”, “not applicable”, “I prefer not to answer” (because of Human Subjects concerns), or because the question was skipped; mean imputation was used as an unbiased estimate for all the missing values to all questions with no usable responses. That is, the mean of the observed cases was used to replace missing values as recommended by Allison (2002), Schafer (1997), and the Census Bureau (1999) because although mean imputation will downwardly bias the variances and covariances through decreasing their associated standard errors, it will not bias the means (Allison, 2002; Census Bureau, 1999; Schafer, 1997). Allison, Schafer, and the Census Bureau also indicated, as a consequence of the downward bias to standard errors, tests of significance of regression would have been upwardly biased.

For research question one a total sample size of 552 (138 per four groups) was targeted because power would have been at least .79 for a targeted effect size of .20 based on the q-distribution (Cohen, 1988). For research questions two and three power would have been .80 for sample sizes of 164 at \( \alpha = .05 \) for a targeted effect size of .06. For research question four, power would have been .80, at \( \alpha = .05 \) for a sample size of 186.

For research questions two, three, and four, unbalanced ANOVAs were used. For research question two a 4 x 2 unbalanced ANOVA with the four levels being face to face, telephone, mail, and Web-based, and the two levels being male and female were used. For question three, a 4 x 2 unbalanced ANOVA with the four levels being face to face,
telephone, mail and Web-based, and the two levels being Caucasian and non Caucasian were used. For question four, a 4 x 2 x 2 unbalanced ANOVA with the four levels being face to face, telephone, mail, and Web-based, two of the levels being male and female, and the other two levels being Caucasian and non Caucasian was used. A further breakdown of ethnicities was not attempted to avoid zero cells.

Attempts to detect falsification by face to face or telephone interviews then occurred. First, an analysis of elapsed time for each interview occurred. If particular interviews had been completed in an unlikely time period—say 30 seconds— that particular interview was flagged as a possible falsified interview. Second, the “flagged” interview was carefully scrutinized to see if there was a pattern of answers possibly indicative of falsification; for example, if 95 percent or more of an interviewers answers had been “ones,” or answers had reflected a repeating pattern such as: “1, 2, 3, 4, 5; 1, 2, 3, 4, 5” and so on. Moreover, the detection of unlikely patterns of data was undertaken with all face to face and telephone interviews. If an interviewer had a “probable falsification,” the rest of that particular interviewer’s workload would have been carefully analyzed for more probable falsifications. If an interview had been found to be “probably falsified,” that interview’s data would not have been included in any subsequent analyses. Because of Human Subjects concerns, no action (especially punitive) would have been taken against any interviewers who probably falsified data. Since there was no evidence of any falsifications, no data was deleted for falsification reasons.

Some of the questions may not have applied to certain students which may have resulted in confused answer(s) (Fowler & Gallagher, 1997; Marquis, Marquis & Pollich,
1983). Therefore, at the end of each question a “not applicable” option was placed. Using this technique, questions not applicable to a sizable proportion (e.g., 20 percent) of the participants would have been deleted from the analyses.

The remaining data was analyzed to attempt to detect for Extreme Response Style (ERS). There would have been three possible manifestations of ERS—Bipolar, Unipolar, and simultaneously occurring Bipolar and Unipolar ERS. As an illustration, first, the mean of each individual participant’s IM item’s variances across all items, were calculated for each participant so that the variances of each participant could be rank ordered from least to most variable, irrespective of SDB. Thus, because 819 participants, or cases, were calculated there could have been 819 individual ranks. Second, a participant using ERS could have answered questions using “ones and fives” corresponding to ERS toward both poles. If this had occurred, the variances of each ERS participant(s) should have exceeded the variances of the non ERS participants. Therefore, the variances associated with this type of ERS were located (ranked) in the upper tail of the distribution of participants’ individual variances. That is, the variances associated with this type of ERS should have tended to be greater than the non ERS variances.

Third, if participants exclusively used one pole of the scale in answering questions, there should have been a) low variability and b) their answers should have been located at a single pole of the survey. To efficiently check for this second type of ERS two steps were taken:

1. Calculate the number of participants who were in the lower tail of the variances’ distribution.
2. Inspect the surveys and interviews falling at the lower tail of the variance distribution.

The plan was to delete the cases that were five percent or less of the total usable surveys, these cases would have been deleted from all subsequent analyses. After all, a non ERS participant could have answered mostly “threes,” which was the neutral position and subsequently would have displayed low variability. However, this participant’s mean variance across individuals would have been located at the lower tail of the ranked variances distribution. Since ERS was not found, an unbalanced ANOVA instead of an ANCOVA, using ERS as a covariate, with method of data collection as the independent variable and IM as the dependent variable, was conducted even though Ryan’s MCP test does not require a significant F-test. Nevertheless, the information provided about the variances would have been useful if the Games-Howell (GH) MCP was necessary later.

Participants may have slanted their responses to appear like an “average man.” (Aiken, 1997; Maynes, 1965). Quetelet had originated this kind of research centuries earlier (Aiken, 1997; Converse, 1970; Lazarsfeld, 1961). Therefore, the medians were compared across methods. Medians are less sensitive to extreme values than means (Glass & Hopkins, 1986; 1996; Huck, Cormier & Bounds, 1974; Mendenhall, Reinmuth, Beaver & Duhan, 1986). Therefore, the Kruskal-Wallis test was used to assess whether the medians were equal.

Coefficient alpha was obtained for the IM subscale of the BIDR-6 at .80. This gauged the internal consistency of this particular study.
For research question one, depending upon the results of the preceding analyses, the Ryan MCP Test was performed on the means of IM. If the variances were greatly unequal, the Games-Howell (GH) MCP was used instead of the Ryan, as recommended by Toothaker (1993, p. 62-63; 1991, p. 98-99, 105).

For only face to face interviews and telephone interviews, the latter halves of the surveys were compared with the first halves of the surveys to determine whether hesitant ingratiation toward the interviewers was likely to have occurred as demonstrated by Back and Gergen (1963), Sudman, Greely, and Pinto (1965), and Sutherland & Spilka (1964). Specifically, the means and medians of the second halves of the answers to the interviews were compared with the means and medians of the first halves of the interviews to determine if hesitant ingratiation was likely to have occurred. Therefore, a t-test was employed with halves of the test as the independent variable and IM as the dependent variable. The Wilcoxon-Mann-Whitney median test was used to compare medians as recommended by Huck, Cormier, and Bounds (1974), Pett (1997) and Vogt (1999).

The participants who selected yes to both of the questions, “Were there individuals within five feet of you?” and “Did these other individuals influence how you answered the survey questions?” for the Web-based and mail survey methods were compared. Specifically, the means and medians were compared with the participants who selected “no” to both questions and those who selected “yes” to both questions. If, for whatever reason, a participant answered no to the first question and yes to the second question, the participants were classified as if they answered yes to both questions because these participants were admitting to having been influenced. This “influence
suspected” group’s means and medians were compared to the other two groups using an unbalanced ANOVA and the Kruskal-Wallis median test.

Because Clark, Stanton and Rao (1996) found that African Americans use the extremes of a scale more than Caucasians, the variances of African Americans and Caucasians were compared. Therefore, an F-test of independent variances was performed with the African American Chi-Square distribution with its associated degrees of freedom in the numerator and the Caucasian Chi-Square distribution with its associated degrees of freedom in the denominator. Furthermore, an additional F-test compared the African American Chi-Square distribution with its associated degrees of freedom in the numerator and all other participants’ Chi-Square distribution with its associated degrees of freedom in the denominator.

Because Latinos have a concept called “simpatia,” which means concern for the group as a whole, they are theorized to answer surveys differently (Shultz & Chavez, 1994). If any identifiable Latinos participated in the surveys or interviews, their central tendency statistics like mean, median and mode were compared with Caucasians central tendency statistics. Since there are not many Latinos in the population (1.17%), a simple comparison of means, medians and methods was employed instead of any statistical tests. Moreover, the variances were compared using an F-test with the Latino Chi Square distribution with its associated degrees of freedom in the numerator and the Caucasian Chi-Square distribution in the denominator. Furthermore, an additional F-test compared the Latino Chi-Square distribution with its associated degrees of freedom in the
numerator and all other participants’ Chi-Square distribution with its associated degrees of freedom in the denominator.
CHAPTER IV
RESULTS

Sampling Data Characteristics

There were 828 undergraduates who participated in this study. A total of nine participants were excluded from the analyses. Because a Web-based participant answered “decline to answer” for every question, this participant was excluded from the analyzed sample. Another three participants were excluded because they answered either “not applicable” or “decline to answer” for more than 50% of the survey or interview questions. Five more participants were excluded because they were not members of the target population (undergraduates at Kent State’s main campus). The researcher decided to exclude these nine participants without knowing under what methods the nine participants had been assigned. Consequently, the total available sample size was 819 (828 – 9). The total sample sizes, per method, were 174 face to face interviews, 173 telephone interviews, 302 mail surveys, and 170 Web-based surveys for a total combined, available sample size of 819. The literature has shown the response rates for these four investigated methods are different (Adams, 1989; Bailey, 1994; Belson, 1986; Bourque & Fielder, 1995; Bradburn & Sudman, 1988; Cobanolu, Ward, & Moreo, 2001; de Leeuw, 1992; Dillman, 2007; 2000; 1978; Dillman, Gallegos, & Frey, 2001; Fowler, 1993; Golden, 1982; Goyder, 1987; Groves, 1989; Hacket, 1981; Hagan, 2003; Higgins, Dimnik, & Greenwood, 1987; Hox & de Leeuw, 1994; Kennedy & Vargus, 2001; McCabe, Boyd, Cooper, & Crawford, 2003; Sax, Gilmartin, and Bryant, 2003; Shermis & Lombard, 1999; Zelen, 1975). Consequently, this study, to simultaneously remain
scientific and economical, capitalized on the documented differential response rates.

Therefore, different sampling frames were used for each method, but the statistics were unbiased because every participant was first randomly selected and *afterwards* randomly assigned to one of the four data collection methods. Therefore, the response rates do not add up to 100 percent. The response rates, total usable responses, and sampling frames respectively were: face to face 5.92% (174/2,940), telephone 14.18% (173/1,220), mail 12.43% (302/2,430), and Web-based 11.46% (170/1,484). The average usable (excluding the nine unusable responses from the respondents not analyzed) rate, which includes all methods, was 10.14% \((819 \text{ [number of usable participants]} / 8,074 \text{ [total sample size]} = 2,940 \text{ [face to face sampling frame]} + 1,220 \text{ [phone sampling frame]} + 2,430 \text{ [mail sampling frame]} + 1,484 \text{ [Web-based sampling frame]})

**Sample Characteristics**

Of the 819 (100%) available participants in this sample, 247 (30.2%) reported they were male while 572 (69.8%) reported being female. Kent State University records confirmed this data. Regarding ethnicity, 682 (83.3%) were Caucasian, 75 (9.1%) were African American, 19 (2.3%) were Hispanic, 10 (1.2%) were international or foreign, 10 (1.2%) were Asian, 3 (0.4%) were Native American, 17 (2.1%) were other, and 3 (0.4%) did not indicate their ethnicity. Again, Kent State University records confirmed this data.

This study was limited by the practicality of using university-provided email addresses, postal addresses, and telephone numbers. According to the final Kent State main campus records, the address list was 97.63% complete and the attrition rate was 2.37% (L.A. Carlson, personal communication, February 6, 2008).
According to the Fall 2006 semester records, of the 819 analyzed participants surveyed, 299 (36.5%) were freshmen, 157 (19.2%) were sophomores, 158 (19.3%) were juniors and 205 (25%) were seniors. The frequency and percentages of the ages of the analyzed participants are shown in Figure three. 170 Web-based participants were included; 174 face to face interviews were included; 173 telephone interviews were included; and 302 mail surveys were included in the research sample. Therefore, with the total sample size of 819, using the Ryan Multiple Comparison test of four groups, power should have been .88 ($\alpha = .05$), assuming a targeted effect size, $\eta^2_p$, of at least .20. Since the BIDR measures IM on a ratio scale the Ryan was used. As Toothaker stated, “if you do all possible pairwise comparisons . . . and want good power . . . the Ryan test is the bottom-line choice of MCPs [Multiple Comparison Procedures].” (Toothaker, 1993, p. 57). Moreover, “The MCPs presented thus far [including the Ryan] do not need a significant overall F test before they are computed” (Toothaker, 1993, p. 41; Toothaker, 1991, pp. 47-55), for a one factor ANOVA with four levels. A priori MCPs are also generally more powerful than post-hoc MCPs (Toothaker, 1993, 1991; Glass & Hopkins, 1996, 1986). Furthermore, if the variances are unequal the Games-Howell (GH) MCP will be used instead of the Ryan as recommended by Toothaker (1991, pp. 98-99, 105).

Results of Research Questions

The research questions of this study were as follows: Question 1: Does the study show differences in IM (the deceiving others factor of SDB) by method of data collection (face to face, telephone, mail and Web-based)? Specifically, the mean of the face to face interviews should be greater than (>) the mean of the telephone interviews which, in turn,
is greater than (>) the mean of the mail surveys which, in turn, should be greater than (>)
the Web-based surveys. Question 2: As suggested by the literature, does this particular
order (for question 1) of means (face to face interviews > telephone interviews > mail
surveys > Web-based surveys) of the methods of data collection in terms of IM apply to
both genders? Question 3: Does this particular order of means (found in question 1) apply
to different ethnicities? Question 4: Does an interaction exist among method of data
collection, gender, and ethnicity in terms of SDB? That is, did the particular order of
means depend upon both the participant’s gender and ethnicity.

Like the results of hypothesis one previously mentioned, hypothesis three did not
turn out as expected (face to face interviews > telephone surveys > mail surveys > Web-
based surveys). As with the analyses of all the research questions, these analyses used
Type III Sums of Squares. The results among the face to face, telephone, mail, and Web-
based methods of the Ryan Multiple Comparison Procedure (MCP) test were not
significant (q = .97, p = .25, α = .05, \( \eta_p^2 = .010 \)). The effect size measure used was partial
eta squared, \( \eta_p^2 \), which can be calculated using Equation 1 below.

\[
\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \quad \text{(Equation 1)}
\]

The means were face to face 61.00, telephone 61.41, mail 62.00, and Web-based
59.85. The overall F-test of the one factor unbalanced ANOVA was significant (\( F_{3,816} =
2.73, p = .043, \eta_p^2 = .010 \)).

No significant interaction between method of data collection and gender in terms
of IM was found (\( F_{3,811} = .52, N = 819, p = .67, \alpha = .05, \eta_p^2 = .002 \)). However, the actual
effect size, denoted by \( \eta_p^2 \) (Vogt, 1999, 2005; Grimm & Yarnold, 2000; 2001; Becker,
1999; Kirk, 1982; Tabachnick & Fidell, 1989), was extremely low. The observed power was only .16 (\(\alpha = .05\)) because the effect size was low and sample size was relatively large.

Levene’s test of equality of error variances for method of data collection by ethnicity interaction effect was not significant (\(F_{3, 811} = 1.69, p = .11\)). Because the groups were so disproportionate in size (Figure 2 in the Appendix), both in the population and sample, ethnicity was therefore coded as Caucasian or non-Caucasian. A method (face to face, telephone, mail, and Web-based) by dichotomized ethnicity (Caucasian or non-Caucasian) two factor unbalanced ANOVA was conducted. No significant interaction between method of data collection and ethnicity was found (\(F_{3, 811} = 2.53, p = .06, \alpha = .05, \eta^2_p = .009\)). The actual effect size was too low, .009, and the observed power was .63 (\(\alpha = .05\)).

As in the population and the sample, the ethnic groups were so disproportionate in size (Figure 2), a method by gender by ethnicity (three factor unbalanced ANOVA) interaction could not be judiciously tested. Any ANOVA without empty cells testing an interaction must be crossed by all levels of all variables (Glass & Hopkins, 1996, 1986; Keren & Gideon, 1993; Kirk, 1982; Searle, 1993). Because the specific ethnicities were Caucasian, African American, Hispanic, Asian, international, American Indian, and other; the ethnicities had to be dichotomized into two groups, Caucasian or non-Caucasian, to avoid empty cells, increase power, and avoid a Type II error (Cohen, 1988; Glass & Hopkins, 1996; Halpin, et al., 1991; Keren & Gideon, 1993; Jennings & Green, 1984; Kirk, 1982; Searle, 1993; Spinner & Gabriel, 1981; Wilcox, 1993). Therefore,
ethnicity was coded as Caucasian or non-Caucasian. Levene’s test of equality of error variances for method of data collection by ethnicity interaction effect was not significant ($F_{15,803} = 1.05, p = .40$). No interaction between method of data collection, gender, and dichotomized ethnicity was found using a three factor (4 x 2 x 2 unbalanced ANOVA) with 819 total participants ($F_{15,803} = 1.26, N = 819, \alpha = .05, p = .29, \eta^2_p = .005$). The observed power was only .34. As Glass and Hopkins (1996) stated, “The absence of interaction [italics copied] is the statistical justification for generalizability [italics copied] . . . if there is no interaction between a treatment factor and a relevant characteristic of the subjects (e.g., gender), the overall effect can be generalized to both sexes without qualification.” (p. 485). Since 1) the Ryan test of methods was not significant; 2) the two way interactions (method by gender, method by dichotomized ethnicity) were not significant; 3) and the three way interaction (method by gender by dichotomized ethnicity) was not significant; 4) and partial eta squared ($\eta^2_p$), an effect size measure, was always small, this study’s claim that method did not influence IM could be made more convincingly.

Possible Threats to Validity—Results of Controlling Analyses

Falsification.

Forsman and Schreiner (2004) defined falsification as, “the interviewer may fabricate answers to…the questions in the questionnaire.” Given the results of the analyses detailed below, it is highly unlikely falsification occurred.

The elapsed times of the face to face and telephone interviews (from start to finish, in minutes) of all interviews indicated 49.4% ($N = 86$) of the face to face
interviews were one minute long. Seven interviewers had one-minute interviews, ranging from 3 to 12 one-minute long interviews per individual interviewer. The range of all the face to face interviews was 1 minute to 16 minutes, as shown in Table 3. There were no unusual patterns in the face to face data. In fact, of the entire dataset (including all four methods), there were no unusual patterns in the data detected except one participant in the Web-based method answered 7 (“decline to answer”) for all questions. This participant’s data was deleted before any analyses were executed. Given the information above, falsification did not likely occur. More importantly, the number of face to face interviewers (29 interviewers) was adequate given the obtained face to face sub-sample size of 174. Subsequently, there was little pressure to falsify. The maximum number of participants interviewed by a single face to face interviewer was maintained at 24 interviews, and the minimum was one, with a mean of six interviews for each face to face interviewer. Moreover, in a face to face interview, a participant could be asked (and answer) 30 to 33 questions, depending on skip patterns, in 60 to 89 seconds.

Falsification was unlikely to have occurred in the telephone method for two reasons. First, the maximum number of participants interviewed by a single telephone interviewer was kept low (20), and the minimum was one, with a mean of nine interviews for each telephone interviewer. Second, as shown in Table 4, the data suggests falsification did not occur. The elapsed times of the interviews appeared to be not too short, given the hasty and rapid nature of falsification (Bushery, Reichert, Albright, & Rossiter, 1999; Cannell, Miller, & Oskenberg, 1981; Forsman and Schreiner, 1991). An analysis of elapsed times indicated 1.72% (N=3) of the telephone interviews were five
minutes long. The range of completion of times of all the telephone interviews was 5 to 26 minutes. As a whole, the 20 different telephone interviewers had a mean of 9.90 minutes, a median of 9 minutes, a standard deviation of 3.40 minutes, a standard error of the mean of .26 minutes, a skewness of 1.54, and a kurtosis of 3.29.

Given the facts in the two previous paragraphs, it is highly unlikely falsification occurred. The logic is if interviewers falsify, they are likely to do so because of time pressures. Analogously, a thief would not remain at the crime scene for 48 or more hours. Since the average telephone interviewer’s work load was only 8.7 interviews, there was little pressure to falsify. Moreover, only one Web-based survey had a peculiar pattern of data as previously discussed — the participant answered “decline to answer” to all questions. It is impossible for falsification to occur in a self-administered method survey because falsification can only be committed by an interviewer—not a respondent. Therefore, a Web-based response, which has no interviewer, cannot be falsified. Since this pattern of data was detected and subsequently deleted before any statistical analyses were conducted, it would not have affected any of the method means.

**Summary of Falsification Results.**

As reported above (a) falsification was unlikely to have occurred, (b) there was no compelling reason for interviewers to falsify because, as indicated in the previous paragraph, the interviewer workloads were kept low, and (c) the only unusual pattern of data was one student’s answers in the Web-based method. Again, since no interviewer is present in a Web-based survey, falsification could not have occurred.
Participants choosing “not applicable” as an option had little effect on the research results. The “not applicable” alternative was chosen most for question 20 at 4.5%, question 2 at 2.9%, question 5 at 1.3%, question 4 at 1.1%, and question 19 at 1.1%, with the rest of the questions at less than 1% (Refer to Figure 4). The question numbers, text and non-applicable percentages are reported below, respectively:

20. I have received too much change from a salesperson without telling him or her, 4.5%;

2. I have never damaged a library book or store merchandise without reporting it, 2.9%;

4. I sometimes drive faster than the speed limit, 1.3%

5. I never cover up my mistakes, 1.1%;

19. When I was young, I sometimes stole things, 1.1%.

The rest of the questions had less than a 1.0% not applicable rate. This was considerably less than the 20% pre-determined limit. Therefore, no participant’s data was deleted solely because of participants answering “not applicable” to questions. In fact, the overall item non-response (“decline to answer”, “not applicable”, and questions left blank) rate for the entire dataset (the number of empty data cells / total number of data cells irrespective if they were empty or not) was .008. All surveys and interview questions are listed in Figure 4.

To assess if participants were overly focused on the responses options (the leaves) instead of focusing on the entire questions (the stems and the leaves), three rigged distracter questions were asked to lure hasty or impetuous responses by participants.
Each rigged question had only one correct answer. None of the distracter options were chosen more than 15.6% of the time. Additionally, the distracter option receiving the 15.6 percent response rate, was in a question about math. A good percentage of undergraduates in the United States have poor math skills (Klein et al., 2005). Furthermore, Kent State’s research department of Research, Planning, and Effectiveness records showed 66.1% of first-time freshman at Kent State were taking developmental math during Fall 2006 (L.A. Carlson, personal communication, October 7, 2008). Of the two non-mathematical rigged questions, the response with the highest frequency reached only 2.9 percent. One would expect a response rate of at least 20 percent by chance alone (100 percent / 5 options = 20 percent) if people were responding carelessly.

**Extreme Response Style.**

Another threat to the validity of this study was Extreme Response Style (ERS) bias. It is the tendency of participants to answer predominantly at the extremes of a scale rather than simply trying to comprehend the questions and determine which option best reflects their attitudes or feelings to the questions. To address the possible confounding influences of ERS bias, two strategies were used. First, instead of a 7-point scale, a 5-point Likert scale was used to circumvent, or at least minimize, ERS bias. As mentioned in Chapter III, according to the literature on scaling, the resulting reliability may have decreased or even remained the same but is highly unlikely to have increased the reliability of a scale (Clarke, 2000, 2001; Lubke & Muthen, 2004; Bernstein & Teng, 1989; Cox, 1980; Olsson, 1979; Green & Rao, 1970; Benson, 1971; Ramsay, 1973; Symonds, 1924; Peter, 1979; Taber, 1977; Komorita & Graham, 1965; Matell, 1971;
Carroll, 1945; Garner, 1960; Lissitz & Green, 1975; Komorita, 1963; Martin, 1978; Matell & Jacoby, 1971; Guilford, 1954; Crandall, 1975; Green & Goldman, 1993; Daamen, 1992; Stober, Dette, & Musch, 2002; Tversky & Kahneman, 1974; Hayes, 1998; Askenasy, 1985). Paulhus (1998) as well as Kwan and Vitelli (1998), indicated the BIDR can be scored dichotomously if a 7-point scale is used but not continuously. They also indicated the BIDR can be scored continuously if a 5-point scale is used. Fortunately, for this study the Cronbach Alpha coefficient was .80. Thus, the first strategy was to minimize this bias by using a 5-point scale. The second strategy was assessing if there could still have been possible confounding influences of ERS bias. Two types of ERS bias were investigated. The first type, Bipolar ERS, is the tendency to answer at both extremes, either a 1 (“Definitely Untrue”) or 5 (“Definitely True”). The other options, the Non-Bipolar ERS options, were 2 (“Occasionally True”), 3 (“Moderately True”), and 4 (“Mostly True”). The other possible ERS bias type is Unipolar ERS. Unipolar ERS is the tendency to answer using a single pole. That is, instead of predominantly using either a 1 or 5, the participant overly uses solely 1’s or, conversely, the participant overly uses solely 5’s in answering questions. Extremely high or low variability is necessary, but not a sufficient manifestation of both Bipolar and Unipolar ERS bias, respectively. Thus, the average variance per student across all the questions of interest must be very high (for bipolar ERS) or very low (for Unipolar ERS). Omitting the 21 students with the lowest mean ERS variances and the 21 highest ERS variances (corresponding to below the 2½ percentile and above the 97½ percentile across students in terms of ERS variability) did not change the essential results of the main analyses. The resulting probability value was
This omission of both potential types of ERS responses is illustrated in Table 5.

The simultaneous manifestation of ERS was the most conservative and realistic since it was much more likely that both types of ERS bias could be affecting the data. Conversely, it is less tenable and conservative to presume only one type of ERS (Bipolar or Unipolar) bias existed and the other did not exist in the data.

To assess the threat of low Unipolar Extreme Response Style (ERS) bias adversely affecting research Question 1, the Ryan MCP test was conducted on the data, deleting the participants who exhibited Unipolar ERS. This analysis can test whether this particular type of an ERS could have adversely affected the results for research Question 1. That is, this same research test’s results will be compared to the new ERS test’s results, which can effectively eliminate the operation of an ERS effect. ERS can adversely affect the scores of a study, but in this case it did not, as shown in Table 5.

The result of the Ryan MCP test was not significant (q = .95, p = .25, \( \alpha = .05 \)). The means were face to face 61.08, telephone 61.47, mail 62.11, and Web-based 59.88. Levene’s test of equality of error variances was not significant (\( F_{3, 761} = 1.16, p = .30 \)).

If researchers compare the overall initial results to the results of analysis attempting to nullify a Unipolar ERS bias (the tendency to answer using a single pole), they would principally be the same as shown in Panels C & D of Table 5. Again, Unipolar ERS bias is when the participant overly uses solely 1’s or, conversely, the participant overly uses solely 5’s in answering questions. Extremely low variability is necessary, but not a sufficient condition for a manifestation of low ERS bias. Thus, the
average variance per student across all the questions should be very low for Unipolar ERS. Participants at the 5th percentile or lower in terms of low variance, were deleted from the analysis. Thus, forty-two students were omitted (42 is \(\approx 5\%\) of 819), with the lowest mean variances at or below the 5th percentile. As shown in Table 5, omitting low ERS response did not change the eventual results of the main analyses nor their significance status (significant or not). For example, the results of the Ryan MCP test were not significant (\(p = .25, \alpha = .05\)). The means were face to face 61.08 (S.D. = 10.74), telephone 61.47 (S.D. = 10.32), mail 62.11, (S.D. = 11.59), and Web-based 59.88 (S.D. = 11.16). Levene’s test of equality of error variances was not significant (\(F_{15, 761} = 1.16, p = .30\)), and the overall F-test of the one-factor unbalanced ANOVA was not significant (\(F_{3, 761} = 2.58, p = .053, \eta_p^2 = .010\)). When the 42 lowest values were omitted from the analyses, the probability value for the overall F value was not significant (\(F_{3, 761} = 2.58, p = .053, \eta_p^2 = .010\)). The results of the Ryan MCP were also not significant (\(p = .25\)). As shown in Table 5, omitting high ERS response did not change the eventual results of the main analyses or their significance status (significant or not). The results of the Ryan MCP test were not significant (\(p = .61, \alpha = .05\)). The means were face to face 60.91 (S.D. = 9.15), telephone 61.15 (S.D. = 9.43), mail 61.57 (S.D. = 9.88), and Web-based 60.27 (S.D. = 8.97). Levene’s test of equality of error variances was not significant (\(F_{15, 761} = 1.50, p = .10\)), and the overall F-test of the one-factor unbalanced ANOVA was not significant (\(F_{3, 761} = 1.09, p = .35 \eta_p^2 = .004\)). When the 42 participants with the highest variances were omitted from the analyses, the probability value for the overall F value
was not significant \((F_{3,761} = 1.09, p = .35, \eta^2_p = .004)\). Moreover, the results of the Ryan MCP were not significant \((p = .61)\).

**Summary of ERS Results.**

First, because a 5-point scale was used instead of a 7-point scale, ERS bias could not manifest itself. Second, in deleting sources of ERS bias, there were no principal differences in “non” significance. Overall, it is plausible ERS bias had little, if any, effect on the results of this study.

Another specific threat to the validity of this study—the setting in which a mail or Web-based survey is completed can affect participants’ answers to surveys—concerns the setting of the study as Beebe, Harrison, McRae, Anderson, and Fulkerson (1998), Dunnette and Heneman (1956), and Koson (1970) pointed out. They stated the social context of the setting in which a mail or Web-based survey is completed can affect participants’ answers to surveys. Therefore, Beebe, Harrison, McRae, Anderson, and Fulkerson (1998) recommended participants fill out Web-based surveys at least 5 feet from any bystander beside them. A mail survey can be difficult for another person beside the participant to read, especially if they are more than 5 feet away from the person taking the survey. Additionally, as suggested by Beebe, Harrison, McRae, Anderson, and Fulkerson (1998), the participant could block the view of the non-participant with a part of their body or something else. This particular study asked each Web-based and mail participant if (a) there were people closer than five feet from them, and (b) if these people influenced how they answered these survey questions. Two hundred fifty-nine combined Web-based and mail participants (132 Web-based [77.65%] and 127 mail [74.71%])
answered no to the question of whether there were any other people within five feet of them while they completed the survey. However, 70 Web-based (19.41%) and 66 mail (21.76%) participants answered yes, indicating there was another individual within 5 feet of them while they completed the survey. Fifty-four (75.71%) of the participants who answered yes and indicated there were people within 5 feet of them also indicated that bystanders did not influence how they answered the survey. Fourteen participants (20.00%) of those who answered yes and indicated they were within 5 feet of another bystander, also indicated their answers were influenced by these other individuals.

To conduct a simple t-test, all (75) participants under all four methods who admitted they were influenced in any way by other people, were compared to 75 of 334 randomly selected participants who indicated they were not influenced by other people. Only six people were excluded from these analyses because they did not answer either or both questions. No significant difference was found between participants who admitted being influenced and those who did not ($t_{2,148} = .77$, $p = .44$, $\alpha = .05$, $\eta_p^2 = .126$). Using the same logic and data as the previous t-test, the Mann-Whitney nonparametric median test was not significant either (Mann-Whitney $U = 9588.5$, $Z = -.17$, $p = .87$, $\alpha = .05$).

To test for the “average man” influences, as posited by (Aiken, 1997; Maynes, 1965), a Kruskal-Wallis test was done with method being the independent variable and IM being the dependent variable. No significant differences were found ($\chi^2_{K-W} = 3.36$, $p = .34$) using a sample size of 819. When the sample size of the methods were equalized to 170 each (680 total), the Kruskal-Wallis test was still insignificant ($\chi^2_{K-W} = 1.73$, $p = .85$).
One of the questions of the IM subscale, “I always declare everything at customs.” was excluded because it was considered invalid to some participants. Moreover, because there was one question omitted from the 20-item IM subscale for validity reasons, the formula for Cronbach’s alpha shows the internal consistency would have increased from .80 to .81 if the question had been included. Furthermore, internal consistency is considered a lower bound estimate of the actual reliability (Crocker & Algina, 1986, p. 119; Allen & Yen, 1979, p. 80; Gorsuch, 1983, p. 154).

Clark, Stanton, and Rao (1996) found African Americans generally tend to use the extremes of a scale more than Caucasians. Therefore, the variances of the two ethnic groups, African Americans and Caucasians were compared in an F-test of independent variances. Using all 75 African American participants and 75 randomly selected Caucasians participants, the African American Chi-Square distribution was placed in the numerator and the Caucasians Chi-Square distribution was placed in the denominator. The variances of these two groups were not significantly different ($F_{2, 150} = 3.85, \alpha = .05, p = .052, \eta^2_p = .025$).

An additional F-test compared the African American Chi-Square distribution in the numerator and a random selection of 75 of all the other 669 participants’ Chi-Square distribution in the denominator. There was not a significant difference ($F_{2, 150} = 1.04, \alpha = .05, p = .31, \eta^2_p = .167$).

Because Latinos have a concept called “simpacia,” which means concern for the group as a whole, they may have been apt to answer surveys differently (Shultz & Chavez, 1994). The variances of Latinos and Caucasians were compared with an F-test.
The Latino Chi Square distribution was placed in the numerator and the Caucasian Chi-Square distribution in the denominator. Latino variances were not significantly different from Caucasian variances \(F_{2, 36} = .02, \alpha = .05, p = .88, \eta_p^2 = .047\).

An additional F-test compared the Latino Chi-Square distribution in the numerator and a random selection of all other participants’ Chi-Square distribution in the denominator. There was not a significant difference \(F_{2, 36} = .20, \alpha = .05, p = .66\).

Therefore, the Games-Howell MCP was used because it adjusts for different sample sizes and variances (Klockars & Sax, 1986, p. 80-81; Toothaker, 1991, p. 98, 105). The Games-Howell Latino mean differences with other ethnicities were .94 \((p = .90)\) with Caucasians, 3.49 \((p = .84)\) with African Americans, -3.20 \((p = .96)\) with international or foreign; -2.40 \((p = .85)\) with Asians; 19.42 \((p = .09)\) with Native Americans; and .48 \((p = .93)\) with other. The Kruskal-Wallis median test was not significant \(\chi^2_{K-W} = 7.62, t = .18\).

Supplemental Analyses and Results

Because of the relatively large number of non-significant results already discovered in this study, one more test was added to see if the basic pattern of non-significant results still held. It was hypothesized there would not be a significant difference between verbal (face to face and telephone) and non-verbal (mail and Web-based) methods because there was no significant difference between the four methods. An independent samples t-test confirmed no significant difference between verbal and non-verbal methods \(t_{2, 678} = .65, p = .52, \eta_p^2 = .001\).
A Summary of the Results

Answers to Research Questions

1. There was no evidence of differences between methods in terms of SDB using the Ryan and Games-Howell MCPs, which both use Bonferroni-type tests. Unbalanced ANOVA F-tests also corroborated the MCP results. The effect size, $\eta_p^2$, was .01. The method statistics and the item statistics are located in Table 7.

2. There was no evidence of a method by gender interaction effect. On average the IM scores across the four methods of data collection did not depend on participants’ gender. The effect size, $\eta_p^2$, was .002.

3. There was no evidence of method by ethnicity interaction effect. On average the IM scores across the four methods of data collection did not depend on participants’ ethnic background. The effect size, $\eta_p^2$, was .009.

4. There was no evidence of method by gender by ethnicity interaction effect. On average participants’ IM did not depend on method of data collection, gender, and dichotomized ethnicity. The effect size, $\eta_p^2$, was .005.

Possible Confounding Influences as Indicated by the Literature:

1. Falsification did not likely occur.

2. Item non-response was not problematic.
3. Students did not likely focus more on the options (leaves) than the questions as a whole.

4. Extreme Response Style did not overly influence this study’s results.

5. The social context of the setting of the study was not influential.

6. The “average man” effect was not consequential.

7. African Americans did not answer the survey any more extremely than Caucasians.

8. African Americans did not answer the survey any more extremely than all other ethnicities combined.

9. Latinos did not answer the survey any more extremely than Caucasians.

10. Latinos did not answer the survey any more extremely than all other ethnicities combined.

11. Hesitant ingratiation did not occur. In fact, the opposite effect occurred. No matter what method was used, the first half of the survey was always greater than the second half.

12. There was no difference in SDB, comparing verbal (face to face and telephone interviews) and non-verbal (mail and Web-based) methods.

13. This study had respectable reliability with a Cronbach alpha of .80.
Before discussing the results of this study, a brief definition and illustration of effect sizes seems appropriate for novice researchers and laypeople. Partial eta squared, $\eta_p^2$, is a measure of effect size whereby the effect of interest is in the numerator while the total variance, including error, is in the denominator (Vogt, 1999, 2005; Grimm & Yarnold, 2001, 1999; Kirk, 1982; Tabachnick & Fidell, 1989).

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$$

(Equation 1)

Stevens (1990) and Cohen (1988) defined small effect size for $\eta_p^2$ as around .1, medium as around .25, and large as being greater than .4. An effect size close to 0.00 usually signifies no treatment effect.

Research Hypotheses

The first research hypothesis stated there would be differences between methods in terms of Social Desirability Bias (SDB) with Impression Management (IM) means in this descending order from greatest to least: face to face > telephone > mail > Web-based. This hypothesis was not supported. The effect size was a miniscule .005, partial eta squared, $\eta_p^2$. Furthermore, .005 is smaller than Stevens’ (1990) low effect size benchmark of .1.

The second research hypothesis stated there will be no Method by Gender interaction. An interaction effect occurs if the combined effect, such as data collection Method (Factor M), Gender (Factor G), is significantly more than the sum of its additive
effects and random error as shown in Equation 2. Equation 2 was simplified for
instructional purposes.

\[
\text{Interaction Method & Gender} = \text{Factor Method} - \text{Factor Gender} - \text{random error} \quad (\text{Equation 2})
\]

If there was an interaction effect, the left hand side of Equation 2 would have
been significantly larger than the additive effects of the two main factors Method and
Gender along with random error. In this study, as expected there was no interaction effect
between Method and Gender. In fact, the effect size for the interaction of Method and
Gender, \(\eta_p^2\), was .011.

The third hypothesis maintained there would be a significant interaction between
method and ethnicity. This was not supported. The effect size, \(\eta_p^2\), was .029.

The fourth hypothesis held there would not be a significant Method by Gender by
Ethnicity (three factor) interaction effect on IM. This hypothesis was supported. The
effect size, \(\eta_p^2\), was .016.

Not one of the four research questions resulted in a single rejection of a null
hypothesis. As discussed in Chapter III, the sample sizes were adequate. Had larger but
still modest effect sizes of .2 or even .14 occurred, this study’s results may have been
simpler to interpret. However, all effect sizes, \(\eta_p^2\), for all research questions were
considerably below Stevens’ (1990) and Cohen’s (1988) criterion for even a small effect
size of .1. Even the largest effect size of .029 was less than 30% of the .1 benchmark!
Controlling Hypotheses

A unique strength of this study was it addressed all known counter hypotheses. There were 14 counter—meaning controlling hypotheses—addressed by this study. Two of the counter hypotheses were avoided (not allowed to influence the current study’s results), but 12 had to be tested for. Counter hypotheses 1 through 12 were all not significant at the .05 probability level, and thus were not supported (pages 105-106). There were 12 counter hypotheses, and their findings are listed below. Therefore, 14 known counter hypotheses (12 + 2) were either avoided or statistically tested for. The 12 counter hypotheses are restated along with their results below.

1. Hypothesis: Falsification will not occur. Result: As expected, there was no evidence of falsification.

2. Hypothesis: Item non-response will not be problematic. Result: As expected, item non-response was miniscule.

3. Hypothesis: Participants will not solely focus on the options (the leaves) of the questions. Result: As expected, there was no evidence participants focused only on the options.

4. Hypothesis: Extreme Response Style (ERS) will not influence this study. Result: As expected, ERS did not affect the results of this study.

5. Hypothesis: The social context of the setting of the study will not be influential. Result: As expected, the evidence suggests the social context did not influence the results of this study.
6. Hypothesis: The “average man” effect will not be consequential. Result: As expected, there was no evidence of an “average man” effect.

7. Hypothesis: African Americans will answer the survey more extremely than Caucasians. Result: Contrary to the literature-based hypothesis, African Americans did not respond any more extremely than Caucasians.

8. Hypothesis: African Americans will answer the survey more extremely than the other ethnicities combined. Result: Contrary to the literature-based hypothesis, African Americans did not answer any more extremely than other ethnic groups combined.

9. Hypothesis: Latinos will not answer the survey any differently (extremely) than Caucasians. Result: Contrary to the literature-based hypothesis, Latinos did not answer differently (extremely) than Caucasians.

10. Hypothesis: Latinos will not answer the survey any differently (extremely) than other ethnicities combined. Result: Contrary to the literature-based hypothesis, Latinos did not answer differently (extremely) than other ethnic groups combined.

11. Hypothesis: Hesitant ingratiation will occur. Result: The opposite effect occurred. No matter what method was used, the first half of the questions was *always* greater than the second half.

12. Hypothesis: The study will have a respectable Cronbach’s alpha Result: Cronbach’s alpha was a respectable .80.
Integration of this Study Results with the Literature on Method Effects on SDB

Convergent Findings

Like Miles and King (1998), Dahl (1992), and Edwards (1957), this study found no interaction effect between method and gender. There was no study in the SDB literature to suspect a method by gender interaction effect on SDB, and in the current investigation no interaction was found. These findings suggest researchers may ignore gender differences when investigating method effects on SDB.

There was no study in the SDB literature supporting a method by gender by ethnicity (three factor) interaction on SDB, and in the current study no interaction was found. These facts suggest researchers may ignore any combined effect of method by gender by ethnicity when investigating effects on SDB. Askenasy (1978) mentioned three main effects of SDB—method, gender and ethnicity in his conference paper abstract—but no interaction effects were mentioned in this abstract. However, only the abstract was obtainable. Therefore, because the current study was the only study that explicitly investigated a three factor interaction between method, gender, and ethnicity, no firm conclusions can be made. Nevertheless, based on both the non significant results of this current study, the small effect size found was, $\eta_p^2 \approx .016$, and on the dearth of literature on this specific three factor interaction, researchers should disregard a method by gender by ethnicity (three factor) interaction effect on SDB.
Divergent Findings

The first research hypothesis posited the IM means will result in the following order (from greatest to least): face to face, telephone, mail, and Web-based. According to the SDB literature cited in this study the higher means are more biased than the lower means. Despite the many previous studies supporting method-related differences proposed in this study’s hypothesis, this study found no statistically significant method linked differences in SDB for the first research question.

The third research hypothesis of a method by ethnicity interaction was not rejected. Unlike the Aquillino (1992, 1994), Rogers (1976), and Zhang and Gerstein (1999) studies this study found no support for two factor interaction of method of data collection and ethnicity. Fowler (1997) and Greenfield et al. (2000) found no interaction. The Aquillino (1994), Rogers (1976), and Zhang and Gerstein (1999) studies used random assignment while the other studies did not. All three studies that used random assignment support an interaction effect. Moreover, with the exception of the Rogers (1976) study, the context of the study was the use or abuse of drugs or alcohol. Furthermore, with the exception of the Aquillino (1994) study, the methods compared were face to face and telephone.

Implications of Findings

First, the fact that this study failed to find SDB differences in IM among the four methods (face to face, telephone, mail, and Web-based) may be highly useful to survey researchers. As the current study suggests, data collection method may have no effect on SDB. Thus, SDB may not be as crucial as it once was or appeared to be in deciding
which of the four data collection methods is used. This suggests that, all other things
being equal, researchers can use the most practical or inexpensive method without regard
to one of the most often mentioned threats to the accuracy of surveys: SDB.

Second, it is possible a specific question order of an SDB inventory could have
affected many studies on SDB. The order the questions appear has been shown to affect
respondents’ answers (DeMoranvi, Bienstock, & Luthans, 1983; Ramirez & Straus, 1998;
McAllister & Wattenberg, 1991; Benton & Daly, 1960; Moore, 1996, 2002; Howe &
Baldwin, 1996; McFarland, Schurr, & Henriksen, 1992; Lacy & Dean, 1969; Jones &
Rorer, 1992; Tourangeau, Couper, & Conrad, 2004; Nisbett & Wilson, 1992; Sutherland
& Spilka, 1992; Knowles & Coker, 1983; Trope & Liberman, 1993; Strack & Schwarz,

Only two studies were found (Knowles, Coker, Scott, Cook, & Neville, 1996;
Schurr & Henrikson, 1983) about survey research methods, other than this study, that
documented a randomization of their questions before administration of the interview or
survey. More importantly, no study on SDB, other than this one, ever randomly selected
questions. Since the questions were ordered identically across methods after initial
randomization and all the other studies in the SDB literature did not randomize their
questions, not randomizing questions may have biased many of the other studies’ results
on SDB. Because no other study controlled for the question order effect, there could be a
question order bias tainting the SDB literature. However, since this study randomized its
questions, a question order effect is not plausible. Therefore, researchers investigating
SDB should prevent, test for, or nullify *question order* effects from influencing their results.

Third, as mentioned on page 106, hesitant ingratiation was the one controlling hypothesis that was, in fact, rejected. Regardless of what method was used, the second half of the questions was always significantly lower than the first half, suggesting the second half questions appeared to have been answered with less social desirability. However, this was not found in any other study about different methods affecting SDB. Thus, the question order effect may have affected *other* studies’ results but *not* this one because it was the only study to randomize its questions.

Fourth, people who respond to interviews or surveys are not, as Belson (1986) stressed, “recording machines to be turned on by an interviewer’s doorstep questioning” (p. 19). Humans process information they receive, and not necessarily what the survey researcher predicts, desires, or finds acceptable (Knowles & Byers, 2002; Bocker & Keil, 1983; Meston & Heiman, 1983; Rorer, 2008; Nisbett & Ross, 2008; Crowne & Douglas, 2005; Eriksen & Kuethe, 1981; Krosnick & Li, 1983; Tourangeau & Couper, 2001; Norman & Bobrow, 2006; Gilbert & Krull, 1995; Krull & Dill, 2003; Zuckerman & Knee, 1994; Gilbet, 2005; Doby, 1995; Schwarz, 1989; Staats, 1989; Fisher, 1969; Gilbert & Krull, 2008; Lehman & Krosnick, 2005; Rebernik & Sirec, 1996; Baumard, 1996; Kruglanski, 1993; Inkso & Butzine, 1973; Martin, Seta, & Crelia, 1990; Bargh & Pietromonaco, 1962; Krosnick & Abelson, 1987; Orne, 2006; Martin, 1998). The fact that the second halves of this instrument were always lower than the first halves, suggests it is *possible* the human participants learned to extinguish their SDB. Although they could
not have known the specific purpose of this study, because they were never informed, it is possible they may have assumed about half-way through this survey instrument it was permissible or even encouraged not to be as socially desirable.

Fifth, another possibility is that surveys and interviews have became so ubiquitous in America, SDB may actually have decreased, as shown in Meta Analysis by Dindia and Allen (1992) and Richman, Kiesler, Weisband, and Drasgow (1998). Both Meta Analyses indicated SDB has decreased over time. Moreover, American culture is not as monolithic and refined as it was before the 1960s. It is now more multicultural and coarse.

Sixth, as previously mentioned, SDB may be on the decline because of the sheer ubiquity of surveys and interviews. Certainly, this study supports the findings of the Meta Analyses by Dindia and Allen (1992) and Richman, Kiesler, Weisband, and Drasgow (1998) that SDB may be decreasing over time.

Delimitations

Since this evaluation of differing methods of data collection was mainly concerned with internal validity, generalizations cannot be made outside of students attending Kent State University’s main campus. However, since this study defined and used a random selection of a specific target population, utilized random assignment to groups, and applied Multiple Comparison Procedures (which are appropriate given the nature of the study), it could be considered more rigorous than any of the studies mentioned. Furthermore, this study analyzed subgroup differences such as gender and ethnicity—adding to some extent to the external validity of this study.
Limitations

This study had its limitations. First, although this study analyzed subgroup differences such as gender and ethnicity, because of Kent State’s disproportionate ethnic population group sizes favoring Caucasians, generalizations to other ethnicities should be made with much more caution. There were disproportionately fewer non Caucasians than Caucasians in the sample. The highest specific minority was African American at 9%. All other ethnicities in the population, besides Caucasian and African American combined, were only an additional 10%:

100% Total - 81% Caucasian - 9% Afr. American = 10% all other ethnicities combined.

Second, this study was limited by the practicality of using university-provided email addresses, postal addresses, and telephone numbers. According to Kent State records, the address list was 97.37% complete (L.A. Carlson, personal communication, February 6, 2009). Therefore, the use of the university listing was a negligible limitation.

Third, instead of refuting all the methodological shortcomings of at least 38 studies mentioned earlier, the principle of Ockham’s Razor, which states, “all other things being equal, the simplest theory is the best” (Vogt, 1999, 2005), will be invoked. That is, if the surveys and interviews are asking the same questions, why would there be a difference? Instead, a publication bias may exist in the SDB literature. For simplicity, Vogt (2005) in his “Dictionary of Statistics and Methodology” defined publication bias as, “bias that arises when studies showing statistically significant results are more likely to get published.” (p. 252). Vogt (2005) also defined Fail Safe N as, “in Meta Analysis, the number of studies confirming the null hypothesis that would be necessary to change
the results of a Meta Analysis study that found a statistically significant relationship” (p. 119). However, in this study, specific Fail Safe N’s were used. Wolf (1986) as cited in Cooper (1979) defined a meta-analytic Fail Safe N as the “number of additional studies . . . necessary to reverse the overall probability . . . from . . . our combined test to a value higher . . . than our critical value for statistical significance.” (p. 38). Orwin’s Fail Safe N indicates only four non significant studies are required to overturn the literature’s position of differences in SDB by method. Cooper’s Fail Safe N, which is more liberal than Orwin’s Fail Safe N, indicates only negative seven non significant studies are required. Thus, no studies are required by Cooper’s Fail Safe N to overrule the SDB literature, since negative seven is less than zero. Both these Fail Safe N’s were based on obtaining a conservative targeted effect of .2, but as stated in Chapters I and III, the evidence in previous research indicated a mean effect size of .32. Thus, this study was conservative. The Fail Safe N calculations and the lack of rigor [a maximum of 17.1% (6 of 35) studies randomly assigned, and only one study randomly selected] in the SDB literature make a publication bias even more likely.

Fourth, even when considering its delimitations and limitations academics should also consider the current study’s strengths. If researchers compared only the methodological rigor of this study to all the studies cited within this study, this study compares very favorably because it randomly assigned, randomly selected, controlled for or avoided extraneous influences as urged by highly reputable educational researchers (Kerlinger, 1986; Kerlinger & Lee, 1999; Kirk, 1982, 1995; Pedhazur, 1984, 1997), attempted to investigate external validity, adequately reported effect size information,
used more than adequate sample sizes for even modest effect sizes such as .2 or .14, used appropriate yet powerful MCPs to control for Type I error, used local response rate information to help plan the data collection, and even initially randomized questions yet kept the order constant—after initial randomization—to compare four different data collection methods, which no study about SDB has done. For example, for research question one, only 6 (Cannel & Fowler, 1963; Wiseman, 1972; de Leeuw, 1992; Locander, Sudman & Bradburn, 1976; Booth-Kewley, Edwards, & Rosenfeld, 1992; Dahl, 1992) of 35 studies randomly assigned participants. Additionally, the only two studies about method influencing SDB that randomly selected participants are de Leeuw’s (1992) dissertation and the current investigation.

Suggestions for Future Research

Although this study was exploratory, it can be considered more rigorous than any study about method effects on SDB because it used random assignment, random selection, a large sample size, as well as emphasized internal validity through avoiding and testing for threats to specific counter hypotheses discussed in the SDB literature. For example, the Hancock and Flowers (2001) study—which was concerned with only the differences between Web-based and mail surveys—had the largest sample size, 176 participants, of any study besides the current study that also used random assignment. However, the current study used 819 participants, a 465 percent increase. Furthermore, unlike the current study, Hancock and Flowers (2001) did not randomly select nor did they test for or avoid any extraneous counter hypotheses.
Studies should be done with minority groups such as African Americans, Latinos, Asians, and Native Americans. This will further examine the external validity of social desirability. A study comparing how older, middle-aged, and young African Americans’ attitudes toward or expression of SDB, could prove fruitful. The age groups may be responding differently toward SDB elicitation. The older African Americans may indeed have responded more extremely. The same types of analyses could also be done with other minorities such as Latinos, Asians, and Native Americans. They could investigate age-group differences within each minority. Different age groups within or across minorities may express or have differing attitudes toward their own socially desirable behavior. Perhaps that is why African Americans and Latinos did not respond any differently. Alternatively, as America gets more multicultural and coarse, perhaps it is simply another indication socially desirable behavior is decreasing.

Studies possessing something easily and objectively verified would be extremely useful, such as cumulative Grade Point Average (GPA). This would provide a “gold standard,” whereby reported GPA’s by students with their actual GPA’s could be compared. The difficulties encountered are GPA may be affected by response error, but even more troublesome, this would raise both Family Education Rights & Privacy Act (FERPA) and Human Subject concerns.

Studies should be done on the differences between the cognitive processes of the respondents across all four methods. After all, if the cognitive processes of participants are very similar, then the products of the processes are more likely to be similar.
Therefore, focus groups and cognitive interviews should investigate the similarities and differences across the four methods in studying participants’ cognitive processes.

To date there have only been six studies about the factor structure, factor invariance, differential reliability, or differential validity of the BIDR (Beretvas & Meyers, 2002; Cruse, 2008; Decision Support Systems Research, 1999; Green & Goldman, 1993; Leite & Beretvas, 2005; Schmitt, 1994). To increase the efficiency and validity of survey research more studies on the reliability, validity, factor structure, and factor invariance of the BIDR, especially the IM subscale, should be conducted. This may help determine if SDB is stable enough to be studied among methods. If these differences prove to be too great, to be more efficient, researchers should investigate methods or measurement tools fine enough to isolate or more precisely measure SDB, especially IM.

Studies investigating the possibility of respondents’ extinguishing their socially desirable behaviors or considering other known strategies during the course of taking a survey or interview should be undertaken. However, what could be more practical, however, is searching for studies on a well-researched variable (e.g., students’ attitudes toward learning, locus of control studies, students’ perceptions of instructional quality, students’ perceptions of the internal or external control over their own learning, students’ attributions toward their academic achievements) to see if they possess information that show differences between halves of studies, more specifically of the instruments, to further investigate the Experimental Extinguishment Effect (EEE).

The investigation of a variable that has already been painstakingly and comprehensively researched (over time) to investigate the nature of method differences is
warranted. The Multi-Method Multi-Trait (MMMT) strategy can not only control the method of measurement as an issue, but also determine whether the trait itself is stable (e.g., students’ attitudes toward learning, students’ perceptions of instructional quality, students’ perceptions of the internal or external control over their own learning, students’ attributions toward their academic achievements, locus of control studies) across the four methods of data collection: face to face interview, telephone interview, mail survey, and Web survey. After all, if method differences of well-defined and intensely researched variables prove unstable across different methods, how can these methods adequately measure a situation specific bias, such as SDB?

If feasible, researchers should also thoroughly call for, search, collect, and study all non-published studies and its results of SDB, especially method studies. If enough non-published studies are found, a Meta Analysis may compare selected characteristics between non-published studies to published studies to explore whether a publication bias and any subtle moderator effects, such as question order, exists within the data across SDB studies.

Alternatively, based on this study and the Meta Analyses by Dindia and Allen (1992) and Richman, Kiesler, Weisband, and Drasgow (1998), SDB may be decreasing over time. This latest explanation seems the most likely given the available literature, specifically the two Meta Analyses asserting SDB effects have been decreasing over the years and the small effect sizes found in this current study. The suggested Meta Analyses in the previous paragraph should also include longitudinal analyses to further investigate
whether the Meta Analyses by Dindia and Allen (1992) and Richman, Kiesler, Weisband, and Drasgow (1998), support the proposition SDB may be decreasing over time.
APPENDICES
APPENDIX A

DEMOGRAPHICS
Figure 2. Ethnicity Percentage Breakdown
Figure 3. Age Range of Analyzed Participants during the Fall 2006 Semester

According to the available sample of 819 students: 199 (24.3%) were 18 years old; 159 (19.4%) were 19; 114 (13.9%) were 20; 116 (14.2%) were 21; 83 (10.1%) were 22; 28 (3.4%) were 23; 47 (5.8%) were 24-26; 15 (1.8%) were 27-30; 32 (3.9%) were 31-40; 15 (1.7%) were 41-50; and 11 (1.3%) were 51-70 years old. The overall mean age was 22.17, the median was 20, the mode was 18, and the standard deviation was 7.48 years. The skewness of the available sample was 3.35 and the kurtosis was 11.96.
APPENDIX B

SURVEY AND INTERVIEW QUESTIONS
Survey Questions

1. I have pretty awful habits.
2. I have never damaged a library book or store merchandise without reporting it.
3. I have never dropped litter on the street.
4. When I was young, I sometimes stole things.
5. I never cover up my mistakes.
6. I am very confident of my judgments.
7. My first impressions of people usually turn out to be right.
8. I sometimes tell lies if I have to.
9. When I hear people talking privately, I avoid listening.
10. I am fully in control of my own fate.
11. I don’t gossip about other people’s business.
12. I never take things that don’t belong to me.
13. I have taken sick-leave from work or school even though I wasn’t really sick.
14. I never read sexy books or magazines.
15. I never swear.
16. I always obey laws, even if I’m unlikely to get caught.
17. I never regret my decisions.
18. I sometimes lose out on things because I can’t make up my mind soon enough.
19. I sometimes drive faster than the speed limit.
20. I have received too much change from a salesperson without telling him or her.
21. There have been occasions when I have taken advantage of someone.
22. I have done things that I don’t tell other people about.

23. I sometimes try to get even rather than forgive and forget.

24. I have said something bad about a friend behind his or her back.

25. Which of these words is a verb?
   
   Please Choose This Particular Option

26. Which of the following numbers has the greatest absolute value?
   
   -400  -300  - 200  -100  0  333 ½

27. When was the Declaration of Independence for the United States signed?
   
   1576  1676  1776  1876  1976  1787

28. What is your gender (sex)?
   
   Male  Female  Decline to answer  Not applicable

29. Do you have a physical, emotional, or a cognitive disability?
   
   Yes  No  Decline to answer  Not Applicable

30. If participant answered “yes” to question 29, choose one word which best describes your disability:
   
   Physical  Emotional  Cognitive  Or, if other please specify _____________

31. What Ethnicity do you belong to?
   
   White  Black  Hispanic  International or Foreign  Asian  Native American  Other  Unknown

32. While you were answering these interview questions, were there any people less than 5 feet from you?
   
   Yes  No  I was too busy to notice this  Not applicable  Decline to answer
33. If you answered “yes” to question 32, what percentage of the interview so far did
the presence of others influence the way you answered the questions?

100%  90%  80%  70%  60%  50%  40%  30%  20%

10%  0%  Not applicable  Decline to answer  Do not know

*Figure 4. Survey and Interview Questions*
APPENDIX C

RESULTS OF DATA ANALYSES
Table 1

*The Mean Difference Between Methods*

<table>
<thead>
<tr>
<th>Method</th>
<th>Face to face</th>
<th>Telephone</th>
<th>Mail</th>
<th>Web-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to face</td>
<td>-.37</td>
<td>-.27</td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td>Telephone</td>
<td>.10</td>
<td></td>
<td>1.31</td>
<td></td>
</tr>
<tr>
<td>Mail</td>
<td></td>
<td></td>
<td></td>
<td>-1.21</td>
</tr>
</tbody>
</table>
Table 2  
*Means and Standard Deviations of Caucasian and Non-Caucasian Participants*

<table>
<thead>
<tr>
<th>Method</th>
<th>Caucasian (N=141)</th>
<th>Non-Caucasian (N=33)</th>
<th>Caucasian (N=133)</th>
<th>Non-Caucasian (N=40)</th>
<th>Caucasian (N=266)</th>
<th>Non-Caucasian (N=36)</th>
<th>Caucasian (N=142)</th>
<th>Non-Caucasian (N=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to face</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>60.48</td>
<td>63.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.29</td>
<td>10.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>61.36</td>
<td>61.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.02</td>
<td>10.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>62.18</td>
<td>60.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>11.07</td>
<td>13.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>60.74</td>
<td>55.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.91</td>
<td>10.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3

*Duration of the Face to Face Interview Across Interviewers*

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Standard Error of the Mean</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.00 minutes</td>
<td>4.00 minutes</td>
<td>5.70 minutes</td>
<td>1.06 minutes</td>
<td>1.61</td>
<td>2.60</td>
</tr>
</tbody>
</table>

For each interviewer, the mean was 6.00 minutes, the median was 4.00 minutes, the standard deviation was 5.70 minutes, and the standard error of the mean was 1.06 minutes. Skewness was 1.61 and kurtosis was 2.60.
Table 4

*Duration of Telephone Interviews Across Interviewers*

<table>
<thead>
<tr>
<th>Range of Single Telephone Interviewer’s Mean</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Error</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shortest Individual Call</th>
<th>Longest Individual Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.00 – 13.17 minutes</td>
<td>9.90 minutes</td>
<td>3.40 minutes</td>
<td>.26 minutes</td>
<td>1.54 minutes</td>
<td>3.29 minutes</td>
<td>5 minutes</td>
<td>26 minutes</td>
</tr>
</tbody>
</table>

The research shows falsification is linked to implausibly brief elapsed interview times (Bushery, Reichert, Albright & Rossiter 1999; Cannell, Miller & Oskenberg, 1981; Forsman & Schreiner, 1991). The logic is if interviewers falsify, they are likely to do so because of time pressures. Analogously, a thief would not “camp” at the scene of his crime for 48 or more hours. The elapsed time for telephone interviews indicated the range of a single telephone interviewer’s mean time taken was between 7.00 and 13.17 minutes and the mean, median, standard deviation, and standard error of the mean were 9.90, 9.00, 3.40 and .26 minutes, respectively. The skewness and kurtosis were 1.54 and 3.29, respectively. The shortest and longest individual telephone interviews were 5 and 26 minutes long, respectively. The mean workload was 8.7 interviews per interviewer.
Table 5

**Eliminating or Controlling Extreme Response Style (ERS) Bias**

<table>
<thead>
<tr>
<th></th>
<th>Panel A: ANOVA Results of baseline (No ERS eliminated (N = 819))</th>
<th>Panel B: ANOVA Results of Bipolar ERS eliminated (N = 777)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>df&lt;sub&gt;numerator&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>2.73</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2.33</td>
<td>3</td>
</tr>
</tbody>
</table>
Panel C: ANOVA Results of Low ERS eliminated (N = 777)

<table>
<thead>
<tr>
<th>F</th>
<th>df&lt;sub&gt;numerator&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>p</th>
<th>( \eta_p^2 )</th>
<th>Levene's Homogeneity</th>
<th>Prob. of</th>
<th>Ryan's of Levene (REGWQ)</th>
<th>Prob. of Ryan</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.58</td>
<td>3</td>
<td>761</td>
<td>0.05</td>
<td>0.01</td>
<td>1.16</td>
<td>0.3</td>
<td>0.95</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Panel D: ANOVA Results of High ERS eliminated (N = 777)

<table>
<thead>
<tr>
<th>F</th>
<th>df&lt;sub&gt;numerator&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>p</th>
<th>( \eta_p^2 )</th>
<th>Levene's Homogeneity</th>
<th>Prob. of</th>
<th>Ryan's of Levene (REGWQ)</th>
<th>Prob. of Ryan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.09</td>
<td>3</td>
<td>761</td>
<td>0.35</td>
<td>0.004</td>
<td>1.5</td>
<td>0.1</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Panel E: ANCOVA Results adjusted for Item Variances (Covariate) (N = 819)

<table>
<thead>
<tr>
<th>F</th>
<th>df&lt;sub&gt;numerator&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>p</th>
<th>( \eta_p^2 )</th>
<th>Levene's Homogeneity</th>
<th>Prob. of</th>
<th>Levene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.59</td>
<td>3</td>
<td>814</td>
<td>0.19</td>
<td>0.006</td>
<td>1.51</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>
Table 6

Response Rate Information by Method

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sample Size</th>
<th>Sample Frame Size</th>
<th>Percentage$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-Face</td>
<td>174</td>
<td>2,940</td>
<td>5.92%</td>
</tr>
<tr>
<td>Telephone</td>
<td>173</td>
<td>1,220</td>
<td>14.18%</td>
</tr>
<tr>
<td>Mail</td>
<td>302</td>
<td>2,430</td>
<td>12.43%</td>
</tr>
<tr>
<td>Web-Based</td>
<td>170</td>
<td>1,484</td>
<td>11.46%</td>
</tr>
<tr>
<td>Total</td>
<td>819</td>
<td>8,074</td>
<td>10.14%</td>
</tr>
</tbody>
</table>

$^1$The percentage is calculated by taking the sample size/sample frame size.
Table 7

*IM Summary Statistics Across Methods and Questions*

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Stand Dev</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to Face</td>
<td>61.00</td>
<td>61.00</td>
<td>60.00</td>
<td>10.31</td>
<td>-.26</td>
<td>.95</td>
<td>174</td>
</tr>
<tr>
<td>Telephone</td>
<td>61.41</td>
<td>61.00</td>
<td>47.00</td>
<td>10.12</td>
<td>.00</td>
<td>-.16</td>
<td>173</td>
</tr>
<tr>
<td>Web</td>
<td>59.85</td>
<td>60.00</td>
<td>65.00</td>
<td>10.93</td>
<td>-.36</td>
<td>.33</td>
<td>170</td>
</tr>
<tr>
<td>Mail</td>
<td>62.00</td>
<td>62.00</td>
<td>58.00</td>
<td>11.34</td>
<td>-.14</td>
<td>-.05</td>
<td>302</td>
</tr>
<tr>
<td>Total IM</td>
<td>61.26</td>
<td>61.14</td>
<td>65.00</td>
<td>10.69</td>
<td>-.20</td>
<td>.40</td>
<td>819</td>
</tr>
<tr>
<td>Question 1</td>
<td>3.89</td>
<td>4.00</td>
<td>4.00</td>
<td>.95</td>
<td>-1.01</td>
<td>1.08</td>
<td>816</td>
</tr>
<tr>
<td>Question 2</td>
<td>3.68</td>
<td>4.00</td>
<td>5.00</td>
<td>1.55</td>
<td>-.78</td>
<td>-.97</td>
<td>795</td>
</tr>
<tr>
<td>Question 3</td>
<td>2.81</td>
<td>3.00</td>
<td>1.00</td>
<td>1.43</td>
<td>.05</td>
<td>-1.38</td>
<td>814</td>
</tr>
<tr>
<td>Question 4</td>
<td>4.18</td>
<td>5.00</td>
<td>5.00</td>
<td>1.26</td>
<td>-1.52</td>
<td>1.11</td>
<td>812</td>
</tr>
<tr>
<td>Question 5</td>
<td>2.70</td>
<td>3.00</td>
<td>2.00</td>
<td>1.15</td>
<td>.16</td>
<td>-.85</td>
<td>806</td>
</tr>
<tr>
<td>Question 8</td>
<td>3.39</td>
<td>4.00</td>
<td>4.00</td>
<td>1.12</td>
<td>-.64</td>
<td>-.35</td>
<td>815</td>
</tr>
<tr>
<td>Question 9</td>
<td>2.93</td>
<td>3.00</td>
<td>2.00</td>
<td>1.21</td>
<td>.03</td>
<td>-1.00</td>
<td>819</td>
</tr>
<tr>
<td>VARIABLE</td>
<td>Mean</td>
<td>Median</td>
<td>Mode</td>
<td>Stand Dev</td>
<td>Skewness</td>
<td>Kurtosis</td>
<td>Sample Size</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Question 11</td>
<td>2.99</td>
<td>3.00</td>
<td>4.00</td>
<td>1.19</td>
<td>-.11</td>
<td>-.94</td>
<td>817</td>
</tr>
<tr>
<td>Question 12</td>
<td>4.05</td>
<td>4.00</td>
<td>5.00</td>
<td>1.21</td>
<td>-1.19</td>
<td>.31</td>
<td>816</td>
</tr>
<tr>
<td>Question 13</td>
<td>3.08</td>
<td>4.00</td>
<td>1.00</td>
<td>1.58</td>
<td>-.19</td>
<td>-1.55</td>
<td>816</td>
</tr>
<tr>
<td>Question 14</td>
<td>2.84</td>
<td>3.00</td>
<td>1.00</td>
<td>1.55</td>
<td>.18</td>
<td>-1.49</td>
<td>814</td>
</tr>
<tr>
<td>Question 15</td>
<td>1.73</td>
<td>1.00</td>
<td>1.00</td>
<td>1.12</td>
<td>1.45</td>
<td>1.07</td>
<td>815</td>
</tr>
<tr>
<td>Question 16</td>
<td>3.31</td>
<td>4.00</td>
<td>4.00</td>
<td>1.28</td>
<td>-.43</td>
<td>-.93</td>
<td>815</td>
</tr>
<tr>
<td>Question 19</td>
<td>2.05</td>
<td>2.00</td>
<td>2.00</td>
<td>1.23</td>
<td>.93</td>
<td>-.27</td>
<td>808</td>
</tr>
<tr>
<td>Question 20</td>
<td>4.14</td>
<td>5.00</td>
<td>5.00</td>
<td>1.23</td>
<td>-1.46</td>
<td>1.06</td>
<td>780</td>
</tr>
<tr>
<td>Question 21</td>
<td>3.88</td>
<td>4.00</td>
<td>4.00</td>
<td>1.11</td>
<td>-1.12</td>
<td>.77</td>
<td>811</td>
</tr>
<tr>
<td>Question 22</td>
<td>2.66</td>
<td>3.00</td>
<td>1.00</td>
<td>1.37</td>
<td>.08</td>
<td>-1.38</td>
<td>815</td>
</tr>
<tr>
<td>Question 23</td>
<td>3.67</td>
<td>4.00</td>
<td>4.00</td>
<td>1.15</td>
<td>-.83</td>
<td>-.01</td>
<td>817</td>
</tr>
<tr>
<td>Question 24</td>
<td>3.23</td>
<td>4.00</td>
<td>4.00</td>
<td>1.26</td>
<td>-.57</td>
<td>-.80</td>
<td>815</td>
</tr>
</tbody>
</table>
REFERENCES


McHorney, C. A., Kosinski, M., & Ware, J. E. (1994). Comparisons of the costs and quality of norms for the SF-36 Health Survey collected by mail versus telephone interviews: Results from a national survey. *Medical Care, 32*(6).


