ABSTRACT

STANDISH, III, WILLIAM RARIDEN. A Validation Study of Self-Reported Behavior: Can College Student Self-Reports of Behavior Be Accepted as Being Self-Evident? (Under the direction of Paul D. Umbach).

The dearth of institution-reported data describing college student behaviors forces higher education researchers to rely upon self-reported behaviors collected via surveys and summarized by the survey statistic. If bias is present and substantive, the invalid, self-reported input data threatens the validity of research conclusions that guide co-curricular and academic policies and best practices at institutions of higher education. This study evaluates the premise of self-reported data validity upon which survey research is based. Potential sources of bias include (1) response bias from inaccurate self-reporting and (2) nonresponse bias where characteristics of respondents differ from those of nonrespondents. The assumption of survey statistic validity depends upon it being an unbiased estimate of behavior in a target population that relies on survey responses being free from systematic biases, an assumption that has not been adequately tested in the literature. This validation study of self-reported behaviors compares institution-reported, transactional data to corresponding self-reported academic performance, class attendance, and co-curricular participation from a sample of 6,000 students, using the Model of the Response Process by Tourangeau (1984, 1987). Response bias, observed as measurement error, is significant in 11 of the 13 questions asked and evaluated in this study. Socially desirable behaviors include campus recreation facility (CRF) use and academic success being overstated as much as three times. Nonresponse bias, observed as nonresponse error, is also significant in 11 of the same 13 questions asked and evaluated with high GPA and participatory students over represented in the survey statistic. For most of the questions, measurement error and nonresponse error

combine to misstate behavior by at least 20%. The behaviors most affected are CRF use, which is overstated by 112% to 248%; semester GPA self-reports of 3.36 versus an actual value of 3.04; and co-curricular participation that misstated by between -21% to +46%. This validation study sufficiently demonstrates that measurement error and nonresponse error are present in the self-reported data collected for the commonly studied topics in higher education that were represented by the 13 questions. Researchers using self-reported data cannot presume the survey statistic to be an unbiased estimate of actual behavior that it is generalizable to larger populations.

A Validation Study of Self-Reported Behavior: Can College Student Self-Reports of Behavior Be Accepted as Being Self-Evident?

by
William Rariden Standish, III

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Educational Research and Policy Analysis

Raleigh, North Carolina

2017

APPROVED BY:

Dr. Paul D. Umbach Committee Chair	Dr. Stephen Porter
Dr. Michael Mullen	Dr. Carrie Zelna

DEDICATION

This work is dedicated to my family,

especially my parents, my grandparents, my children Paige and Will, and my wife Adrian.

BIOGRAPHY

Trey Standish (more formally William Rariden Standish, III) was born in Indianapolis and raised in Charlotte and Glen Ellyn, Illinois. Since earning a Bachelor's of Science in Economics from North Carolina State University, he has remained dedicated to the idea of living in Raleigh where he has resided since 2006 with his wife Adrian and children Paige and Will. In between, he earned a Master's of Arts in Higher Education Administration and a Master's of Arts in Sports Management from Appalachian State University in Boone, NC, a Master's of Arts in Applied Economics from The University of North Carolina – Greensboro in Greensboro, NC, and forecasted McDonald's Happy Meal toys from Chicago, IL.

Values instilled by his family and experiences at North Carolina State University cause Trey to believe in the mission of public higher education as historically applied by the state of North Carolina. Students that attend the universities and the communities surrounding them are incalculably improved by their investment. Trey's career began in student affairs, and he continues his service as he quantitatively supports North Carolina State University as an institutional researcher, enrollment planner, and currently the Senior Research Data Analyst in the office of Enrollment Management and Services. In these roles, he forecasts student enrollment, evaluates the experiences of students, and assists university stakeholders with program development.

In his spare time, Trey enjoys spending time with his family and pursues several hobbies including woodworking, reading, gardening, and watching NC State basketball. He is particularly proud of earlier-in-life accomplishments including competitive chess, completing an iron distance triathlon, and independent international travels.

ACKNOWLEDGMENTS

The process of writing this dissertation has challenged me in almost every way. Its completion is both a relief and a great reward. I could not have completed this this work without the support of my university, communities, and family whom I would like to acknowledge.

First, I would like to thank my committee for their insight and guidance. My chair, Dr. Paul Umbach, has an encyclopedic mastery of the literature and has efficiently guided me to the best references and resources to assist my studies. Dr. Stephen Porter has provided theoretically based direction that immensely improved this work. Dr. Carrie Zelna was among my first staff contacts as an undergraduate student and remains a valued mentor in my student and professional life. Dr. Michael Mullen is a continued voice of encouragement and inspiration for my professional pursuits. I would also like to thank all the faculty and staff from many academic units that contributed to my dissertation's completion. The student support resources offered by North Carolina State University are invaluable and adaptive to student needs.

I would not have been able to complete this dissertation without the support of my employing offices at NC State. My supervisors in both the Office of Institutional Research and Planning and Enrollment Management and Services have been incredibly understanding in the time and resources they have enabled me to use. In particular, I would like to thank Karen Helm for reminding me of the degree's importance and Nancy Whelchel (with the help of her survey team) for her help in creating, validating, and implementing my survey. Everyone in the Office of Institutional Research and Planning, University Recreation, and the

Division of Academic and Student Affairs has provided timely and accurate data without which I could not continue my study.

I would also like to thank my family and friends who maintained a supporting and understanding community from which I could work. Each in your own way have contributed to my dissertation's completion. In particular, I would like to thank my mother for allowing me to vent frustrations, and my parents, siblings, and other close relatives who have worked around my schedule. For many years, I remained hard to find and distracted from my relationships while I wrote the dissertation. I have missed you all and look forward to spending more time with you all.

Finally, I would like to thank my wonderful wife Adrian for all that she has done during the years I dedicated to this study. She has lived with and parented with a husband whose mind was constantly split between academics and being a husband and father. She has been a constant source of support, motivation, and encouragement throughout my studies, and I would not have been able to complete this project without her. She has reviewed and edited this work many times and driven us so many places while I worked from the passenger seat. I cannot adequately express my gratitude and am thankful to be her husband. Most of all, I look forward to leaving this study behind to spend more time with her and our children, Paige and Will.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This validation study tests the validity of self-reported student behaviors using institution-reported validation data. A sample of North Carolina State University (NCSU) undergraduates received a survey instrument to self-report co-curricular participation, class attendance, and academic performance. Student self-reports are merged with corresponding institution reported records to test for the presence of biases and the suitability of the survey statistic as an estimate of actual sample and target population behavior. This topic is relevant to higher education research because it tests the assumption that self-reported data on college student behavior is a valid measure of actual behavior.

Importance of Study

How student participation and engagement with collegiate curricular and co-curricular programs affect student outcomes is a frequently researched topic in the field of higher education. The National Survey of Student Engagement (NSSE) is a leading source of self-reported data on student participation; the NSSE includes the caveat that the survey's findings are dependent on the validity in self-reported behaviors (Kuh, Hayek, Carini, Ouimet, Gonyea, & Kennedy, 2001). Other national and institution-specific surveys on college student behaviors make similar assumptions (Porter, 2011). Historically, higher education researchers relied on self-reported behaviors because data storage and hardware limitations made incorporating actual participation records into statistical software packages impractical or impossible (Porter, 2011). However, the cost of computer storage space and processing power continues to decrease making institution-reported behavioral data ever more accessible for higher education research. In the case of this study, student-level campus recreational facility (CRF) use data has been archived since 2006 allowing an actionable amount of CRF use data to accrue. Each CRF transaction records the student who entered, the time of entry, and their purpose of entry. In the context of this study, there are no technical barriers to using CRF transactional data to compare actual CRF use to self-reported student descriptions.

Survey methods are an indispensable tool in the study of higher education. The validity of self-reported, survey data is instrumental to interpreting research conclusions that describe the effect of behavior on student outcomes. Sampling and survey data collection are cost effective methods for estimating the occurrence of specific behaviors among the target population (Pike, 1999a). Higher education researchers estimate gains from curricular interventions by collecting baseline data and comparing them to subsequent reports (Pike, 1999a). Other researchers use survey data to measure institutional climate and conduct needs assessments that shape services to benefit their campus community. Survey methods can collect summative data for accountability and formative data to improve institutional programs. Whatever a researcher's goal, customized survey instruments collect behavioral measures that specifically target the researcher's questions (Kuh et al., 2001). For these researchers, the validity of survey data is a meaningful issue that is rarely explored (Pike, 1995).

Institutional researchers, cross-institutional groups, and faculty researchers study psychological development, learning, and public policy by collecting survey data to describe student behavior (Pike, 1999a). For institutions of higher education, participant self-assessment methodologies are a best practice (and an accreditation agency requirement) for evaluating curricula on improving student outcomes (Astin & Antonio, 2012; Eaton, 2012). Federal, accreditor, and state agencies are examples of cross institutional groups which rate and regulate institutions of higher education with ever-expanding scope and influence (Dougherty, Natow, Bork, Jones, & Vega, 2013; Eaton, 2012). To evaluate student safety and satisfaction, these agencies mandate campus climate surveys about sensitive behaviors; requirements to survey students about sexual assaults are a recent example (Koss, Wilgus, & Williamsen, 2014). Finally, public or institutional policy is guided by the work of ad hoc researcher teams' aggregated individual surveys used to estimate the occurrence of behaviors theorized to impact outcomes. For example, universities may require students to live in campus residence halls if survey data suggests it improves the likelihood of degree attainment. Collect direct measures of past behavior is prohibitively costly so researchers instead use survey samples to make inferences about target populations (Tourangeau, Rips, & Rasinski, 2000). The validity of self-reported behavioral data is germane to the use of samples as substitutes for populations.

The Model of Response Process (Tourangeau, 1984; 1987) is a four-part theoretical framework for accessing the validity of self-reported data. A valid survey response requires the respondent to (1) comprehend the survey question as intended; (2) correctly retrieve the information requested; (3) make good judgments when imputing and reporting missing memories; and (4) then report their response accurately and truthfully (a deeper discussion of the Model of Response Process will follow). Participants are known to unwittingly provide inaccurate information on survey questions for several reasons. Many respondents are unable to accurately retrieve memories; previous studies find respondents describe past behavior with circumstantial evidence rather than actual memory (Tourangeau, et al, 2000). Additionally, respondents often intentionally misreport behavior for sensitive subjects, which is sometimes referred to as intentional misreporting. As Tourangeau, et al. (2000) explains, "This notion of sensitive questions presuppose that respondents believe there are norms defining desirable attitudes and behaviors, and that they are concerned enough about these norms to distort their answers to avoid presenting themselves in an unfavorable light." (p. 257). Social desirability bias describes the systematic phenomenon of respondents inflating the incidence of socially accepted behavior by lying or biasing their responses in the acceptable direction. Within the context of the Model of Response Process (Tourangeau, 1984, 1987), this study tests the validity of self-reporting behaviors with an emphasis on how accurately memories are retrieved, how social desirability bias distorts the survey statistic, and the generalizability of the survey statistic to the target population.

Statement of Problem

Though much research has addressed outcomes of co-curricular and academic experiences of college students, the premise upon which those studies have been conducted needs further validity study. Because researchers have been unable to collect original data participation records for such experiences, much reliance has been placed upon self-reported survey data. However, because participants may not provide reliable data when surveyed, this methodology needs further examination. This validation study uses institution reported data sets to tests the presumed validity of self-reported college grades, college class attendance, and co-curricular participation. Results help evaluate the trustworthiness of past studies that use survey data to evaluate student outcomes.

Purpose of Study

The purpose of this study is to validate college student self-reported behaviors with an emphasis on social desirability bias and the accuracy of recent memories compared to those of months or years. Past validation studies are conducted neither for large numbers of students nor for behavioral topics often used to evaluate higher education outcomes. As such, the literature establishes an unmet need for this validation study.

Research Questions

The research questions use the Model of the Response Process by Tourangeau (1984, 1987) to evaluate the accuracy of self-reported survey data. Specifically, the research questions are:

1. To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?
    a. What student characteristics are associated with the most accurate and inaccurate self-reporters? Do demographic or background characteristics affect accuracy in self-reporting? Do students with high collegiate GPAs or standardized test scores more accurately self-report behaviors? Do participatory students provide more accurate self-reported behavioral data?
    b. Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?

c. Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?

d. Are sample means of self-reported behaviors valid estimates of target population means?

2. Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?

   a. Do social norms cause students to systematically over report the incidence of socially desirable behaviors such as physical activity, class attendance, and high grades earned? Are self-reported respondent means of socially desirable behaviors less valid estimates of behaviors than those from more socially neutral behaviors?

   b. Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?

The first research question (RQ1) tests the retrieval and judgment components of the Model of the Response Process by testing the validity of self-reported data as a representation of actual behavior across the target population. Student characteristics are tested for a statistically significant correlation with the student's accuracy in self-reporting, including measures of student ability, observed co-curricular participation, and other institution reported characteristics. The second research question (RQ2) tests the reporting components by testing for systematically invalid variance that can be predicted by the perceived social desirability of the behavior and the respondent's propensity for self-monitoring (e.g., presenting one's self more favorably than is actually true).

Finding systematic biases is generally cause for concern but not a prima fascia invalidation of the survey statistic. Random differences between the survey statistic and actual population means are inevitable in the sampling process and can be corrected with statistical techniques. In some cases, the systematic bias may improve the survey statistic because respondents make adjustments that more accurately measure the construct as intended by the researcher (Kuncel, Crede, & Thomas, 2005). In other cases, the magnitude and direction of misreporting is consistent across all respondents, resulting in a level shift of the survey statistic that does not affect the correlative relationship (Kuncel et al., 2005). In most cases, however, systematic bias is a threat to the validity of a study's findings, and the systematic biases in self-reporting must be identified and the degree of threat bias poses to the validity of the research conclusions determined.

Significance of Proposed Research

The validity of the survey statistic as an estimate of actual target population behaviors is theoretically well explored, but there remains an unmet need for validation studies on self-reported college student behavior (Porter, 2011). Despite this, the study of higher education has progressed onward because "the tacit agreement in postsecondary research seems to be that validity is assumed until proven otherwise" (Porter, 2011, p. 73). This study questions the presumed validity of self-reported behavior by testing the hypothesis that respondents more accurately self-report data for more recent and noteworthy events and events for which they are not motivated to misrepresent their responses by inflating socially desirable behavior. This research is unique because it draws upon access to a directly measured validation data set, which describes student level co-curricular CRF use, Health/Exercise Sciences (HES) class attendance, and collegiate grades. The institution reported data allows for location of predictors of accurate reporting with an MCSDS score, characteristics of students, and behaviors of students. Though prior studies have isolated portions of the data available here, there are no known examples of validation studies utilizing such diverse data.

Tourangeau (1984, 1987) theorizes that taxing respondents' memory causes less accurate retrieval of information, particularly for memories of periods further in the past and for more detailed events; less accurate retrieval materializes as less accurate reporting. The current study tests Tourangeau's theory by asking respondents to describe behaviors retrieved from recent events (7 days) to distant events (months or years). Furthermore, Tourangeau et al. (2000) hypothesize that respondents tend to edit survey responses for questions on socially desirable items like class skipping and exercising regularly. This study tests the hypothesis by asking respondents to self-report behaviors that can be biased toward social norm conformity. Confirmation of these theories may call into question the validity of complicated and/or socially desirable behaviors commonly self-reported in surveys.

At the local level, I test the research questions using data focused upon behaviors occurring at NC State and the Carmichael Complex. The Carmichael Complex is a hub of academic and co-curricular activity that affects the educational experience of all NC State undergraduates. Over 90% of freshmen visit the CRF in their first year, and the majority participate in formal co-curricular programs ranging from fitness classes to intercollegiate club sports to leadership courses. Furthermore, NCSU requires all bachelor's degree students to complete two HES academic courses, nearly all of which meet in the CRF. Offices for academic and student affairs professionals are located within the facility to enable frequent student to faculty/staff contact. Almost all prior CRF assessment uses surveys describing CRF facilities as assets for building community, wellness, and facilitating engagement (Bryant & Bradley, 1993; Lindsey & Sessoms, 2006; Hall, 2006; Haines, 2001; Miller, Noland, Rayens, & Staten, 2008). Finally, the CRF is capital intensive and implies accountability from stakeholders (Mahoney, 2011; Upcraft & Schuh, 1996; Supiano, 2008).

Theoretical Framework

The research in this study is conducted within the Model of the Response Process by Tourangeau (1984, 1987) as applied to factual questions (Tourangeau et al., 2000). Widely accepted among survey researchers, the Model of Response Process stratifies considerations of survey respondents into four components: comprehension, retrieval, judgment and reporting, and response selection. A thorough understanding of these four components is critical to the current study (a more detailed discussion follows in chapter 2).

Comprehension. Comprehension describes respondent issues related to understanding the survey instructions, understanding the question, and understanding the information requested (Tourangeau et al., 2000). An example of a comprehension component threat to validity is the mismatch between information requested by the researcher and how the request is understood by the respondent; this mismatch often occurs because word definitions are unknown or not commonly perceived.

Retrieval. Retrieval for factual questions addresses the retrieval of relevant information from memory, incorporating a retrieval strategy to aid the recall of memories, and the inferential conclusions reached from partially recalled memories (Tourangeau et al., 2000). An example of a retrieval component threat to validity is a participant's inability to accurately recall events from long term memory; this is especially problematic if the request is complex or requires descriptions of long-past, day-to-day events.

Judgment. The judgment component for factual questions involves how a respondent merges various retrieved memories or supplements partially retrieved memories with imputed information (Tourangeau et al., 2000). Imputed information describes how a respondent uses retrieved information and estimation strategies that draw conclusions on aggregate behaviors by (1) recalling and counting memories; (2) making estimates of rates per time period multiplied by number of time periods; or (3) assigning total rates from vague impressions of past behavior (Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2009). An example of a judgment component threat to validity occurs when respondents cannot retrieve a fact and instead describe their behavior based on an impression of their behavior compared to an arbitrary understanding of average behavior.

Reporting. Reporting and response selection describes how respondents map their best estimates of a behavior to the possible responses offered by the survey instrument (Tourangeau et al., 2000). An example of a reporting and response component threat to validity is the desire to present oneself more favorably than the retrieved memory recalls (i.e. self-monitoring).

Within the Model of the Response Process by Tourangeau (1984, 1987), respondents are not expected to consciously differentiate between the components when formulating their response to a survey question. There is no clear delineation between components, and factors that influence responses in one component may cause responders to reconsider their response from a prior component. Respondents do not treat the components as linear steps and instead move iteratively between the components, often reconsidering their initial response in light of considerations developed in later components (Tourangeau et al., 2000). If respondents are comfortable with their answer or do not seriously consider the survey request, they may not use all four components when answering a survey. The layout of the survey instrument and question design affects the respondent's navigation through the four components, making the accuracy of the self-reported behaviors a function of interaction between the survey instrument and the respondent. While one component doesn't exist in isolation, the research questions of this study focus on the retrieval, judgment, and response and reporting components of the Model of Response Process.

Prior Research

In survey research, error refers to differences between information that the researcher desires to collect from a sample and the information that is actually true about the target population. Errors are categorized as measurement error and errors of nonobservation. Errors of nonobservation are unrelated to the survey instrument but are attributable to differences between the sample selected and target population (Groves et al., 2009). Measurement error is attributed to the survey methods/instrument and refers to differences between information that the researcher desires to collect and what is actually collected (Groves et al, 2009). Measurement error is expected and unavoidable; however, randomly distributed errors rarely bias the survey statistic because over and under reporting tend to offset. However, the researcher must be concerned by the presence of systematic nonrandom errors because such errors do not offset and can bias the survey statistic, thereby threatening the validity of research conclusions.

The first focus of this study is on testing for measurement errors in self-reported behaviors caused by the survey instrument. Systematic, non-random errors committed across large portions of the sample result in response bias in the survey statistic. Response bias occurs when the survey instrument causes significant portions of the sample to systematically deviate their self-reported measures from the true value in the same direction and/or magnitude. The response bias is measured as the difference between the expected value of an individual's response and the actual value across the target population of the survey. Common causes for the response bias include an inability to retrieve the requested information properly and intentional misreporting of the information to present one's self in a more favorable light (Groves et al., 2009). This study estimates the deviation between the expected value of a response (i.e. sample mean) and the mean of the measure across the target population and attempts to explain the differences by other factors. This is framed within the context of information retrieval problems and intentional misreporting. Researchers have debated the presence and effect of measurement error on the validity of research conclusions, but few have validated self-reported behaviors among college students.

Nonresponse bias is measured as the difference between the actual, aggregated measures of behavior among respondents and nonrespondents that distort survey means. Thus, the second focus of this study is testing for nonresponse errors in self-reported behaviors that result from differences in actual behavior among respondents and nonrespondents. Systematic differences in nonresponse propensities across the sample result in nonresponse bias in the survey statistic. Nonresponse bias is a threat to validity when the characteristics that define nonresponse propensity are correlated with or caused by the surveyed behavior (Groves, 2006). This study estimates the deviation between the actual behavior of respondents (i.e., sample mean) and the mean of the measure across the sample, including nonrespondents. Researchers have explored the theoretical aspects of nonresponse bias on the validity of research conclusions, but data availability constrains the analysis of nonresponse error in self-reported behaviors among college students.

The accuracy of self-reported standardized test scores and grade point averages is well explored because validation data is easily accessible (Kuncel et al., 2005). Self-reported academic performance and measures of ability are susceptible to social desirability bias and are often inflated and rarely underestimated. Students with high GPAs are more likely to respond to survey requests and introduce nonresponse error (Porter & Whitcomb, 2005). Attempts to validate college student academic behaviors like class skipping and studying are hampered by indirectly measured validation data. For example, Schmitt, Oswald, Kim, Gillespie, Ramsay, and Yoo (2003) asked students to self-report class skipping and used their final grade in each course as a proxy for attendance rates because they did not have access to actual attendance records. Other researchers possess directly measured validation data, but the construct being measured is unclear with definitions open to interpretation. For example, Porter, Rumann, and Pontius (2011) validated a nationally administered survey question about student reading requirements by obtaining course syllabi across large swaths of courses but found that the question wording only loosely corresponds to the realities of modern reading assignments. This study uses a direct measure of class skipping for HES courses and administers a survey instrument specifically designed to solicit responses about the available data.

Student development theory consistently finds evidence that co-curricular participation has a positive effect on grades and persistence (Torres, 2011), but the field remains wanting for direct institution reported measures of participation (Upcraft & Schuh, 1996). Surveys are often used to describe student behavior because the lack of institutional data forces higher education researchers to rely on self-reports of participation and participation intensity (Porter, 2011). For example, many studies evaluate the impact of Living Learning Villages (LLV) on student outcomes but rely on survey data to discern the intensity of curricular participation and feelings of community inclusion (Stassen, 2003; Pike, 1999a; Pike, Schroender, & Berry, 1997; Inkelas & Soldner, 2011). Meanwhile, institution-reported data typically describes participation as binary without regard to a number of a student's day-to-day participation intensity. This study uses institution-reported databases to compare actual behaviors with self-reported behaviors on the number of times a student attends CRF-based, and co-curricular participation in activities such as intramural sports, club sports, Outdoor Adventures, etc.

The presence of social desirability bias in self-reported data is well established. Respondents tend to over report socially desirable behaviors like seat belt use and voting while underreporting socially undesirable behaviors like smoking and illicit drug use (Tourangeau & Yan, 2007). Medical research recommendations and societal norms establish physical activity and exercise as socially desirable behavior susceptible to social desirability bias (Motl, McAuley, & DiStefano, 2005). Tourangeau, Smith, and Rasinski (1997) found that exercise is extremely susceptible to social desirability bias with over reporting of exercise rates exceeding that of taboo sexual acts, drug use, cigarette smoking, drinking and driving, and abortions. Because students view learning and developmental gains as socially desirable, those topics are susceptible to social desirability bias as students over report or overestimate the impact of their curriculum (Bowman, 2014; Pike 1999b). Therefore, the validity of findings that extol the benefits from CRF programs and informal recreational activity (Rothewell & Theodore, 2006) require further evaluation using a validation data set to locate biases.

Overview of the Study

In all manner of settings, researchers collect survey data to describe respondents' past behaviors and to make subsequent conclusions about how behaviors affect outcomes. Rather than focus on responses to an individual survey, most survey research summarizes data from completed surveys to create the survey statistic or a concise description of behaviors within the target population. The survey statistic is an input to models estimating the effect of student behaviors on student outcomes; the validity of models is dependent on the validity of the survey statistic. Groves et al. (2009) describes a valid survey construct as one for which the survey instrument clearly articulates a question for which respondents accurately report information, as intended. The survey statistic also requires respondents to possess similar characteristics as Nonrespondents, or it will over represent behaviors correlated with nonresponse propensity (Groves, 2006). Survey research conclusions on the study of higher education are dependent upon the validity of self-reported input data. To assume the survey statistic is a valid, unbiased estimate of behavior in a target population, the individual surveys must be free from systematic bias. This validation study uses a large transactional data set to detect systematic bias in self-reported behavior described by a survey statistic.

Data

Data for this study is collected from several units at NCSU, including The Office of Institutional Research and Planning (OIRP), Division of Academic and Student Affairs (DASA), and data collected from a custom survey instrument. After obtaining Institution Review Board (IRB) approval, I requested data from the respective units and conducted my survey with a 6,000 student sample randomly selected from the target population of all NCSU bachelor's degree students. OIRP uses student identification numbers to combine data sets but protects student identities by stripping personally identifiable information from the research data set.

Study Sample. The target population is the "group of elements for which the survey investigator wants to make inferences by using sample statistics" (Groves et al., 2009, p. 69). This study's research conclusions are directly generalizable to all undergraduate students at NCSU and, to a lesser extent, will inform researchers on how all college students describe their past behaviors. The survey population and the target population are identical and include nearly 23,000 students from all demographic groups and majors at the university. Specifically, the sample population includes all NCSU students enrolled in degree-granting, baccalaureate level academic programs who were enrolled on the tenth day of classes during the spring of 2016 and who first entered NCSU after the summer of 2006. The sample frame is a list of all baccalaureate students enrolled on the tenth day of classes with coverage expected to approach 100%; limited sources of error are derived from ineligible units and undercoverage. The sample is selected from the sample frame using a simple random sampling without replacement.

Validation data. OIRP provides the researcher with student level data in SAS flat files which describe (1) demographic data (including gender, ethnicity, and age); (2) measures of ability (including standardized test scores, high school GPA, collegiate GPA, and number of courses failed); and (3) external validation data (including collegiate GPA; number of courses failed; Health/Exercise Studies Fitness [HESF] course grades and attendance; number of CRF entries; and CRF-based co-curricular participation). The validation data provides both the actual self-reported and the actual population means and measures of variance. The sample size and respondent count provide enough observations for a 95% confidence level for the target population. Typical response rates for dissertation surveys and suggested required sample size of 1,150 is surpassed fivefold.

Survey instrument. This study uses a customized survey instrument to collect information on self-reported student behaviors with an emphasis on socially desirable/undesirable behavior that can be confirmed with external validation data. The instrument ends with a shortened version of the MCSDS to measure a given respondent's likelihood for self-monitoring behavior. The survey was presented as an attempt to gather unknown information about student grades and behaviors involving the CRF. Respondents were asked to retrieve CRF-based behavior from the prior 7 days and since the beginning of the academic year. Overall, academic performance and CRF-based, academic behaviors are retrieved from the beginning of a student's academic career, which could span from 1 month to 9 years. Because this study informs the researchers on biases that may affect research findings, this study's question design mimics survey questions from prior studies.

Quantitative Methods

This validation study uses Model of the Response Process by Tourangeau (1984, 1987) to evaluate the validity of self-reported behaviors described in a survey administered to a random sample of undergraduate students at NCSU. To assess the accuracy of self-reported college student behavior, the self-reported, student-level survey data is merged with student-level validation data provided by OIRP and DASA. Among all respondents, I attribute the percentage of self-reported variance as being valid variance, systematic invalid variance, and random error. Valid variance is self-reported data associated with actual variance observed in the validation data set. Systematic invalid variance is variance in the self-reported data associated with other measured variables that is not observed in the validation data. Random error is variance in the self-reported data that is not observed in the validation data and not predictable using other measured variables. Correlation analysis describes the validity of self-reported behaviors as descriptors of actual behavior. Self-reported behavior may be more valid for specific questions or subgroups of the target population – high GPA students or participatory students, for example – and the correlations will be compared across groups of students. Identification of systematically invalid variance – specifically from social desirability bias – will be identified by estimating the incidence of over reporting socially desirable behaviors using the frequencies of errors, identifying the magnitude of bias, and establishing a relationship between intentional misreporting and MCSDS score. Nonresponse bias is identified by comparing the actual behaviors or respondents to those of nonrespondents. A more detailed discussion of the methods follows in the methodology section.

Definition of Terms

Campus recreation facility (CRF). The CRF is a college or university managed facility that meets the physically active/sport-based leisure needs of students through an intentional curriculum managed by student affairs professionals. In this study, the CRF refers only to the Carmichael Recreation Complex at NCSU with the Department of Health/Exercise Studies managing academic coursework and University Recreation managing co-curricular programming. The CRF hosts independent recreation activities, programmatic recreation activities, club sports for 44 team and individual sports, group fitness courses, intramural sports for 26 teams and individual sports, and for-credit Heath/Exercise Studies courses. The CRF includes over 350,000 square feet of indoor and outdoor space with 6 fitness centers, 11 multi-purpose sport courts, a 25 yard and 50-meter Olympic-sized pool, 6 fitness studios, indoor climbing wall, racquetball courts, outdoor fields, outdoor basketball and tennis courts.

Campus recreation facility use. Entering the CRF signals participation in one of many programs but always requires students to present identification to a CRF employee who records the date and time by swiping the identification card. These swipes are summed into a large transactional data set used for this study. The purpose for CRF entry is explicitly stated for students participating in formal programs like intramural sports, club sports, aquatics, Outdoor Adventures, fitness classes, Challenge Course, and fitness instruction. Class attendance as a purpose of entry is implied by combining a given student’s time of entry and class schedules. For informal recreation, the purpose of the CRF visit is not known.

Health/Exercise studies (HES) course. HES are for-credit courses offered by the Department of Health Science. Prior to 2014, the courses were labeled Physical Education (PE) and the department was called the Department of Physical Education. Almost all courses are based in the CRF with exceptions including Introduction to Bowling (Western Lanes in Raleigh) and Introduction to Downhill Skiing (Appalachian Ski Trail near Boone, NC). Most face-to-face courses meet entirely inside the CRF, but some – tennis for example – build skills for outdoor activities and often meet adjacent to, but not inside the CRF. Distance education courses track course progress remotely and do not require regular class meetings in the CRF. Students are advised to wear athletic attire to HES class meetings, and prior to class, many students enter the CRF to change clothes in the facility's locker room. Fitness courses (HESF) are 100 level courses that build a fitness baseline and teach wellness. Skill courses are HES 200 and 300 level courses that teach a specific skill.

Physical education requirement: All Bachelors level graduates of NCSU are required to pass one HESF fitness course and one HES skill course. There is no suggested time line for course completion. Students with registered disabilities must complete the Physical Education requirement but can enroll in special sections with accommodations.

CRF-based co-curricular program. These are formal, university sponsored programs led by University Recreation employees. Programs are not for-credit but have curricula intentionally designed by student affairs professionals to emphasize inclusion and wellness. All programs are led by professional or student employees who track and report attendance. In fall 2015, these programs included Aquatics, Challenge Course, Club Sports, Fitness, Intramural Sports, and Outdoor Adventures.

Limitations of Study

This single institution study tests the validity of a narrow segment of self-reported student behaviors. The institution-reported validation data describes academic participation through HESF course skipping, but the requirements of HESF courses are unlike those of academic courses in other disciplines. The institution-reported validation describes social participation using CRF entries and CRF-based co-curricular activities that require more physical activity than most other co-curricular programs on campus. As such, this study will forward the field, but the methodology and data limit the generalizability to other academic disciplines or co-curricular behaviors commonly reported in national surveys. The saliency of using HESF courses as a substitute for other academic subjects is debatable, but HESF courses remain the best available to measure class attendance across large groups at this institution. Further study may use computer login data for courses meeting in computer labs, academic building entry swipes for class meetings in other disciplines, electronic appointments at student services like tutoring, or wireless device connections to networks.

Results come with limited generalizability to other settings because NCSU is a selective institution with students who have an unusually high concentration of Science, Technology, Engineering, and Math (STEM) oriented majors and high rates of participation in the traditional residential college experience. Further study should expand this methodology to other types of institutions.

Finally, student data provided by the institution is presumed accurate but may describe instances of non-students borrowing student identification cards to enter the CRF at rates that are unknown.

Chapter Summary

Chapter 1 provided the purpose of the study, definitions to commonly used term, research questions, and research methods and briefly explored the existing literature, theoretical context, and limitations to potential findings. Chapter 2 is a detailed exploration of the existing literature on Tourneau's (1984, 1987) Model of Response Process, survey methods, social desirability bias, and validation studies. Chapter 3 provides an explanation of the research data and research methods and summarizes the available data for study. Chapter 4 analyzes the data to test hypotheses related to answering the research questions. Chapter 5 discusses the significance of the findings and integrates the findings into a discussion of conclusions and limitations in seeking the answer(s) to research questions.

CHAPTER 2: LITERATURE REVIEW

Chapter 2 reviews the literature relevant to the proposed validation study. First, I review the Model of Response Process by Tourangeau (1984, 1987) by exploring the four stage theoretical framework which processes how a survey respondent formulates and reports their past behaviors. Next, I review past validity studies that explore the accuracy of self-reported behaviors related to higher education and physical activity levels. I then explore the concept of social desirability bias and identify past validation studies that suggest that exercise frequency, grade estimation, and class skipping are susceptible to social desirability bias. Finally, I review nonresponse bias and studies that compare behaviors of respondents and nonrespondents. This literature review discusses how researchers assess the validity of self-reported behavioral data. For the purpose of this study, survey validity is defined as evaluating the information reported by respondents as a representation of actual behavior across the target population.

Theoretical Construct

This is a validation study which tests the validity of self-reported college student behaviors using a survey instrument and questions that are typical of higher education surveys. This study focuses on the ability of students to recall and report behaviors with an emphasis on social desirability bias and how the behavior of respondents generalize. I validate self-reported data by creating a survey instrument that requests self-reported data for factual questions whose actual values are known to the researcher. Completion of the survey instrument requires respondents to recall and report academic performance, class skipping, and detailed CRF use over their student careers. Retrieval and recall ability are tested for many time points, including prior 7 days, since the end of prior semester (3 months), since the beginning of the academic year (7 months), and further into the past. CRF use is further broken down into topic areas. Some topics, like informal recreation, are commonplace and club sports seasons are more noteworthy. Some other areas may be considered demanding (and often unpleasant) like NCSU's required two course HES sequence that includes an infamously intense fitness class. All these topics are commonly included in surveys about higher education, so my survey instrument's questions are modeled on questions found in commonly cited publications on student behavior. What follows is a review of relevant literature.

In this study, I assess the validity of self-reported college student behavioral data collected from a customized survey instrument within the theoretical construct of the Model of Response Process (Tourangeau, 1984, 1987). The four-stage model of survey response is used to explain the cognitive processes of survey respondents and identify the origin of misreported student behaviors. The four stages are used by survey respondents when answering factual questions, though not every respondent uses each stage in their self-reporting. First, the respondents must comprehend the survey question through a common understanding of terms used and the information requested by the survey researcher. For questions that describe factually-based behavior (as opposed to opinions), the information is retrieved from memory. Because memories are rarely complete and accurate, retrieved information is combined or supplemented with other information to create a judgment-based question response. Finally, respondents must map their responses to the response options provided on the survey instrument; respondents sometimes edit their true response to be compliant with societal norms that define socially accepted behavior. Each stage possesses threats to self-reported behavior validity that subsequently threaten the validity of findings bias on self-reported data. This study focuses on the retrieval, judgment, and social desirability-based reporting portions.

Tourangeau et al.'s (1984, 1987) model is a widely accepted tool for understanding survey responses, but survey researchers do not necessarily understand how and why some respondents provide more valid data than others. Respondents rarely differentiate between components to organize their response. Respondents do not necessarily progress linearly through the stages and may backtrack to earlier stages and restart their progression. For example, an estimation strategy may trigger a previously unrecalled memory, which begins the response process anew. There is no clear delineation between stages, and respondents may simultaneously exhibit behaviors consistent with different stages as they use their comprehension of a question to influence their judgment of what should be recalled from memory. Respondents are not assumed to answer question to the best of their ability and may demonstrate satisficing behaviors in any or all stages of response (Kronsnick & Alwin, 1987). Similarly, respondents are not equally affected by in-stage factors; for example, internal and external factors make some respondents more willing to self-report undesirable and/or illegal behaviors. Other respondents are better able to retrieve memories accurately because of superior memory recall. The model describes how respondents process their answers, but rarely does the survey researcher understand exactly in which point of the survey response process a given respondent failed to provide valid responses.

Table 2.1.

The Model of Response Process

Component	Specific Process
Comprehension	Attend to questions and instructions Represent logical form of question Identify question focus (information sought) Link key terms to relevant concepts
Retrieval	Generate retrieval strategy and cues Retrieve specific, generic memories Fill in missing details
Judgment	Assess completeness and relevance of memories Draw inferences based on accessibility Integrate material retrieved Make estimate based on partial retrieval
Response	Map judgment onto response category Edit response

Comprehension

"Comprehension encompasses such processes as attending to the question and accompanying instructions, assigning a meaning to the surface form of the question, and inferring the question's point – that is, identifying the information sought" (Tourangeau et al., 2000, p. 9). Higher education researchers create survey instruments to collect self-reported data for program assessment, especially for specialized program areas not covered by national surveys, including campus recreation or physical fitness. Survey instruments clearly describe some behaviors well, including how many times they lift weights per week or check out sports equipment. But some campus recreation behaviors – intramural basketball tournaments, Outdoor Adventure backpacking trips through the Andes Mountains, club sports travelling to out-of-state triathlons, etc. – describe complex behaviors and demand too much specificity for Likert response options intended for broad audiences. These are relevant examples of common comprehension issues like vague question wording and/or use of terms without a common definition.

The first threat to validity from the comprehension stage of response stem from vaguely worded questions. Respondents cannot validly describe their behaviors if survey questions are understood in ways inconsistent with the researcher's intent. Campus recreation surveys also attempt to assess familiarity with preoperational activities, opinions about niche equipment, and baseline fitness to influence curricular design but are often surveying novices unfamiliar with the professional terms. Researchers and respondents have different definitions of what a university owned facility is or how the concept of being physically fit is conflated with participating in campus recreation activities. For example, students and researchers may not mutually understand what the CRF facility describes in the case of a large, on campus field commonly used for Ultimate Frisbee games in no way affiliated with campus recreation. Similarly, respondents may believe researchers care about measuring physical activity and fail to report a CRF visit for static activities like drinks at the CRF cafe or checking out recreational equipment despite the researcher's desire to capture all CRF-based interactions.

The second threat to validity stems from terminologies that all parties do not commonly understand. Researchers are subject matter experts with specialized terminology. Yet they often collect information from populations with diverse vocabularies, so the use of specialized terms often leads to invalid responses. This is particularly problematic for novices who attend CRF instructional programs specifically to gain such operational knowledge but who have not yet done so. Using technical terms to describe techniques or equipment leads to invalid responses. Many institutions have significant international populations where recreation terms may have different meanings. For example, football describes an entirely different activity in the United States than in other nations. Even for those completing the survey in their native language, some campus recreation terms may be confusing. For example, disc golf is a relatively unknown, emerging sport that conjures images of pitching wedges among the unfamiliar.

Retrieval

"The retrieval component involves recalling relevant information from long-term memory. This component encompasses such processes as adopting a retrieval strategy, generating specific retrieval cues to trigger recall, recollecting individual memories, and filling in partial memories through inference" (Tourangeau, et al., 2000, p. 9). The traditional residential higher education process is complex and spans many years. The volume of higher education memories is immense. Because few students accurately recall events from the past, survey researchers invite validity threats when they ask students to recall daily events from years ago. This issue is compounded because cyclical academic schedules invite students to misreport events from one academic year as occurring in a different year.

Compared to benchmark events in a student career, daily events are recalled with less accuracy, yet campus recreation surveys often identify daily users to understand how their early experiences formed later habits. Imagine a fourth year student passionate about basketball; this person played basketball on teams in high school, intramural tournaments, on campus leagues, off campus leagues, informal pickup games (on and off campus), for-credit classes, formal instructional camps, and even attended intercollegiate games as a spectator. This student may accurately report significant events like an injury or a league championship game and may accurately describe themselves as an above average CRF user. But requesting them to accurately self-report common basketball occurrences like weekly participation rates from their freshman year or when they developed specific skills is unrealistic. This student is unlikely to differentiate from past games played inside the CRF versus outdoor courts or what percentages of games were played with friends they first met at the CRF versus preexisting friends. It is more unlikely that this student could recall past events that caused an interest in high levels of engagement with CRFs formal basketball programs (a common question for program evaluators). Commonly used surveys that ask such questions are inviting validity issues. Adding to the issues, national surveys do not ask campus recreation questions, so these researchers create customized questions relevant to their studies but without the extensive validation process.

Judgment [For Factual Questions]

"Retrieval often does not yield an explicit answer to survey questions" so the "judgment component comprises the processes that respondents use to combine or supplement what they have retrieved [from memory]" (Tourangeau, et al., 2000, p. 10). The judgment response process includes evaluation of retrieved memories for accuracy and how the respondent then uses retrieved memories to impute missing information for the creation of conclusions on aggregate behaviors. The cognitive load of questions varies and, for simple questions, most respondents can answer accurately by relying on their memories. For more complicated questions – how often they visit the CRF, for example – respondents must use their personal judgment to estimate the behaviors requested.

Respondents rely on several estimation strategies to impute missing information (Groves et al., 2009). First, respondents recall past experiences from memory and then total up occurrences of the behavior, sometimes with an upward adjustment for memories assumed to be overlooked. The second strategy is rate-based estimation where the respondent estimates the rate of occurrences over a finite time and then extrapolates the rate of incidence over the reference period. For example, a respondent reports their total CRF visits in a semester by estimating that they visit the CRF thrice per week and then multiplying the rate by the number of weeks in a semester. The final strategy is impression based estimation where respondents use a vague impression of their incidence of a behavior and then match it to permissible survey responses. Most respondents integrate the product of all retrieved memories and estimation strategies to formulate their descriptions of past behaviors. Prior research suggests high ability students (as measured by GPA and standardized test scores) recall past behaviors with more accuracy (Dobbins, Farh, & Werbel, 1993; Kuncel et al., 2005; Porter & Umbach, 2006).

When answering questions about CRF use, most students use estimation strategies to formulate their responses. It may be easy to recall and count CRF visits for users who rarely or never visit the facility, but because NCSU requires two HES courses for a bachelor's degree, most persisting undergraduates make multiple weekly CRF visits at some point in their career. The recall and count strategy for commonly occurring behaviors requires substantial cognitive ability (Tourangeau et al, 2000) so, with few exceptions, the students sampled will rely on estimation strategies. Prior research on CRF use focuses on the high volume users or the impact of use on long term outcomes; such requests increase the cognitive load of recalling and emphasize the importance of estimation strategies (Hall, 2006; Lindsey & Sessoms, 2006; Rothwell & Theodore, 2006; Todd, Czyszcson, Wallace Carr, & Pratt, 2009;). A student may meet friends to play basketball every Tuesday and Thursday and use a rate based estimation strategy that multiplies twice per week by the number of weeks in a semester. This strategy over estimates use when the respondent fails to consider missed meetings for holidays, illness, inclimate weather, final exams, or other unusual events. Another estimation strategy is to associate the amount of time they play basketball with a particular life circumstance. For example, on campus residents visit the CRF at higher rates than commuter students; since most students moved from campus housing to commute as upperclassmen, their impression based estimation strategy may be biased by their current life circumstance.

Reporting and Response Selection

Reporting and response selection “involves selecting and reporting an answer…mapping the answer onto the appropriate scale or response option and ‘editing’ the response for consistency, acceptability, or other criteria…it may not be clear to them how to report it” (Tourangeau, et al., 2000, p. 13). For discussions within this study, the literature establishes validity concerns stemming both from respondents’ interpretations of reporting options and from the social acceptability of a behavior affecting the likelihood of self-reporting an honest response (Groves et al, 2009). Most standardized and national surveys research their question validity and provide clearly understood response options. Meanwhile, localized surveys focusing on specific behaviors tend to be less refined and have an increased likelihood of confusing respondents or offering contradictory response options (Porter et al., 2011). Most CRF surveys, which tend to be locally created, ask students to report their answers in a multiple choice format that offer responses on a scale or as discreet categories. It is not trivial to assume that survey response options provide researchers the data they intend to collect.

Prior Research

Higher education researchers and stakeholders seek a better understanding of college student development to assess curricular effectiveness and enact reforms to improve student outcomes. Researchers evaluate benefits of curricular interactions using student-level data which describes demographics, family of origin, pre-collegiate experiences, academic performance, collegiate classroom experiences, co-curricular experiences, collegiate behaviors, and time use (Bowman, 2010). Joining researchers in study are stakeholders whose ever-expanding expectations of transparency and accountability place pressure on researchers to collect higher quality data about college student behaviors and outcomes (Bowman & Schuldt, 2014). Institutional data about student background, demographics, and behaviors apply to large numbers of students but generally provide limited information (Gonyea, 2005). Standardized tests accurately describe large groups of students but are limited to narrowly defined measures of ability that are too specific to adequately explain the entirety of student ability; for example, SAT Math score validly predicts mathematical ability, but mathematical ability is too narrow a measure to predict overall college success (Gonyea, 2005). Modern higher education study requires descriptive specificity from factual student data that limit sources to self-reported behaviors obtained directly from students (Astin, 1993; Tourangeau et al., 2000). The growing popularity of survey research stems from the utility of self-reported data as a source for describing formative and summative outcomes (Ewell and Jones, 1993). Researcher reliance on self-reported data using college student surveys is increasingly important (Porter, 2011), which makes the validity of self-reported student behaviors a compelling research topic.

The study of higher education is highly dependent on accurate self-reported, objective data to formulate research conclusions on college student learning (Bowman, 2010). Many researchers agree that self-reported, factual data is indispensable and presume that it is trustworthy when the survey instrument and implementation are designed for scholarship, whether locally constructed or a national survey (Gonyea, 2005). Instead, Porter (2011) believes researchers should place the onus on proving the validity of self-reported behavior. The preponderance of evidence suggests that students are poor self-reporters of actual behavior and learning and that self-reported data should be presumed invalid until proven otherwise (Bowman, 2014; Gonyea & Miller, 2011). To date, the literature describes a need for dedicated manpower to assess the validity of self-reported data (Bowman & Herzog, 2011; Pike, 1999b; Umbach, 2005), yet the validity of self-reported college student data has received modest attention (Gonyea, 2005). Instead, researchers continue to increase the number of surveys conducted as survey tools become less expensive and more accessible (Pike, 2011; Porter, 2011). Advances in electronic communications continue to expand researchers' ability to implement more surveys with little emphasis on validity issues (Umbach, 2005). This study forwards the field by creating a customized survey instrument that describes behaviors known via institution-reported data sets, including CRF use, class skipping, and co-curricular participation. Compared to prior validation studies, these behaviors are more similar to those collected by researchers using national and local surveys.

Survey Life Cycle

The survey life cycle (Groves et al., 2009) is a useful frame to enable discussion about the phases of survey implantation and how error can affect the survey statistic in each phase. The survey life cycle is divided into the measurement process and the sampling process. The survey life cycle organizes the survey statistic creation with discreet opportunities for error.

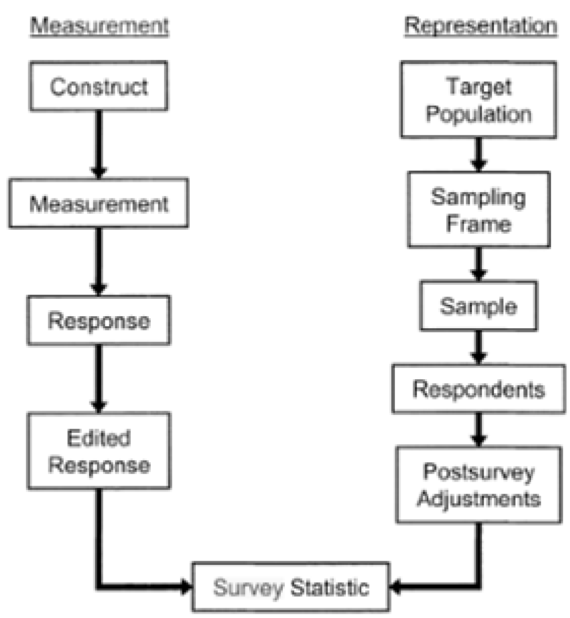

Figure 2.1. The survey life cycle.

Measurement Process of the Survey Life Cycle

The measurement process describes the evolution from the construct the survey researcher wishes to measure to the values derived from the actual responses provided. For example, a researcher may set out to measure the demand for more open swim time in the CRF pool with the intent of conducting a needs assessment for the university aquatic facilities. The researcher develops measures for open swim demand by constructing questions about satisfaction with the current pool policies and how they would use the pool without scheduling constraints. The responses are information provided by those completing the survey instrument. The survey researcher compiles the responses and edits some responses by removing outliers or inappropriate answers that are clearly not representative of typical users. The edited responses are the primary input for the survey statistic.

At each phase of the measurement process there exist opportunities to introduce error. Error in the measurement process is defined as deviations in the answers provided by respondents from the construct the survey researcher intended to measure.

Measurement Error. Measurement error occurs when survey question answers do not collect responses that accurately respond to the construct intended. Measurement error is commonly caused by poor question wording which results in survey researchers and respondents lacking a common understanding of the data collected. Irrelevant questions lead respondents to confuse the survey's intent and should be avoided. Survey questions should be carefully worded and tested to assure that the typical respondent understands text as intended. It is recommended that survey questions be piloted from members of the survey population through pretests or cognitive interviews.

Processing Error. Processing error occurs after collection and during the data cleaning and coding. For closed ended questions, process error is often introduced by inaccurate scales or inaccurate data entry. For open-ended questions, process error is introduced through improper coding or failure to accurately clean outlier responses. In the pool use example, an open-ended question about past use could be intended to collect the numeric representation of annual pool visits, but respondents could enter open text to describe their pool use. The survey researcher may improperly convert the text into an actionable numeric value.

Representation Process of the Survey Life Cycle

The representation process of the survey life cycle describes how a subset of an overall population can be representative of an entire population without surveying all members of the population. The target population includes all members of a population which the survey statistic is intended to describe. In the pool demand example, the target population includes all the institution's enrolled students eligible to use the CRF. The sample frame includes all members of the target population for which there is contact information (in higher education surveys, the sample frame and target population are often identical). The sample is a subset of the sample frame created strategically to best represent the target population. Simple random samples – or sample frames where each unit has an identical, non-zero probability of selection – are used to select the sample from the sample frame. This method is consistent with most existing studies about student behavior, though it is not uncommon for a sample to assign a higher probability of selection to selected demographics. In my example, known pool users may be surveyed at higher rates than nonusers because their opinion on pool hours warrants more influence than nonusers. Respondents are members of the sample who completed the survey. Researchers may make post-survey adjustments by weighting known values or imputing missing values to assure that members of the target population are adequately represented. The adjusted sample is the second input to the survey statistic.

Coverage Error. Coverage error occurs when members of the target population are not included in the sample frame, usually because the contact information is not correct. Survey researchers should be particularly concerned when the lack of contact information is correlated with other attributes in the target population. Among higher education survey research, coverage error is a concern when the institution contact information (usually email addresses) is not used by the student.

Sample Error. Sampling error occurs when the opinions or behaviors of the selected sample do not match that of the sample frame. Sample error exists in all cases where the sample frame and sample are not identical. Larger samples and strategic sample creation (i.e. stratified random sampling) reduce magnitude of sample error. Sampling error can be reduced though the use of confidence intervals.

Nonresponse Error. Nonresponse describes a correlation between the likelihood of responding to a survey and systematic differences in the characteristics of responders and nonresponders. In our pool use example, frequent pool users may be more likely to respond to a pool use survey because they are invested in the pool hours and policies. Meanwhile, students who do not visit the pool fail to respond to the survey because they are less concerned. The nonresponse error describes the underrepresented data from non-pool users. Nonresponse error can describe as abstention from the entire survey (unit nonresponse) or abstention from specific questions (item nonresponse). Nonresponse error is more likely in cases where the respondent characteristics are dissimilar to the target population.

Adjustment Error. Survey researchers may weight or edit the responses provided to compensate for detected sources of coverage, sampling, or nonresponse error. These adjustments are meant to improve the validity of the sample mean as a representation of the target population. These adjustments may decrease the validity and introduce new biases, especially in cases where missing data is not identified and properly imputed or excluded from the analysis.

Response Bias and Measurement Error in Measurement Process

Factual, self-reported data is data the researcher obtains directly from the participant rather than from another source that describes objective behavior and characteristics (Tourangeau et al., 2000). These self-reports are compiled into summary statistics used by quantitative researchers. Quantitative researchers use the variance across all the responses to a question, which means that validity studies evaluate the survey statistic as a valid descriptor of actual behavior rather than focus on an individual's response. For self-reported, factual data, the variance within the survey statistics is categorized as (1) valid variance associated with actual differences among respondent behavior; (2) systematic invalid variance explained by respondent, survey, or process factors but not descriptive of actual behaviors; and (3) random error that does not describe actual behavior and is not correlated with any other factor (Kuncel et al., 2005). Researchers who make research conclusions without verifying survey data validity rely on the assumption that variance from self-reported data is predominantly valid variance and void of systematic invalid variance.

Random error is ever present in most quantitative data measurements but is not correlated with any other measured variables. People are inaccurate self-reporters and introduce random error when making various judgments about themselves (Bowman, 2010). Unlike systematically invalid variance, random error reduces the power of self-reported data but presents a far lesser threat to research findings because randomly over reported rates are usually offset by randomly underreported rates within the survey statistic (Groves et al., 2009). Errors in self-reported data increase with the complexity of the data requested (Gonyea, 2005). Random error is ever-present because respondents inaccurately recall past behavior, but it is mitigated through simple recall requests and well-designed survey questions (Cole & Gonyea, 2010; Gonyea, 2005).

Systematically invalid variance has many causes stemming from the survey process and the respondents. Examples of systematically invalid variance caused by the respondent include an inability of respondents to accurately report due to impaired ability to recall events (Kuncel et al., 2005); a Halo Effect error in which specific factor responses are correlated with both the overall experience and factor requested (Bowman, 2014); or intentional misreporting in which respondents deliberately provide false information (Tourangeau et al., 2000). Another cause is satisficing, or respondents failing to expend sufficient effort to understand questions completely and/or recall enough material to provide a fully descriptive answer (Groves, et al., 2009). Examples of systematically invalid variance caused by the survey process include order effects, which are defined as survey question order affecting answers provided because the information that is used to process a prior question alters information provided in a subsequent response (Bowman & Schuldt, 2014). Furthermore, survey response options may cause respondents to misreport if the options do not provide sufficient granularity for respondents to accurately express answers (Tourangeau et al., 2000). For a final example, if the researcher and respondent commonly understand a research question, an accurate self-report is valid; but if the question is understood differently, then accurate description of behavior does not validly measure the researcher's construct (Porter, 2011). The focus of this study is locating systematic invalid variance that results from the survey process and/or intentional misreporting motivated by respondent factors.

Respondents may intentionally misreport to protect themselves from perceived repercussion (Tourangeau et al., 2000). Intentional misreporting is observed through skipping a question (item nonresponse), failing to complete a survey (unit nonresponse), or responding with a false answer because item nonresponse raises the suspicion of researchers (Tourangeau & Yan, 2007). Efforts to detect intentional misreporting, use validation data, or to imply consequence for misreporting are commonly referred to as validation studies. Prior validation studies provide consistent evidence that respondents intentionally misreport to inflate image and to avoid repercussions from truthful answers to sensitive questions (Dobbins et al., 1993; Kreuter, Presser, & Tourangeau, 2008; Tourangeau & Yan, 2007). Though ever present, self-reporting error is not an impediment when researchers remain vigilant to minimize systematically invalid variance and random error (Umbach, 2005).

Researchers frequently use survey methods to collect data on college student socializing, use of free time, class attendance, study habits, and other behaviors that are not well tracked. However, respondents often view surveys on these topics as requests for sensitive information, thus making their responses vulnerable to motivated misreporting (Bowman & Hill, 2011; Bowman & Schuldt, 2014; Porter, 2011). Because objective data is difficult to obtain, researchers continue to rely on self-reports despite the preponderance of evidence suggesting that students cannot accurately self-report this information (Bowman, 2014; Bowman & Hill, 2011; Gonyea & Miller, 2011; Porter, 2011). To point, Bowman (2010) estimates that two thirds of educational gains found from self-reported data cannot be replicated when using objective longitudinal data as a result of motivated misreporting which causes systematically invalid variance. Validation studies find that college students can accurately answer clearly constructed questions about recent and/or landmark events (Kuh, 2004; Garry, Sharman, Feldman, Marlatt, & Loftus, 2002; Sifert & Asel, 2011), but the absence of validation data hampers the study of validity for self-reported day-to-day, intrapersonal and interpersonal events (Bowman, 2011; Porter et al., 2011). Prior validation studies use data collected by Admissions' or Registrars' offices to test respondent ability to accurately self-report grades; correlations between self-reported and actual data range from loose (0.45) to strong (0.98) (Cole & Gonyea, 2010; Herzog, 2011; Kuncel et al., 2005).

Satisficing

Satisficing is an alternative explanation for why respondents fail to provide accurate descriptions of behavior (Barge & Gehlbach, 2012). Groves, et al. (2009) explain that "Satisficing respondents do not seek to understand the question completely, but just well enough to provide a plausible answer; they do not try to recall everything that is relevant, but just enough material on which to base an answer; and so on" (p. 224). In other words, providing a quality response requires more energy than some respondents are willing to expend on a survey. "Respondent may simply be less thorough in comprehension, retrieval, judgment, and response selection. They may be less thoughtful about a question's meaning; they may search their memories less comprehensively; they may integrate retrieved information carelessly; and they may select a response imprecisely" (Kronsnick, 1999, p. 547-8). While researchers may attribute invalid responses to an inability to recall accurately, responders may, instead, be unmotivated.

To provide high quality, actionable survey data, there must be relatively few errors in the survey process (Biemer & Lyber, 2003). Most surveys, including national surveys like NSSE, do not provide measures for response quality (i.e. satisficing), leaving interpretation vague though satisficing symptoms are observable through low response rates, missing data from skipped questions, and the introduction of nonresponse bias (Chen, 2011). According to Kronsnick (1999), satisficing respondents may select the first reasonable response option rather the best one, or they may describe the first viable memory rather the most relevant, and some respondents may randomly select an option without consideration. To prove the point, Kronsick (1999) described a -0.22 unit level correlation between the closed ended questions "I enjoy socializing" and "I don't enjoy socializing" (which should approach -1.00). Satisficing respondents may provide neutral responses inoffensive to the status quo by responding with a "don't know" rather than risk offense with negative descriptions.

Kronsnick (1999) theorizes that the social norm to be polite causes satisficing because potential respondents find it easier to cheerfully comply with a survey request they intend to satisfice rather than cause interpersonal tension by objecting. Other respondents comply in good faith but satisfice as the survey turns burdensome (Kronsnick, 1999). Identifying satisficers is problematic, but Chen (2011) suggests that a survey include nearly identical questions on inversed scales to find potential statistics that fail to provide consistent answers. Frequent item nonresponse is also symptomatic of satisficing (Chen, 2011). Kronsnick (1999) recommends counting the number of words provided in open-ended questions as a measure of response effort.

Within the context of this study, two types of satisficers may be observed. First, are respondents who provide minimal effort and give poor quality answers without any systematic method. These sastisficers provide inaccurate answers that are not consistently biased unless the question design causes commonly occurring behavior like respondents choosing the first viable response option. The second satisficer types are motivated by describing their generalized identity rather than taking the time to recall and report behaviors as they occurred. For example, habitual exercisers may over report to describe themselves as a high volume exerciser rather than expend the effort to recall details of their recent activity. In a parallel example, Abelson, Loftus, and Greenwald (1992) found that satisficing habitual voters describe themselves as voting in the most recent election (even if they did not) because it is easier to rely on their assumption that they voted than to recall the specifics of the last election.

Social Desirability Bias

Social desirability bias results from systematic and intentional misreporting, which results from the respondents' tendency to inaccurately describe their behavior to present themselves more favorably. A subset of sensitive questions, socially desirable questions ask about intrusive topics that may inadvertently motivate respondents to misreport in an effort to present them in a more favorable light or protect egos (Gonyea, 2005; Tourangeau & Yan, 2007). The theory of social desirability bias presumes that there are social norms that define desirable behavior, and many respondents answer questions to present themselves as complying with social norms (Kreuter et al., 2008). Social desirability bias concerns researchers when intentional misreporting causes the survey statistic to become an inaccurate descriptor of actual behavior (Groves et al., 2009). Respondents tend to over report socially desirable behaviors like seat belt use and exercising and underreport socially undesirable behaviors like smoking and illicit drug use (Tourangeau & Yan, 2007). Social desirability bias is a function of the respondents' characteristics, the survey's mode of delivery, and survey question design (Dobbins et al., 1993).

Respondent characteristics that affect the likelihood of intentionally misreporting behavior in a socially desirable direction are tied to a given respondent's propensity for self-monitoring. Self-monitoring is the ability and willingness for an individual to read situational social cues and then provide an unauthentic response that conforms to the expectations communicated in the social cues (Snyder, 1987). High self-monitors (HSMs) "excel at reading social cues and at regulating their self-presentation to fit the expectations of others" and are more likely to alter self-reported behaviors consistent with preferred social norms (Dobbins et al., 1993, p. 322). HSMs are considered good actors who rely more on situational factors to guide their outward behaviors and less upon their actual feelings and attitudes. Low self-monitors (LSMs) lack either the motivation or ability to alter their behavior and/or are less able to accurately assess social cues.

Situational changes to self-monitoring by exaggerating incidence of socially desirable behaviors in some, but not all settings is a hallmark of HSMs. HSMs will pick and choose the occasions for which they inflate or alter their emotional response, and LSMs present themselves consistently across all situations (Dobbins et al., 1993). HSMs intensify their authentic behaviors, conceal inappropriate behaviors without a false representation, or falsely represent their actual response (Snyder, 1987). On a survey, HSMs select a socially correct response option even if it is not an accurate description of truth (Mueller-Hanson, Heggestad, & Thorton, 2003). For example, an HSM may over report exercising because exercise is socially desirable, or they may underreport class skipping because it is a socially undesirable behavior. In a subsequent section, I will discuss the MCSDS which is an accepted measure of propensity for self-monitoring (Crowne & Marlowe, 1960). Identifying HSM respondents may not improve the construct validity of a survey because remedies are not apparent (Ellingson, Sackett, & Hough, 1999).

The second influencing factor of social desirability bias is the mode of survey delivery which affects the frequency of socially desirable misreporting by altering the respondents' perceived consequence for accurately reporting undesirable behavior. On average, socially undesirable behaviors are reported at higher rates using modes that limit human interaction (e.g., web vs. Computer Assisted Telephone Interview); after controlling for human interactions, respondents self-report socially undesirable behaviors at similar rates across all modes (Kreuter et al., 2008). Theoretically, interactions with a human interviewer increase the likelihood that respondents will feel threatened or embarrassed by accurate descriptions of socially undesirable behavior (Tourangeau et al., 2000). The respondents' perceived threat is greater if the interviewer and respondent continue to interact after the survey is complete (Tourangeau & Yan, 2007).

The final factor of social desirability bias is survey question design. A socially desirable question alters how students self-report such that errors trend in a positive direction (Porter, 2011). Questions viewed as intrusive, including taboo topics, topics normally avoided in everyday conversations, are susceptible to social desirability bias; as a general rule, they involve topics that the government would not normally ask (Tourangeau & Yan, 2007). Schmitt et al. (2003) theorize that the instrument reduces misreporting when it includes follow up questions that require respondents to elaborate on previously reported behaviors. Other studies find that respondents more accurately self-report behaviors when they believe inaccurate reports are detectable through lie detection (Tourangeau et al., 1997) or by known validation data (Schmitt & Kuncel, 2002).

Social desirability bias threatens research findings if the behavior described by the survey statistic differs substantively from the respondents' actual behavior. "Social desirability bias can compromise researchers in one of two ways: (1) over reporting of socially desirable behavior (such as performing community service) and underreporting of socially undesirable behavior (such as smoking), and (2) attenuation, inflation, or moderation of relationships between variables" (Gonyea, 2005, p. 82). The impact of social desirability bias in college student behavioral studies is unknown because validation data sets are difficult to obtain. Examples of social desirability bias observed in college student studies include inflation of self-reported grades (Dobbins et al., 1993; Gonyea & Miller, 2011; Zimmerman, Caldwell, & Bernat, 2002), standardized test scores (Kuncel et al., 2005), academic engagement (Bowman & Hill, 2011; Schmitt et al., 2003), and abstinence from vice (Kreuter et al., 2008; Tourangeau et al., 1997). Student development researchers estimate the effect of a behavior on outcomes using self-reports; research findings may be misleading if socially desirable bias causes misreporting of a behavior determined to improve an outcome (Bowman, 2011; Bowman & Hill, 2011; Tourangeau & Yan 2007). Social desirability bias is more likely to systematically affect college student subgroups with less desirable traits. For example, low GPA students are susceptible because high GPAs are desirable, and the upper limit on the GPA scale only allows low GPA students to significantly inflate their self-reported GPA (Dobbins et al., 1993)

Marlowe Crown Social Desirability Scale

The MCSDS is a generally accepted tool used to detect propensity for self-monitoring (Ballard, 1992; Gonyea, 2005; Leite & Beretvas, 2003; Snyder, 1987). Respondents complete the MCSDS by answering 33 true-false questions that describe socially desirable behavior that is untrue of most people or socially undesirable behavior that is commonplace; the likelihood to self-monitor is scored by totaling the presumed untrue socially desirable responses provided (Crowne & Marlowe, 1960). For example, "My table manners at home are as good as when I eat out in a restaurant" (question 8) is unlikely to be true for the majority of people, so respondents marking true are more likely HSMs. Conversely, "I am sometimes irritated by people who ask favors of me" (question 30) is likely true for most people, so respondents marking false are more likely HSMs. Researchers usually co-administered the MCSDS with their survey instrument and then tested for correlations between the MCSDS results and self-reported scores on socially desirable questions.

Statistically significant correlations indicate the presence of social desirability bias and suggest that the survey items are not valid (Gonyea, 2005). Dobbins et al. (1993) assume that LSMs provide more accurate data and suggest weighting results by the MCSDS score.

The MCSDS is widely accepted as a leading tool to measure the likelihood of self-monitoring. The MCSDS tests the need for social approval among college students with questions that describe cultural approval items with minimal relationship to psychological abnormalities (Crowne & Marlowe, 1960). The scale measures the link between the need for approval being and providing socially desirable responses on surveys (Synder, 1987), but it does not determine if respondents provide fake answers (Mueller-Hanson et al., 2003). Synder (1987) compared MCSDS scores across groups who are likely HSMs (actors), likely neutral self-monitors (fraternity members), and likely LSMs (psychiatric patients) to determine if these groups' MCSDS scores are correlated with their expected concern for what others think. He further cross validated the MCSDS by asking fraternity members about themselves and their fraternity brothers to find that low MCSDS scorers were less likely to alter their behavior when social norms dictate a certain response. Snyder (1987) found that the MCSDS was "internally consistent, temporarily stable, and uncorrelated with self-report measures of related concepts" (p. 536).

The MCSDS faces some criticism as an indicator of social desirability bias. Tourangeau and Yan (2007) believe that the scale is unable to distinguish between respondents who are actually compliant with social norms, respondents who have sincere but inflated views of themselves, and those who are HSMs. Further, Mueller-Hanson et al. (2003) believe that the "correct" MCSDS are obvious and that adept self-monitors will provide misleading answers to the self-monitoring scale in addition to other sensitive questions of interest. Moreover, Synder (1987) suggests that the MCSDS measures an individual's need for approval but not his or her ability to hide misrepresentations on surveys. Self-monitoring and social desirability scales – including the MCSDS – are not highly correlated with each other, so there is no evidence that they measure only the likelihood to fake survey responses (Mueller-Hanson et al., 2003). These studies identify a need to test the relationship between the MCSDS score and faking survey responses.

Researchers often condense the MCSDS to a shortened form that reduces the completion time, but retains the majority of the scale's validity. The long form increases the burden on respondents, yet some questions possess limited usefulness for detecting HSMs (Ballard, 1992). Ballard (1992) conducted a principle component analysis to identify 13 factors that account for nearly all of the full scale's variance and shortened the MCSDS to retain one question from each factor (see figure 3.1 for questions). Her shortened MCSDS retains 93% of the full scale's validity with 40% of the questions. Ballard (1992) compared her shortened MCSDS to other shortened forms to find similar results.

Exercise Identity

Brenner and DeLamater (2014) bring identity theory as an alternative to social desirability bias to explain systematic over reporting of physical activity. Identity theory was first presented by Stryker (1980) to describe the phenomena of self-reporting behaviors that are consistent with self-created image rather than a factual description of memory. The respondent is either unable to recall his or her actual behavior or is unwilling to put forth the effort (i.e., satisficing) and instead falls back upon identity-based estimation rather than enumeration (Brenner & DeLamater, 2014). In the exercise identity case, respondents are motivated to over report physical activity because of their self-image rather than the social expectations of the interviewer (Brenner & DeLamater, 2014). This approach moves responsibility for the misreport from an attempt to impress the interviewer to the respondent's internal conversation about their identity that is not affected by survey administration methodology (Brenner & DeLamater, 2014). In the context of this study, habitual exercisers over report physical activity because they default to a response consistent with their identity rather than recalling actual activity.

Brenner and DeLamater (2014) tested the exercise identity theory by comparing self-reports to a validation data set of university reported CRF entries. Fifty of 156 respondents over reported their level of physical activity in the most recent week, only 11 of the 156 under reported physical activity, and the remaining 95 respondents accurately described their physical activity. Alongside questions about their physical activity in prior weeks, Brenner and DeLamater (2014) administered the Mann-Whitney-Wilcoxon test for exercise identity. Of the 89 respondents who tested positively for an exercise identity, 43 over reported physical activity versus only 4 who underreported. Meanwhile, only 7 of the non-exercise identity respondents over reported their physical activity, and a stochastically offsetting 7 underreported. Exercise identity respondents who did not visit the CRF in the reference period were the most likely to over report physical activity. Exercise identity theorizes that misreporters are lying to themselves rather than conforming with the socially desirability bias theory where the respondent is lying to the interviewer.

Validation Studies

An absence of validation data in higher education leaves an unmet need for validation studies on college student time use and self-reported behaviors. Researchers evaluate the validity of self-reported data by comparing self-reports to objective data sets with verified values. “Validity studies are a critically important element in higher education research. These studies help ensure that higher education researchers are collecting accurate and appropriate information that can be used to understand important questions, such as how college affects students” (Pike, 2011, p. 54). Common validation studies in higher education include grades and standardized test scores which compare self-reported data to admissions’ and registrars’ records. Known validation studies provide ample evidence of motivated misreporting among college students with almost all misreporting errors in the socially desirable direction (Tourangeau & Yan, 2007). Validation studies find that students are better able to recall events that are more recent, current, and relevant (Klesges, Eck, Mellon, Fulliton, Somes, & Hanson, 1990; Seifert & Asel, 2011) and behaviors rather than outcomes (Bowman, 2011). Researchers clearly define a need for validation studies, but the scarcity of validation data sets prohibits validation studies on additional topics.

The preferred method for validation studies is to match self-reported data with a corresponding validation data set and then identify the frequency, direction, and magnitude of unit level misreports (Gonyea, 2005). Most validation studies focus on prior academic performance using easily obtained academic data like standardized test scores and grade point averages (Gonyea, 2005). Very few studies have examined the validity of self-reported behavioral data because validation data sets for generalizable college student populations are rare (Gonyea, 2005). In cases where self-reported data is available, only basic respondent characteristics – gender, ethnicity, academic ability, and academic outcomes – are tested for correlations with accurate self-reporting (Kuncel et al., 2005). Another need in higher education research is to identify variables that moderate the relationship between reporting accuracy and respondent characteristics that help researchers distinguish between random error and systemic errors in self-reporting (Kuncel, et al., 2005). Of specific interest are self-reports of socially undesirable behaviors which are susceptible to social desirability bias (Krueter, et al., 2008).

Porter et al. (2011) attempted a study to validate NSSE questions about course syllabi and faced data limitations that typify validation studies. Students were asked to report on the number of reading and writing assignments required in their classes for a given semester. The researchers collected and compiled nearly all their course syllabi to establish the actual answer for validation against self-reported answers. However, the complete course requirements were only available for 42 seniors. The correlation between the self-reported books read and the syllabi validation data set was only 0.38 with 21% of the students correctly identifying the number of readings. The researchers found it difficult to align the NSSE assignment definitions with syllabi information; the lack of commonly understood terms for homework assignments further exemplified one shortcoming of the survey but hampered the validation of self-reported course requirements. The study further suggests that students cannot accurately self-report behaviors while also exemplifying the challenges of obtaining validation data sets.

Self-Reported Physical Activity

Survey questions about illicit drug use, sexual activity, and alcohol abuse are well studied and known to be affected by social desirability bias. In comparison to the aforementioned vices, questions about visiting the gym and missing HESF class meetings appear less sensitive and salient to social desirability theory. However, the only known study that compares social desirability bias of self-reported exercise rates to other undesirable behaviors was conducted by Tourangeau et al. (1997). They found that exercise is among the topics most susceptible to social desirability bias. Using a bogus pipeline method, control and experiment groups were orally administered an identical survey. The experiment group was successfully convinced that the researcher possessed a highly accurate, next generation lie detection machine. This belief increased the incentive to accurately self-report. When asked if they exercise four or more times per week, the control group responded in the affirmative 45% of the time, a much higher rate than the experiment group's 23% (significant at the .01 level). The estimated effect of social desirability bias for the exercise question exceeded questions about taboo sexual acts, drug use, cigarette smoking, drinking and driving, and abortions. Of the 20 intentionally sensitive questions, the exercise question's estimated magnitude of social desirability bias is only eclipsed by a question asking the respondent if they drink more alcohol than the average person. This survey is conducted in person, and findings may not be generalizable to online surveys.

Epidemiologists and other medical researchers robustly study physical activity's effect on health outcomes, but like educational researchers, their research conclusions are dependent upon self-reported data. Validation studies of self-reported physical activity are highly demanded but lightly studied. Existing validation studies suggest self-reported physical activity is systematically over reported and lowly correlated with actual activity, though little is known about respondent motivation to misreport (Brenner & DeLamater, 2014; Sallis & Saelens, 2000; Shephard, 2003). Like this study, epidemiological validation studies compare self-reported, survey data to objective measures of physical activity. Examples of validation data include sign in logs from swim and tennis clubs one year after participation (Chase & Godby, 1983), consuming radioactive isotopes that decompose to measure energy expenditures throughout a week (Adams, Matthews, Ebbeling, Moore, Cunningham, et al., 2005), pedometers and accelerometers that track motion over a day (Shephard, 2003); journals and text message prompts that track very recent physical activity (Brenner & DeLamater, 2014), correlations between leisure time choices and physical activity scales (Motl, et al., 2004), correlations with physical fitness (Shephard, 2003), and even by spying on subjects (Klesges, et al., 1990). With the exception of Klesges et al. (1990) and this study, subjects are aware of their participation in a validation study and are potentially expending atypical effort to track physical activity, thereby altering the quality of their self-reports.

Correlations between survey reported physical activity and validation data range from 0.02 to 0.78 with respondents over reporting physical activity many times more often as underreporting (Durante & Ainsworth, 1996; Sallis & Saelens, 2000; Shephard, 2003). Subjects are generally poor self-reporters though research design affects the correlation; for example, respondents more accurately recall time spent in vigorous, aerobic workouts and/or physical activity from the prior seven days (Sallis & Saelens, 2000). Alternatively, respondents inaccurately report long past physical activity and struggle with complex questions about their activities (Shephard, 2003). Across studies, two thirds to three quarters of respondents over report their level of physical activity (Brenner & DeLamater, 2014; Chase & Godbey, 1983; Durante & Ainsworth, 1996; Klesges et al, 1990) though Adams, et al. (2005) found unbiased self-reports of total calories expended. Estimating the magnitude of response bias is hampered by inconsistent methods, definitions, and survey instruments (Sallis & Saelens, 2000; Shephard, 2003).

Despite somewhat consistent findings about self-reported validity, the field is unsettled in explaining why respondents misreport. Prior research attempts to explain the phenomenon using respondent traits, temporal factors, and social desirability.

Age. Adult respondents and older children are theorized to possess better recall ability than children do. Among youth, older subjects are more likely to accurately self-report their activity levels (Kleges, et al., 2004). Age is also positively associated with self-reported accuracy among adults with the highest correlations (.75 to .78) between self-reporting and actual activity occurring among elderly subjects (Sallis & Saelens, 2000).

Gender. Women are theorized to face more societal pressure to maintain ideal body weight and are theorized to inflate self-reported physical activity to conform with societal expectations. Warnecke, Johnson, Chavez, Sudman, & O’Rourke (1997) found that women are more likely to score as HSMs on the MCSDS and are correspondingly more frequent inflators of physical activity. Other studies find that stereotypical gender roles affect what is considered physical activity and how it is reported. If researchers link house cleaning to physical activity, women are more likely to over report physical activity because women face more social pressure to inflate time spent house cleaning; as such, researchers may introduce gender-based social pressures that cause misreporting (Adams, et al., 2005). Warnecke, et al. (1997) found a contrary trend that women were not likely to include housework in their self-reports though men were likely to inflate self-reports by including yard work. Finally, women scoring high on general lifestyle activity measures are more likely to underreport physical activity compared to men and less active women (Klesges, et al., 1990).

Ethnicity. Because social desirability bias is based upon conformity with societal expectations, theory suggests that individuals from different cultures face physical activity expectations unique to their culture. Warnecke, et al. (1997) found that ethnicity affects the likelihood to edit responses. African American respondents are more likely to include housework as a physical activity than non-Hispanic whites. Additionally, African Americans and Mexican Americans are more likely to be HSMs than Puerto Ricans and non-Hispanic whites.

Body Type. Overweight and obese individuals are theorized to face weight related pressures to present themselves as physically active. Adams et al. (2005) found that women with a BMI of 27 or higher inflated physical activity at higher rates than lower BMI women or men. Klesges, Baranowski, Beech, Cullen, Murray, & Rochon, (2004) also found that the presence of social desirability pressures African American girls at risk for obesity to over report physical activity. Contrarily, Klesges, et al. (1990) found that obese subjects underestimate their activity level, and recommended weight subjects over estimate.

Skill Level and Interest. Theoretically, an individual's skill level measures their dedication to an activity, which predicts their likelihood to track training regimes and then informs self-reporting. Chase and Godby (1983) found a tennis player's skill level is not correlated with accuracy of self-reports, but skilled swimmers are more likely to accurately self-report swim times. The interest level of swimmers and tennis players has no effect on self-reporting accuracy. Brenner and DeLamater (2014) found that individuals who identify themselves as exercisers are more likely to over report their physical activity.

Temporal Factors. Across all studies, researchers found individuals to be more accurate self-reporters of recent physical activity compared to physical activity that occurred further in the past. Self-reports from the last seven days (or less) are three times more accurate than those from more distant days or months (Brenner & DeLamater, 2014; Sallis & Saelens, 2000; Shephard, 2003), though self-reports from as soon as one hour after physical activity remain susceptible to over reporting (Klesges, et al., 1990). After the passage of three months, respondents are unable to recall physical activity with any accuracy (Shephard, 2003). Shepard (2003) also found that asking respondents about the number of times per week that they engage in physical activity yields the most accurate self-reports.

Social Desirability. Some researchers have found that an individual's likelihood to edit responses to conform with social norms is highly predictive of their likelihood to over report physical activity (Klesges, et al., 2004; Warnecke, et al., 1997). The MCSDS score predicts the likelihood to over report the frequency and duration of physical activity (Adams, et al., 2005) and underreporting of sedentary time (Shephard, 2003). However, Motl, et al. (2005) found insignificant correlation between self-reported importance of exercise and scores on social desirability scales.

Epidemiological research provides relevant insights to guide the best practices in collecting self-reported data on physical activity. First, care must be taken to simplify questions as responses to simple questions are most accurate (Shephard, 2003). College students do accurately report physical activity after 7 days, but the validity of self-reports erodes after 3 months (Shephard, 2003). Physical activity intensity listed on Likert Scales are dependent upon the perception of the respondent and do not give reliable results because respondent perception is based upon their peak activity level (Shephard, 2003). Finally, requests about occurrences of physical activity in times per week is the best method for capturing physical activity as a whole, but summing questions about specific activities (hiking, biking, etc.) results in over reporting (Shephard, 2003). Medical researchers tolerate correlations of 0.3 to 0.5 (Shephard, 2003).

The literature includes many studies that collect self-reported data describing college student CRF use to establish a relationship between CRF use and improved outcomes. CRFs are capital-intensive investments that demand accountability, yet the typical CRF use study is conducted by campus recreation professionals using self-reported data without regard to systematic errors in self-reporting. Self-reported CRF use data concludes that the vast majority of undergraduate students use the CRF (Artinger, et al., 2006; Haines, 2001), and CRF use improves social engagement outcomes (Zizzi, Ayers, Watson, & Keller, 2004), academic outcomes (Rothwell & Theodore, 2006), and wellness outcomes (Miller et al., 2008). Researchers either ask questions about CRF use as a weekly, rate-based estimate (Lindsey & Sessoms, 2006; Todd et al., 2009; Young, Ross, & Barcelona, 2003) or as visits in the last seven days (Miller et al., 2008). Use questions often ask respondents to summarize their interactions during a typical CRF visit (Artinger et al., 2006; Zizzi et al., 2004). To date, only two known studies use university reported CRF use data rather than self-reported data; their research conclusions are consistent with studies using self-reported CRF use data (Belch, Gebel, & Maas, 2001; Huesman, Brown, Lee, Kellog, & Radcliffe, 2007).

Self-Reported Class Absenteeism

College student class absenteeism is a topic of interest among academic engagement researchers, but few studies are conducted because class attendance records for more than a handful of class sections are rare. The only known example is Schmitt et al.'s (2003) study on propensity to self-monitor and self-reported rates of class absenteeism. Respondents estimate their number of classes skipped and complete a self-monitoring assessment test. The likelihood to self-monitor is significantly and inversely correlated with self-reported class skipping rates. However, the researchers lacked validation data on actual class attendance and instead used the grades earned as proxy for actual class attendance rates. Follow up questions asking respondents to elaborate on their answer are a strategy for mitigating social desirability bias in other studies but did not reduce inaccurate reports of class skipping. A more directly measured validation data set would enhance Schmitt et al.'s (2003) conclusion that self-reported class absenteeism is subject to social desirability bias.

Nonresponse Bias and Nonresponse Error in Representation Process

Bias remains a threat to the survey statistics even in cases where self-reported data is an unbiased and accurate description of actual behavioral. Nonresponse bias threatens validity when actual survey variable characteristics of respondents are significantly different than those of nonrespondents and the survey mean is not generalizable to larger groups. The threat of bias is observed as nonresponse error in cases where the actual value among survey respondents is significantly different than the sample frame mean and the characteristics that predict nonresponse propensity correlate the research topic(s) (Groves, 2006). Nonresponse propensity is the unit level likelihood that an eligible unit does not comply with a survey request and is a function of individual and situational factors. Nonresponse bias is attributed to (1) the failure of the researcher to deliver the survey request to certain segments of the sample frame who are more likely to possess a given characteristic; and/or (2) a refusal to participate among members of the sample frame who possess a given characteristic; and/or (3) an inability to participate for members with a given characteristic (Groves, et al., 2009). Respectively, examples include (1) not subscribing to a landline telephone for a survey; (2) disinterest in complying with a simply accomplished survey request; and (3) facing a survey request in an unfamiliar language. Nonresponse bias is present in almost all self-reported data but has a minimal impact on research conclusions in cases where nonresponse error is not correlated with the surveyed topic (Groves, 2006).

To assure data integrity, researchers must study the correlation between nonresponse propensity and the survey topics before using the data in academic study. The threat of nonresponse bias is strongest when the characteristics that make an individual unlikely to comply with a survey request correlate with the survey topic. The desirable scenario is for actual behaviors being surveyed to be unrelated with the likelihood that a potential respondent complies with the survey request, resulting in an actual respondent mean equal to the sample frame mean. The threat to validity escalates as the surveyed behaviors become more correlated with the characteristics that define nonresponse propensity. Groves (2006) provides the Alternative Causal Model to diagram the relationship and to assess the threat of nonresponse bias in a variety of settings. Figure 2.2 describes three models relevant to this study of nonresponse bias:

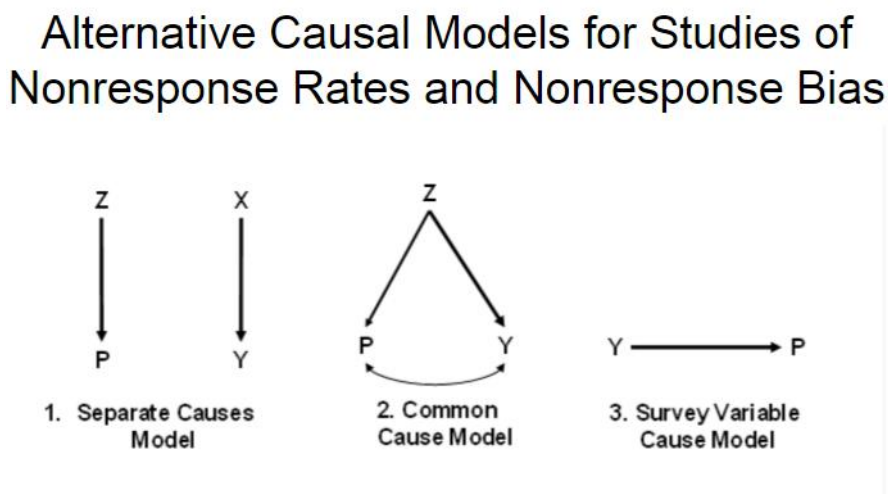

Figure 2.2. Alternative causal model (Groves, 2006, p. 651).

In this model, X and Z are vectors of causes for survey response behaviors, P describes the likelihood to respond to a survey (i.e., nonresponse propensity), and Y is the aggregated values of behaviors being surveyed. The Separate Causes Model (model 1) presents a low risk of nonresponse bias because the attributes that define nonresponse propensity are independent of the behaviors studied. This low correlation likely results in minimal nonresponse bias. The Common Cause Model (model 2) does not necessarily bias the survey statistic, but nonresponse bias threats require investigation because the vector of behaviors that predict nonresponse propensity also describe the studied behaviors. For example, team sport enthusiasts may be more likely to respond to a survey request about Intramural

Athletics participation and also be more likely to participate, so the threat is an overestimate of Intramural Athletics participation. Finally, the Survey Variable Cause Model (model 3) describes a non-ignorable threat of nonresponse bias where the behavior of interest to the study determines the likelihood that an individual will submit a survey response and that respondents are somewhat homogeneous in actual behavior. For example, a survey about CRF use distributed in person, within the CRF, results in non-ignorable nonresponse bias because the survey statistic behaviors exclude data from all CRF nonusers. To evaluate the threat of nonresponse bias, researchers should identify characteristics that are predictive of nonresponse propensity and their correlation to survey variable(s).

However, nonresponse from a refusal to participate is the focus of this study and is observed as a failure to comply with a survey request observed as item nonresponse or unit nonresponse. Item nonresponse describes survey respondents partially complying with the survey request when they fail to answer a portion of the questions on the survey instrument. Item nonresponse may be caused by random occurrences, commonly experienced difficulties with the certain questions, or systematic respondent reactions. For example, respondents may not notice that the back side of a paper survey instrument contains questions, or they may systematically refuse to answer a question when their response describes incriminating behavior (Tourangeau, et al., 2013). Item nonresponse from the latter example is more concerning because systematic nonresponse behavior is more likely to bias the survey statistic (Groves, 2006). Unit nonresponse describes members of the sample frame who refuse the entirety of the survey request and are completely unrepresented in the survey statistic. Unit nonresponse from a refusal to participate owes to the potential respondent's disinterest or perceived burden (Groves, 2006). Groves, et al. (2009) describe four levels of refusal to participate that, taken together, provide one framework of nonresponse propensity theories. The levels include (1) social environment level factors that describe the climate surrounding the survey process; (2) person level factors that describe characteristics about the potential respondent that factor into their response decision; (3) interviewer level factors that describe how personal interactions between potential respondents and interviewers affect response propensity; and (4) survey design level factors that describe reactions to the survey instrument's layout. Broadly, the study of nonresponse analyzes the relationship between item/unit nonresponse and each of the four levels. This study evaluates person level factors that affect nonresponse propensity.

The Leverage-Salience Theory (Groves, Singer, & Corning, 2000) describes the relationship between survey process choices and each potential respondent's nonresponse propensity. Nonresponse propensity describes the likelihood that an individual potential respondent will refuse the survey request. Leverage-Salience theorizes an individual's likelihood of nonresponse results from the aforementioned social environment, person, interviewer, and survey design level factors. Changes to survey implementation uniquely affect each individual's nonresponse propensity in ways that cannot be directly observed, leaving survey researchers to instead estimate the impact of survey implementation changes with changes in response rates across the sample. When survey implementation changes result in higher response rates, the Leverage-Salience theory postulates that the change lowers the nonresponse propensity for most potential respondents.

Response rates describe the percent of potential respondents who respond to a survey question. The American Association for Public Opinion Research (AAPOR) publishes standard definitions of response rates for a variety of settings and research designs. The most applicable definition calculates the response rate for each question as the sum of question completers divided by all members of the sample frame eligible to respond to the question, regardless of viewing the survey request (AAPOR, 2016). The sample frame is fixed because it is defined by the number of emails sent. Not all units are eligible to respond to all questions; for example, those who report never taking an HESF 100 level course are piped past questions about their in-class behavior. Higher response rates indicate that input from a larger share of the sample is included in the survey statistic; as such, there is a widespread belief that higher response rates imply that the survey statistic is less threatened by nonresponse bias (Johnson & Wislar, 2012). However, nonresponse bias has an inconsistent relationship with response rate, so response rates are not necessarily a reliable measure of nonresponse bias (Johnson & Wislar, 2012; Groves & Peytcheva, 2008; Groves, 2006). Estimating the impact of nonresponse bias requires more nuanced analysis than a statement of response rates.

Porter and Whitcomb (2005) offer four methods for the study of nonresponse bias. First, Nonrespondents can be offered follow-up surveys, employing enhanced modes of contact to decrease nonresponse propensity. For example, nonrespondents to an online survey are approached by in-person interviewers distributing paper-and-pencil versions of the survey. Unfortunately, they warn that habitual non-compliers will refuse the initial and all follow up survey requests, so this method will never capture their impact on nonresponse bias (Olson, 2006; Porter & Whitcomb, 2005). Second, a time of response analysis can be utilized where survey means from early respondents are comparable to those from later respondents based upon the assumption that later responders are not substantively dissimilar from nonresponders. However, in a study of recent college graduates, Olson and Kennedy (2006) provide evidence that survey means from the first wave of responders are not significantly different from those responding toward the end of the survey window. This approach also fails to consider habitual nonrespondents who never submit responses. Third, a panel analysis can be used in which the same samples are sent survey request after survey request and habitual compliers' responses are compared to those complying with a fraction of the requests. The assumption is that partial compliers have higher nonresponse propensities and offer responses that approach those of nonrespondents with similar propensities. Again, this approach does not consider habitual non-compliers.

Finally, Porter and Whitcomb (2005) recommend a record linkage approach where respondents and nonrespondents are linked to objective, external databases to compare the characteristics of respondents and nonrespondents. This approach offers complete coverage of the sample and overcomes the disadvantages of the three former approaches. However, even though such validation data sets are rarely available, my study was able to acquire sufficient validation data and employs the fourth approach. For validity studies with known actual survey variables, nonresponse bias is best evaluated as the relationship between the mean of actual values among respondents and actual values among nonrespondents in the sample frame (Olson, 2006). When the difference is close to zero, nonresponse bias is a minimal concern because the actual mean of respondents is not substantively different from the actual mean of all potential respondents. However, when the difference is significantly different from zero, response bias threatens the validity of self-reported data when that data is used as a generalizable estimate of population behavior. Few studies offer other methodologies to study nonresponse bias because possessing actual measures of both respondent and nonrespondent behaviors is rare (Porter & Whitcomb, 2005).

Nonresponse Research

Prior research on nonresponse bias is limited by the availability of validation data but does document consistent relationships between potential respondents' characteristics and nonresponse propensity. Several studies find that there are consistent relationships between demographics and differences in nonresponse propensity. For example, female college students are more likely to respond to survey requests than men (Adams & Umbach, 2012; Clarkberg, Robertson, & Einarson, 2008; Crawford, Couper, & Lamias, 2001; Groves, et al., 2000; Porter & Umbach, 2006; Porter & Whitcomb, 2005; Sax, Gilmartin, & Bryan, 2003). However, Sax, Gilmartin, Lee, and Hagedorn, (2008) found that the gender-related differences in nonresponse propensity are not present in online survey modes, and Olson and Kennedy (2006) found no gender-based response rate differences in multivariate models that account for other college student personality and behavioral descriptors. White students are more likely to comply with survey requests (Sax, et al., 2003) than African American students, and international students are less likely (Clarkberg, et al., 2008; Porter & Umbach, 2006), though Sax, et al., (2008) found that adding more robust background information about potential respondents makes ethnicity an insignificant predictor of nonresponse propensity. Students with a high level of financial need are also less likely to respond (Porter

& Umbach, 2005; Sax, et al., 2008). Porter and Umbach (2005) detect lower response rates from students attending institutions in urban settings and/or public institutions, and Sax, et al., (2008) found that older students and community college students intending to transfer to a four-year college have higher response rates. NCSU is a public university in an urban setting with an unusually high percentage of male students that suggests atypically high levels of aggregate nonresponse propensity.

Prior research also suggests that academically successful students are more likely to respond to survey requests. Students with higher collegiate GPAs are more likely to respond to survey requests (Clarkberg, et al., 2008; Dey, 1997; Olson & Kennedy, 2006) as are potential college student respondents with higher high school GPAs (Sax, et al., 2003). Meanwhile, Aviv, Zelenski, Rallo, and Larsen (2002) found that high GPA students are more likely to be represented in the first wave(s) of survey respondents, and Adams and Umbach (2012) found that students earning high grades in a course are more likely to respond to student evaluations of teaching survey requests. However, when metrics describing personality traits are placed alongside GPA in nonresponse propensity models, there is no statistical significance for GPA in nonresponse propensity modeling (Dey, 1997; Porter & Whitcomb, 2005). Porter and Umbach (2006) also found that students scoring high on the SAT and those attending selective institutions are more likely to respond to NSSE survey requests. Regardless of cause, responses from low achieving students are often underrepresented in survey statistics (Sax, et al., 2008).

Prior research also shows that engagement in the campus and surrounding community are predictive of nonresponse propensity. Students who are socially engaged in the campus community are more likely to comply with survey requests (Clarkberg, et al., 2008; Groves, et al., 2000; Olson & Kennedy, 2006; Porter & Whitcomb, 2005). Theoretically, saliency should predict a positive relationship between engagement and the likelihood to complete a survey that describes engagement (Groves & Peytcheva, 2008) because engaged students are more eager to discuss something relevant to their experiences. Potential respondents are also more likely to comply with a survey request when they have a prior relationship with the survey researcher (Groves & Peytcheva, 2008) or the survey request is pertinent to their life experience (Groves, et al., 2000). In other words, it is likely that engaged students respond to university survey requests because they recognize the researcher or investigating department. However, the finding that engaged students are more likely to complete surveys is not uniform. Kuh (2003) found that engaged students are less likely to complete the NSSE survey, and members of Greek organizations are less likely to respond to general survey requests (Clarkberg, et al., 2008). Nevertheless, the concern is that engaged and academically able students are over represented in survey results (Porter & Whitcomb, 2005).

The aforementioned literature uses a variety of methods to explore the relationship between nonresponse propensity and respondent characteristics. However, because none of the aforementioned studies possesses sufficient validation data, none employs the unit-level, record linkage approach recommend by Porter and Whitcomb (2005). Olson and Kennedy (2006) are able to use this method to provide the most relevant study identified. They distributed a seven topic, unit level survey with survey values corresponding to institution-reported data on academic performance and alumni giving at the University of Maryland. For recent graduates, the validation data set includes cumulative GPA, record of a

D/F/Withdraw at any point in student career, graduation with honors, prior donations to the university, and current dues paying status to the university's alumni association. Olson and Kennedy (2006) compared actual values for individual respondents and the respondent means to the sample frame's mean to study the interaction of measurement error and nonresponse error. Consistent with the literature, the response rate for graduates with GPAs in the A range was 35.5% and significantly higher than the 28.8% response rate for C range GPA graduates. The low GPA respondents were also less likely to accurately describe their past behavior. Overall, respondents over report their academic success by self-reporting a 3.2 cumulative GPA versus an actual value of 3.0. Only 45.5% of respondents self-reported earning a D/F/Withdraw, though institution-reported values found an actual rate of 62.1%. Alumni involvement is the most influential predictor of likelihood to complete a survey about alumni giving with respondents over reporting their giving rates. Overall, the respondent means describe more academic success and generous alumni giving than is true of the sample frame's mean or corresponding institution-reported values at the unit level. My study uses data and methods that align with Olson and Kennedy (2006) with two distinctions. First, Olson and Kennedy (2006) rely upon validation data consisting entirely of graduates from a selective institution, implying that all members of the sample frame are academically successful with the GPA measuring the degree of success. Second, Olson and Kennedy (2006) do not track the response accuracy of an individual across all questions to seek relationships between respondent characteristics and systematic response biases including social desirability bias. Results from this study will better generalize to unsuccessful students and explain how respondent characteristics predict self-reporting accuracy.

Survey Instrument Design

Survey software such as Qualtrix and Survey Monkey lower the cost of delivering online survey instruments to college students. Accordingly, higher education scholars and institutional researchers deliver more surveys and become more dependent on survey data despite widespread signs of survey fatigue (Sarrah & Tukibayeva, 2014). Online survey instruments should be designed with care to reduce survey fatigue and improve the quality of data they yield. Because online surveys are easier to break off than other survey collection methods (Tourangeau, Conrad, & Couper, 2013), they should provide respondents an expected time required for completion and a sense of progress so that respondents allot adequate time to answer all questions (Dillman, Tortora, & Bowker, 1998). Progress bars that allow students to see their quick progression through the instrument are an effective strategy for reducing break off and item nonresponse (Villar, Callegaro, & Yang, 2013). Prior to the survey window, question design should be evaluated by survey experts and members of the survey population (Gonyea, 2005). Closed ended response categories that quantify how respondents behave should be offered in real, measurable units (e.g. number of hours per week, times per month, and so on). Creating a flexible aspect ratio application for smart phones and tablets diversifies the survey modes and increases response rates (Standish, Joines, Young, & Gallagher, submitted). And, as with all surveys, prenotification letters and reminder emails increase response rates (Groves et al, 2009).

This study uses an externally provided data set to validate self-reported behavior, and the survey instrument should be constructed to maximize its validity. Umbach and Kuh (2006) offer five conditions that should be met for participants' responses to be considered valid. First, the information requested must be known to respondents if they were able to perfectly recall past events. Second, the questions must be phrased clearly and unambiguously, which is achieved through careful survey instrument design and testing questions prior to use. Third, the questions should refer to recent activities because the accuracy of memories fades with time. Fourth, the questions should be delivered such that respondents believe that the question merits a serious and thoughtful response. And fifth, an honest response to a question should not be commonly interpreted as a threat to the respondent's integrity, which causes them to respond in a socially desirable way. Because this study tests social desirability bias, the survey instrument will be designed with careful consideration to the first four conditions to assure the validity of findings regarding the fifth condition.

Chapter Summary

This chapter reviewed prior research relevant to answering this study's research questions. Specifically, I introduced Tourangeau's Model of Response Process (1984, 1987) as the theoretical framework through which the research is viewed. I then provided a theoretical discussion of survey validity, social desirability bias, nonresponse bias, and self-monitoring. These topics were followed by a review of applied research, which provides evidence that class skipping, academic performance, and rates of physical activity are susceptible to intentional misreporting and that nonresponse bias is a persistent threat to validity. Below is a review of my researcher questions followed by my hypothesized finding.

1. To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?
    a. What student characteristics are associated with the most accurate and inaccurate self-reporters? Do demographic or background characteristics affect accuracy in self-reporting? Do students with high collegiate GPAs or standardized test scores more accurately self-report behaviors? Do participatory students provide more accurate self-reported behavioral data?
        i. I expect students with higher ability levels (e.g. higher grade point averages and standardized test scores) to more accurately self-report their behaviors. I also expect students who participate in CRF-based, co-curricular activities to be more invested in their image as physically active; thus, CRF-based behavior is more likely to be over reported by frequent CRF users.
    b. Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?
        i. I hypothesize that precise response options that minimize the effort required to answer will elicit the most accurate responses.

c. Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?

    i. I hypothesize that more recent behaviors will be more accurately reported than those from prior semesters or academic years.

d. Are sample means of self-reported behaviors valid estimates of target population means?

    i. I hypothesize that the survey responses, in aggregate, will over report the actual rates of socially desirable behavior and that the survey statistic will be a biased estimator of behavior across the target population.

2. Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?

    a. Do social norms cause students to systematically over report the incidence of socially desirable behaviors such as physical activity, class attendance, and high grades earned? Are self-reported respondent means of socially desirable behaviors less valid estimates of behaviors than those from more socially neutral behaviors?

        i. I hypothesize that socially desirable behaviors will be over reported at higher rates than socially neutral behaviors.

b. Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?

    i. I hypothesize that the rates of exercise, class attendance, and co-curricular behavior will be over reported in a socially desirable direction. The over reporting will correlate with the MCSDS.

The focus of this study is to address the validity of self-reported data about college student behaviors both recent and in the distant past. The literature shows that bias affects the validity of self-reported data on academic performance, academic engagement, and exercise. There are no known studies that replicate this study by tracking the accuracy of day-to-day behaviors of college students. This study is comprehensive study on the validity of self-reported academic and social engagement for self-reported grades, class skipping, and participation at the CRF. Chapter 3 will discuss the methods employed.

CHAPTER 3: METHODS

This chapter describes research design, specifies research questions intended to be answered, establishes the study sample creation, describes the survey instrument and data requested for analysis, and then concludes with a discussion of data concerns, opportunities, and limitations. Chapter Three's objective is to describe the methodological process used to answer the study's research questions.

Research Design

Higher education researchers have a continuing interest in validation studies that detect the presence of nonrandom, systematic bias in self-reported behavioral data. The methodology presented in this section describes a validation study to detect nonrandom misreporting by comparing self-reported data on college student CRF entries, HESF class skipping, and collegiate grades earned. This study compares self-reported data against a large, institution-reported validation data set. Random errors in self-reported data occur when misreporting is uncorrelated with other factors, and their effect can be controlled for with the appropriate quantitative method(s) (Kuncel et al., 2005). However, explanatory factors that predict self-reported deviations from validation data suggest systematic errors that threaten the validity of research findings (Kuncel et al., 2005). Though it is technically possible for systematic misreporting to create a new construct which better responds to the researcher's data needs, it is usually true that systematic misreporting threatens the validity of research findings (Kuncel, et al., 2005).

In cases where unit level validation data is available, Tourangeau et al. (2000) suggests a unit level merger of self-reported and institution-reported data to estimate their correlation.

In this study, OIRP provides university-reported administrative data that I merge with 12 million CRF entry records from DASA, which, when combined, allow me to assess the accuracy of self-reported behavioral data. This research tests the theory that ability, social desirability bias, and student attributes (i.e. MCSDS score, demographics, standardized test scores, interest in physical activity, etc.) explain systematic misreporting of self-reported factual data. Tourangeau et al., (1997) found that exercise is susceptible to social desirability bias; CRF entries are similar in nature to exercising and sufficient for testing for social desirability bias. In prior studies, age, gender and ethnicity are predictive of self-reported accuracy of physical activity (Chen, 2011; Warnecke, et al., 1997), and academic ability is predictive of self-reported accuracy of academic performance (Kuncel et al., 2005). Satisficing respondents are known to provide less accurate descriptions of past behavior because they are not motivated to answer questions thoroughly (Kronsnick, 1999). There is also a theoretical justification to use exercise identity to explain misreporting that is consistent with an individual's exercise identity rather than actual behavior (Brenner & DeLamater, 2014). The institution reported student characteristics offer opportunity for studying these relationships.

Research Questions

The research questions use the Model of the Response Process by Tourangeau (1984, 1987) to evaluate the accuracy of self-reported survey data. Specifically, the research questions are:

1. To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?
    a. What student characteristics are associated with the most accurate and inaccurate self-reporters? Do demographic or background characteristics affect accuracy in self-reporting? Do students with high collegiate GPAs or standardized test scores more accurately self-report behaviors? Do participatory students provide more accurate self-reported behavioral data?
    b. Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?
    c. Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?
    d. Are sample means of self-reported behaviors valid estimates of target population means?
2. Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?
    a. Do social norms cause students to systematically over report the incidence of socially desirable behaviors such as physical activity, class attendance, and high grades earned? Are self-reported respondent means of socially desirable behaviors less valid estimates of behaviors than those from more socially neutral behaviors?

b. Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?

RQ1 tests the retrieval and judgment components of the Model of the Response Process by testing the summary statistics' accuracy of the sample statistic against those for the target population, sample, and respondents. Student characteristics are tested for a statistically significant correlation with the student's accuracy in self-reporting; potential student descriptors include student ability, observed behavioral CRF use patterns, and other university-measured characteristics. RQ2 tests the retrieval and reporting components by testing for systematically invalid variance that can be predicted by the perceived social desirability of the behavior and the respondent's propensity for self-monitoring.

Institution Reported Data

This is a validation study that estimates the validity of self-reported behavioral data by comparing responses to a custom survey instrument to corresponding institution reported data. In March of 2015, I conducted a survey using a simple random sample of 6,000 students that OIRP pulled from the undergraduate population at NCSU. OIRP also provided the student-level, institution reported data which allowed me to validate the self-reported data and find relationships between student characteristics and/or question topics that are related to more accurate self-reports. The validation data set drives the survey instrument content, and the survey questions are designed to be consistent with CRF use questions in prior student behavioral surveys.

All data will be provided as an SAS version 9.4 flat file. To protect subjects' identities, OIRP created a student identification numbering system that allows the data to be merged with other data sets at the student level without revealing individually identifiable information. The final data set is stripped of personally identifiable information and is stored on NCSU owned virtual machines that are password protected and firewalled by NCSU Office of Information Technology. The demographic data is collected and cleaned by OIRP for reporting purposes and is the official record of enrollment at NCSU.

Student background data. OIRP provides demographic data (including gender, ethnicity, age, etc.), measures of ability (including standardized test scores, high school GPA, and collegiate GPA), and portions of the validation data set (collegiate GPA, classes failed, and class schedules). DASA provides student-level data descriptions of student participation in co-curricular activities along with time stamped CRF entries. The demographic data describes student attributes and background at the closest time point available to the first day of the survey window. Demographic and background data are used as explanatory variables that predict misreporting of behavior. Included in the data are the student's age, gender, ethnicity, residency, socioeconomic status (SES), major, and full time equivalent (FTE) enrollment. The student's birthday is provided by OIRP, and age is calculated on the first day of the survey window. Gender and ethnicity are reported in a standardized Integrated Postsecondary Educational Data System (IPEDS) format and describe students on census day (the tenth day of classes) during the semester in which they complete the survey. Also determined on census day, residency describes their home county for residents of North Carolina; for domestic, residency describes home state or territory for non-North Carolina residents; and residency describes country of origin for international students. SES is described by the student's Pell Grant eligibility and unadjusted income and expected family contribution on census day for the fall semester of the academic year of the survey. Family of origin's unadjusted income is taken from submitted Free Application for Financial Student Aid (FAFSA) forms and used to compute expected family contribution. Many students – predominately from high-income families – do not submit FAFSAs, and their income will be missing. Because Pell Grant eligible students rarely bypass the FAFSA, I use Pell Grant eligibility as a binary measure of need that accurately covers nearly all students. Student FTE is determined by the number of student credit hours enrolled on census day for the semester in which the survey is delivered.

Student ability describes pre-college and post-enrollment measures of academic potential and actual performance. In prior research, low achieving students inflate self-reported grade point averages and standardized test scores, and high achieving students accurately report. It is unclear if high achieving students more accurately retrieve memories or if only low achieving students feel social pressure to inflate their GPA. This study tests this finding and further expands the relationship to explain systematic misreporting. The measures of ability collected include SAT scores, weighted high school GPA, and cumulative college GPA. Pre-college measures of ability include SAT score and weighted high school GPA and are taken from the student's first semester of enrollment as a degree-seeking, bachelor's level student. In cases where the SAT score is used in the admissions decision, the SAT score describes the total of the math and verbal score. In cases where the ACT score was used in the admission decision, the SAT score describes an ACT to SAT concordance table conversion (College Board, 2009). In cases where both the SAT and ACT are used in the admissions decision, the SAT score is the higher of the actual SAT score and converted ACT score. In cases where neither the SAT nor ACT scores are used in the admissions decision, the SAT score will be missing (usually international students). Weighted high school GPA describes the students' final GPA with weights for honors and advanced placement courses. The post-enrollment measure of ability is collegiate GPA taken from a file extract frozen from the first day of the survey window. Cumulative college GPA describes the student's official GPA as listed on their transcript and describes all grades earned with adjustments for the course repeat policy, medical withdrawals, and other considerations.

Other measures of student background describe student participation in official DASA programs to test for a relationship between participation and systematic bias in self reporting. Brenner & DeLamater (2014) establish a relationship between exercise identity and inflated reports of physical activity. This study lacks direct measures of exercise identity, but the data provides participation records for exercise-related co-curricular programs that measure interest in exercise programming. There is no known research on the relationship between other co-curricular activity participation and self-reporting accuracy, though I theorize that involved students more accurately track their use and uninvolved students inflate their participation in formal programs. Other participation measures collected include membership in a Greek organization and residential LLV involvement. Membership in a Greek organization describes at least one semester of membership in a socially oriented fraternity or sorority from spring 2009 to current (incomplete records are available prior to 2009). Residential LLV involvement describes at least one fall semester living in a formal residential community with participation expectations.

Validation data. OIRP provides external validation data including grade information (cumulative grade point averages, semester grade point averages, courses failed, and HESF 100 level course grades) and data to help calculate HESF class skipping (class schedules for HESF 100 level courses). DASA provides use rates for formal co-curricular programs (participation rates for club sports, intramural sports, formal fitness classes, and Outdoor Adventures) and data to help calculate HESF class skipping (time stamped CRF entry data). The time stamped CRF entry data is merged with class schedules to determine when HESF 100 level courses met and which students skipped the class meetings. Participation in club sports describes inclusion on a roster for at least one intercollegiate, non-varsity sporting event (team sports and individual sports); participation in intramural sports describes entry in at least one sporting event or tournament (team sports and individual events); and formal fitness programs describe attending at least one not-for-credit group fitness class taught by DASA employees.

For grade related external validation data, direct values are provided for cumulative grade point averages, semester grade point averages, number of classes failed, and grades in HESF 100 level courses. These measures are collected and cleaned by OIRP and are the official record of performance at NCSU. The cumulative grade point average is the official grade point average on the first day of the survey window. The semester grade point averages are from end of term files that describe the grades on the last day of each semester. For approximately one percent of students, the semester grade point from the semester's last day does not match the official grade point average because of grade changes, late grades, and incomplete grades. The number of courses failed counts “Unsatisfactory”, “D+”, “D”, “D-“, and “F” grades on the student’s transcript, including those subject to the course repeat policy. The HESF 100 level grade describes letter grade awarded to students who completed HESF (formerly called PE) 100 (Cross Training), 101 (Fitness and Wellness), 102 (Fitness Walking), 103 (Water Aerobics), 104 (Swim Conditioning), 105 (Aerobics and Body Conditioning), 106 (Triathlon), 107 (Run Conditioning), 108 (Water Step Aerobics), 109 (Step Aerobics), 110 (Adapted Physical Education), 111 (Indoor Group Cycling). All students are required to take at least one HESF 100 level course to graduate with a bachelor’s degree, though many will not have completed this course at the time of survey completion.

The CRF use data includes all time stamped CRF entries from Summer Session 1 2006 to present and time stamped participation records for formal co-curricular programs offered by Campus Recreation at NCSU. In each academic year, there are approximately 1.1 million CRF entries for a total of 12 million entries across all years. Each time a student enters the CRF, their student identification card is presented to a DASA employee at the entrance. The employee checks that the identification photo matches the student and then swipes the identification card through a card reader linked to recreational student fee payment records. The card reader records the exact time that each swipe is stored. There is no record of student use for informal events that occur outside the facility’s walls (e.g. tennis, athletic field use, greenway running track, etc.). When students participate in formal co-curricular activities – club sports, fitness classes, intramural sports, aquatics, Outdoor

Adventures, etc. – their entry is attributed to the co-curricular activity and recorded in a separate data base.

HESF class attendance is computed by combining students' CRF entry time stamps with student class schedules. On a given day, if the majority of students enrolled in an HESF class section enter the CRF within 30 minutes of the scheduled start time, I assume the class section convened for that day. If a student enters the CRF within 90 minutes of a section's start time on a day their section convened, I assume the student entered the CRF specifically to attend their class meeting (it is not uncommon for students to practice a skill or independently exercise prior to HESF class meetings, thus the 90 minutes). A student is assumed to have skipped their class if their class section convened on a given day and the student did not enter the CRF within 90 minutes of the start time. For HESF 100 level courses that regularly meet inside the CRF – pool based water aerobics, for example – this method accurately detects the rate of HESF class skipping for all students. For HESF 100 level courses that regularly meeting outside the CRF – mileage based fitness walking, for example – this method will provide irregular estimates of class skipping among individuals and for all students. I will separately analyze class skipping in all sections and for those meeting regularly in the CRF.

Survey Instrument

This study uses a customized survey instrument to collect self-reported student behaviors with an emphasis on socially desirable/undesirable behavior that I am able to confirm with external validation data (see Figure 3.1 for full instrument). The instrument includes a shortened version of the MCSDS to estimate a given respondent's likelihood to self-monitor and ends with unvalidated questions about physical activity attitudes. The survey will be falsely presented as an attempt to gather unknown information about the relationship between CRF use and student grades.

The survey delivery and data collection are intended to minimize errors of nonobservation and conform to best practices described in Chapter 2. A uniform survey instrument was delivered as a web survey using Qualtrics software provided by the OIRP. I chose a web survey delivery because of its low cost and easy accessibility by the study population. The university requires access to technology, so I assume that members of the sample frame have an email account and internet access. To assure a consistent respondent experience, all communications were sent via formatted letter emails, though I responded to a handful of students who emailed procedural questions. The standardized email messages include one prenotification email, up to 3 reminder emails, and a thank you letter for respondents. The survey window was March 14 to March 24, 2016. The cost of delivery was $0 as no incentives were offered and OIRP provides all students a Qualtrics license, technological support, and email lists.

Question design minimizes measurement error by following the best practices outlined by Groves et al. (2009) and Tourangeau et al. (2000). Specifically, each question considers the seven problems that commonly occur in the response process, including (1) failure to encode the information sought, (2) misinterpretation of the questions, (3) forgetting and other memory problems, (4) flawed judgment or estimation strategies, (5) problems in formatting the answer, (6) deliberate misreporting, and (7) failure to follow instructions (Groves et al., 2009). Response options conform with Sudman and Bradburn’s (1982) suggestions

Figure 3.1. Survey instrument.

Page 1: Carmichael Recreation Complex entries question

1) Alternative versions of the CRF use questions randomly offered
   a. In the **<u>past 7 days</u>**, how many times have you entered the Carmichael Recreation Complex for any reason? Specifically, how many times did you enter the indoor areas that require you to present your student identification card? Count each entry separately, even if you entered multiple times per day.
      - Drop down menu from 0 to 20
   b. Approximately how many times **<u>this academic year</u>** – that is, since the start of fall 2015 – have you entered the indoor parts of the Carmichael Recreation Complex for any reason? Specifically, how many times did you enter the indoor areas that require you to present your student identification card? Count each entry separately, even if you entered multiple times per day.
      - Very often
      - Often
      - Sometimes
      - Never
   c. Approximately how many times **<u>this academic year</u>** – that is, since the start of fall 2015 – have you entered the indoor parts of the Carmichael Recreation Complex for any reason? Specifically, how many times did you enter the indoor areas that require you to present your student identification card? Count each entry separately, even if you entered multiple times per day.
      - I never entered Carmichael Recreation Complex
      - Less than once per week
      - Once per week
      - Twice per week
      - Three or more times per week

Page 2: Academic Performance

2) Rounded to the nearest one hundredth (.01), what is your current **cumulative** grade point average (please include only for classes taken at NC State)? If you are unsure, please estimate.
   - (type here)

3) Rounded to the nearest one hundredth (.01), what was your **semester** grade point average in fall 2015 (please include only for classes taken at NC State)? If you are unsure, please estimate.
   - (type here)
4) In how many courses was your final grade an Unsatisfactory, F, D-, D, or D+ (please include only for classes taken at NC State)? If you are unsure, please estimate.
   - 0 (Never received F, D-, D, or D+)
   - 1 course
   - 2 courses
   - 3 courses
   - 4 courses
   - 5 or more courses
5) In how many courses was your final grade an A+, A, or A- (please include only for classes taken at NC State)? If you are unsure, please estimate.
   - Drop down list 0 to 20 or more

Page 3: PE/HESF Course Taking

6) How many for credit Health and Exercise Studies/Physical Education courses have you taken at NC State only? Include all courses at any level.?
   - Drop down list 0 to 4 or more

7) Including courses in which you are currently enrolled, select the most recent of the **Health and Exercise Studies Fitness 100 level** course you have taken. These are the HESF **100 level** (previously called Physical Education) courses that are required to complete general education requirements. If you have not taken an HESF 100 level course, click on "I have not taken..." option; if you cannot remember the course please indicate "Not sure."
   - PE/HESF 100 (Cross Training)
   - PE/HESF 101 (Fitness and Wellness)
   - PE/HESF 102 (Fitness Walking)
   - PE/HESF 103 (Water Aerobics)
   - PE/HESF 104 (Swim Conditioning)
   - PE/HESF 105 (Aerobics and Body Conditioning)
   - PE/HESF 106 (Triathlon)
   - PE/HESF 107 (Run Conditioning)
   - PE/HESF 108 (Water Step Aerobics)
   - PE/HESF 109 (Step Aerobics)
   - PE/HESF 110 (Adapted Physical Education)
   - PE/HESF 111 (Indoor Group Cycling)
   - I have not taken a PE/HESF 100 level course (skip to question 12)
   - Not sure (skip to question 12)

Page 4: Academic, for credit Physical Education course taking and behavior

8) What grade did you earn in [piped in name of question 9 course]? (select one)
   a. Currently enrolled (no grade)
   b. Satisfactory
   c. A-/A/ A+
   d. B-/B/B+
   e. C-/C/C+
   f. D-/D/D+
   g. F
   h. Not sure

9) How many class meetings did you miss for any reason (e.g. you were sick, didn't feel like going, had other obligations)? Only count class meetings you missed entirely; do not count class meetings you were late for or left early.
   a. 0
   b. 1
   c. 2
   d. 3
   e. 4
   f. 5
   g. 6
   h. 7 or more
   i. I took a distance education section

Page 5: Participation and planned participation in Campus Recreation Programming

10) For each activity listed below, please identify if you have participated at any time **while enrolled at NC State?** Please include each event or competition separately; for example, count each game of an intramural basketball season.

   A) Intramural athletics, including only games, tournaments, and other competitions (excluding practice)
   - I have never participated
   - For one semester
   - During multiple semesters
   - In all semesters

   B) Club sports, including only games, tournaments, and other competitions (excluding practice)
   - I have never participated
   - For one semester
   - During multiple semesters
   - In all semesters

   C) Formal fitness classes, including only fitness meetings located inside NC State's Carmichael Complex
   - I have never participated
   - For one semester
   - During multiple semesters
   - In all semesters

D) Outdoor Adventures, counting each instructional activity, day-long trek, or weekend excursion as one event

- I have never participated
- For one semester
- During multiple semesters
- In all semesters

Page 6: Ballard's Shortened Marlowe Crowne Social Desirability Scale

11) I sometimes feel resentful when I don't get my way.
    a. True
    b. False
12) On a few occasions, I have given up doing something because I thought too little of my ability.
    a. True
    b. False
13) There have been times when I felt like rebelling against people in authority even though I knew they were right.
    a. True
    b. False
14) No matter who I'm talking to, I'm always a good listener.
    a. True
    b. False
15) I can remember "playing sick" to get out of something.
    a. True
    b. False
16) There have been occasions where I took advantage of someone.
    a. True
    b. False
17) I'm always willing to admit when I make a mistake.
    a. True
    b. False
18) I sometimes try to get even rather than forgive and forget.
    a. True
    b. False
19) I am always courteous, even to people who are disagreeable.
    a. True
    b. False

20) I have never been irked when people expressed ideas very different from my own.
    a. True
    b. False
21) There have been times when I was quite jealous of the good fortune of others.
    a. True
    b. False
22) I am sometimes irritated by people who ask favors of me.
    a. True
    b. False
23) I have never deliberately said something to hurt someone's feelings.
    a. True
    b. False

Page 7: Exercise Attitudes and Open-Ended Question

24) I believe being physically active is important for my wellness and health. (check one)
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree
25) When you exercise or participate in physical activity, how often do you do so inside NC State's Carmichael Recreation Complex? This only includes areas you are required to swipe your student identification card to enter. (check one)
    - Almost all of the time
    - Most of the time
    - About half of the time
    - Less than half of the time
    - Almost never
26) How do you believe physical activity and exercise are related to academic success?
    - Open text
27) Is there anything else you would like the researchers to know about physical activity and academics?
    - Open text

including (1) to offer all reasonable possibilities as explicit response options, (2) request specific information, (3) use words all respondents will understand, (4) provide memory cues, and (5) use visually consistent presentation across all questions.

Because this study informs the researchers on the validity of their survey data and impact on research conclusions, this study's question design mimics survey questions from prior studies found in recreational sports journals (Haines, 2001; Mahoney, 2011; Zizzi et al., 2004). Table 3.1 shows the survey question from prior studies and its mapping to my survey question in this survey; in most cases, the validation data set limitations prevent a direct representation of the questions.

Alternative response options for question 1 tests the validity of recent use, rate based estimation and Likert scales. Some respondents are randomly selected to describe their CRF use in the prior 7 days with response options ranging from 0 to 20. Other questions format as times per week, including the options "I never entered," "Less than once per week," "Once per week," "Twice per week", and "Three or more times per week." Shephard (2003) finds rate based response options are the best method for capturing valid data on physical activity. Weekly rates are an accessible method for describing rates over long time periods without requiring respondents to recall and count over months or years. Other respondents are randomly selected to describe past CRF use as intensity categories listed on a Likert scales including the options "Very Often," "Often," "Sometimes," and "Never." Shephard (2003) finds Likert scales are dependent upon the perception of the respondent relative to their peak activity level and do not yield reliable estimates of physical activity. Despite this finding, these Likert scale response options are offered because they mimic the format of NSSE and other national surveys (Kuh, 2004). Respondents and higher education researchers have used Likert scale response options without necessarily understanding the actual frequency of the behavior.

Table 3.1.

Survey Instrument to Prior Research Question Map

During an average week, how many times do you visit the student recreation center (Zizzi, et al., 2004, p. 605)	In the **past 7 days**, how many times have you entered the Carmichael Recreation Complex for any reason? Specifically, how many times did you enter the indoor areas that require you to present your student identification card? Count each entry separately, even if you entered multiple times per day.
	For each activity listed below, please identify if you have participated at any time **while enrolled at NC State?** Please include each event or competition separately; for example, count each game of an intramural basketball season.
Do you currently participate in any PHYSICAL EXTRA-CURRICULAR ACTIVITIES (e.g., intercollegiate athletics, intramurals, club teams)? (Zizzi, et al., 2004, p. 611)	A) Intramural athletics, including only games, tournaments, and other competitions (excluding practice)
How often do you participate in the following activities, programs, and/or services at Campus Recreation Center Facilities? Open recreation, intramural sports, club sports, aquatics, personal training. (Mahoney, 2011, p. 114)	B) Club sports, including only games, tournaments, and other competitions (excluding practice)
Have you participated in a fee-based physical activity programs offered at the Student Recreation Center (e.g., spinning, aerobics, yoga, martial arts, etc.)? (Zizzi, et al., 2004, p.606)	C) Formal fitness classes, including only fitness meetings located inside NC State's Carmichael Complex

To test temporal effects on memory, respondents will be asked to retrieve information from the entirety of their collegiate career. Depending upon length of a student's career, they will reach back weeks, and for others, years. Grade information and class skipping patterns may be retrieved from the end of the fall 2015 semester (3 months) to summer 2006 (10 years). The number and/or rate of CRF entries require respondents to retrieve information from recent days (prior seven days), same academic year (1 to 7 months), and further into history (four or more months). Depending upon the span into the past, the demands of memory vary. This allows me to test how the passage of time affects the validity of responses.

The CRF use portion of the instrument includes questions about number of CRF entries, information about the specific purposes for the CRF visits, the number of HESF class meetings the respondent failed to attend, and opinion on the value of exercise. For the question about CRF use, I randomly offered a version asking respondents to recall and count total CRF visits and another asking for an estimated rate of attendance. Respondents offered their participation in Intramural Athletics, Club Sports, Outdoor Adventures, formal fitness programs, for credit HESF courses, and informal recreation. Additional questions collected data for items without verifiable data including the respondent's view of using the CRF as a desirable behavior and their exercise rates and non-CRF based activity. The salience of asking respondents to self-report HESF class meetings skipped as a measure of socially undesirable behavior is debatable but consistent with prior research on other academic disciplines (Schmidt et al., 2003). Perhaps erroneously, I assume students view the HESF courses differently because the content and delivery differ substantially from courses in other disciplines offered at NCSU.

The self-reported grades portion of the instrument includes questions about low grades earned, grades in HESF, and grade point averages. All questions about grades and grade point averages request sensitive information and offer the opportunity for systematic misreporting with a bias towards inflating self-reported socially desirable behaviors. Grade information is an accepted and salient measure of socially desirable behavior (Kuncel et al., 2005).

The MCSDS is a generally accepted tool used to detect propensity for self-monitoring (Ballard, 1992; Gonyea, 2005; Leite & Beretvas, 2003; Snyder, 1987). Respondents complete the MCSDS by answering 33 true-false questions that describe socially desirable behavior untrue of most people or socially undesirable behavior that is commonplace (this study uses a 13 question shortened version offered by Ballard). Likelihood to self-monitor is scored by totaling the untrue, socially desirable responses provided (Crowne & Marlowe, 1960). For example, “My table manners at home are as good as when I eat out in a restaurant” (question 8) is unlikely to be true for the majority of people, so respondents marking true are more likely HSMs. Conversely, “I am sometimes irritated by people who ask favors of me” (question 30) is likely true for most people, so respondents marking false are more likely HSMs. Researchers usually co-administered the MCSDS with their survey instrument and then test for correlations between the MCSDS results and self-reported scores on socially desirable questions. Statistically significant correlations indicate the presence of social desirability bias and suggest that the survey items are not valid (Gonyea, 2005). Dobbins et al. (1993) assume that LSMs provide more accurate data and suggest weighting results by the MCSDS score.

Study Sample

The target population is the "group of elements for which the survey investigator wants to make inferences by using sample statistics" (Groves et al., 2009, p. 69). This study's research conclusions are directly generalizable to all bachelor's level students at NCSU enrolled in the spring semester of 2016. To a lesser extent, the results of this study inform survey researchers on how all college students describe their behaviors in surveys. Compared to all college students nationally, the NCSU student body has a higher share of students who participate in a traditional residential college experience; three quarters of new freshmen live in university-owned residence halls with another 15% living in privately owned dormitories that abut campus. Over 99% of the freshmen cohort are traditionally aged and enrolled full time in their first semester. NCSU is a more selective institution than most with atypically high freshmen profile metrics. Finally, NCSU has an unusually high percentage of engineering students and other STEM majors enrolled in the university.

The survey population is closely aligned with the target population, though students for which there is incomplete validation data available are excluded. Specifically, the sample population includes all NCSU students enrolled in a degree-granting, baccalaureate level academic program on the tenth day of classes in spring 2016 with complete validation data sets. The grade related validation data is available from 1993 to the present, and the CRF use validation data begins in the summer of 2006. Students first enrolling prior to summer 2006 are excluded from the sample frame. Over 95% of the students in the target population are included in the survey population. To avoid respondent confusion about question intent, I remove all fall 2016 students who initially enrolled in any NCSU classes prior to June 1,

2006. Between the sample, all demographic and degree majors are substantially represented with the exception of older students who returned to campus after a stop out period.

The sample frame describes procedures used to contact potential respondents and list all individuals who have a chance of receiving a survey instrument (Groves et al., 2009). In this instance, the sample frame and the sample population are identical because the university-created data set which defines an individual as being a student also requires each student's contact information. The sample frame is a list of all baccalaureate students enrolled on the tenth day of classes and is provided by OIRP at NCSU. Coverage exceeds 95% of the target population with sources of error limited to a small number of ineligible units and undercoverage.

Ineligible units are individuals included in the sample frame despite not being in the target population (Groves et al., 2009). For example, an ineligible unit may describe students who informally withdrew from the university without using the official channels of communication (i.e. asking for a tuition refund). Undercoverage describes individuals who are in the target population but not included in the sample frame (Groves et al., 2009). For example, it may describe students enrolling in classes after the tenth day census. Historic student behavior suggests that less than 1% of students are ineligible units or in undercoverage groups.

Table 3.2.

Data Summary

Data Source	Variables	Availability	Variable Type	Variable Role
OIRP	Cumulative GPA	1987 to current	Continuous	Validation data
OIRP	Semester GPA	1987 to current	Continuous	Validation data
OIRP	Number of A grades	1987 to current	Continuous	Validation data
OIRP	Number of courses failed	1987 to current	Continuous	Validation data
OIRP	HESF 100 level grades	1987 to current	Categorical	Validation data
OIRP	HESF 100 class schedule	1987 to current	Continuous	Validation data
DASA	CRF entries	2006 to current	Continuous	Validation data
DASA	CRF entry purpose	2006 to current	Categorical	Validation data
Survey	Cumulative GPA	Spring 2016	Continuous	Self-reported data
Survey	Semester GPA	Spring 2016	Continuous	Self-reported data
Survey	Number of courses failed	Spring 2016	Continuous	Self-reported data
Survey	Number of A grades	Spring 2016	Continuous	Self-reported data
Survey	HESF 100 level grades	Spring 2016	Categorical	Self-reported data
Survey	HESF 100 level class skipping	Spring 2016	Continuous	Self-reported data
Survey	CRF entries	Spring 2016	Continuous	Self-reported data
Survey	CRF entry purpose	Spring 2016	Categorical	Self-reported data
OIRP	Gender	1976 to current	Binary	explanatory variable
OIRP	Ethnicity	1976 to current	Categorical	explanatory variable
OIRP	Age	1987 to current	Continuous	explanatory variable
OIRP	SAT Score	1987 to current	Continuous	explanatory variable
OIRP	College GPA	1987 to current	Continuous	explanatory variable
DASA	CRF entry rate	2006 to current	Continuous	explanatory variable
DASA	Greek membership	2009 to current	Binary	explanatory variable
DASA	Living Learning Community	2009 to current	Binary	explanatory variable
DASA	Co-curricular participation	2006 to current	Categorical	explanatory variable
OIRP	Family income	2003 to current	Continuous	explanatory variable
OIRP	Pell Grant recipient	2003 to current	Binary	explanatory variable

The sample is selected from the sample frame using a simple random sampling without replacement. A simple random sample without replacement assigns an equal, nonzero probability of selection to all elements in the sample frame and prohibits elements from being selected more than once (Groves et al., 2009). The actual population means and measures of variance for the entire sample frame are known to the researcher. Sample sizes are typically determined using tables that specify the minimum number of units required to be sufficiently representative of the total population. However, instead of sampling to estimate the incidence of behavior across the target population, this study's goal is to collect enough data to assess the validity of student responses. The preferred sample size is as large as possible, but OIRP maintains policies to prevent over surveying of students and the sample size is determined by this policy.

Prior to data collection, it was important to estimate the required sample size for a correlation study. The sample provided is 6,000 students, and the number of responses is sufficient to evaluate the research questions. The correlation between self-reported values and institution-reported values is a numerical index that measures the strength of the relationship between the two variables and is often represented as 'r'. A positive correlation indicates the measures increase (or decrease) in the same direction, and a negative correlation indicates the measures increase (or decrease) in opposite directions. A statistically significant correlation between measures allows one estimate to predict another. In quantitative studies, it is typical to strive for a 95% confidence interval with a width of .1. Table 3.3 shows the sample size required to meet this threshold for each correlation level (Moinester & Gottfried, 2014). Prior to the sample selection, the correlation between the self-reported and institution-reported values was unknown but presumed somewhat highly correlated (i.e. 0.75 or higher). For a 0.75 correlation, Table 3.3 suggests a sample size of 299. A recent online survey of mid-career undergraduate students at NCSU received a 26% response rate (OIRP, 2013) and is considered a guide for my survey's response rate. These ratios (i.e. 299 divided by 0.26) recommend that I survey at least 1,150 individuals. OIRP provides a simple random sample of 6,000, and all sub questions are offered to at least 1,150 students. The adequate sample size combined with above average response rates yielded well over the required 299 responses for every question.

Data Storage

NCSU's OIRP provides each of the aforementioned data sets separately and with a consistent student identification number system. The identification numbers map to, but are not identical to, the university's official student identification. The data provided should not allow any third party to use student identification numbers to identify survey participants or other students. Each data set is a standalone file. I merged the files together to create panel data sets that meet the requirements of each research question.

Some data is further transformed to obfuscate the identities of small minority populations. For example, Pacific Islanders represent less than 1% for the student population and, when paired with Greek participation data, could make subjects individually identifiable. In such cases, the descriptive characteristics are combined with other subpopulations so that no subject can be identified. In this example, Pacific Islanders are combined with Native Americans to make identification impossible.

Table 3.3.

Sample Size Requirements

Sample Size Requirements for Desired 95% Confidence Interval at .05 Level	
|r|	N
0.05	1530
0.10	1507
0.15	1469
0.20	1418
0.25	1352
0.30	1274
0.35	1185
0.40	1086
0.45	980
0.50	867
0.55	751
0.60	633
0.65	517
0.70	404
0.75	299
0.80	205
0.85	125
0.90	62
0.95	22

Data is stored safely on virtual machines that are firewalled and protected by NCSU's Office of Information Technology. All data sets remain in SAS version 9.4 tables that require a SAS license and programming skills to view. Non-identifiable, aggregated data summaries of analysis will be exported into Microsoft Office for the creation of tables and figures. IRB approved an eighteen-student pilot study of question quality but otherwise exempted this study from IRB oversight because the identity of respondents is protected.

Data Summary

Data provided by OIRP and DASA are combined to create a detailed, institutionally-reported description of each student's demographics, background, academic performance, co-curricular participation, and CRF use. Institutional data on academic performance and CRF use are combined with self-reported data on the corresponding topics. Included in the survey data is a shortened version of the MCSDS that estimates each student's likelihood of self-monitoring. The final data set allows me to validate the accuracy of self-reported data with presumed accurate institution reported data. Survey questions are designed to study the validity of the survey instrument with attention to the effect of social desirability bias. The literature establishes a need for a validation study of student behaviors.

Methods

This section establishes how the data and methods are integrated to answer this study's research questions. The research questions use the Model of the Response Process by Tourangeau (1984, 1987) to evaluate the accuracy of self-reported survey data. Specifically, the research questions are:

1. To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?
    a. What student characteristics are associated with the most accurate and inaccurate self-reporters? Do students with high standardized test scores or collegiate GPAs more accurately self-report behaviors? Do participatory

students provide more accurate self-reported behavioral data? Do demographic or background characteristics affect accuracy in self-reporting?

   b. Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?
   c. Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?
   d. Are sample means of self-reported behaviors valid estimates of target population means?

2. Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?
   a. Do students systematically over report the incidence of socially desirable behaviors such as exercise, class attendance, co-curricular participation, and high grades earned?
   b. Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?
   c. For socially desirable topics, are sample means of self-reported, socially desirable behaviors valid estimates of target population means?

Answers to these research questions have implications on the validity of higher education data collected by survey researchers, from customized survey instruments at single institutions to standardized survey instruments delivered across many institutions.

Specifically, this study furthers the field by assessing the validity of self-reported, college student behaviors, especially for sensitive topics.

This is a quantitative study that combines external validation data and a survey instrument to identify sources of variance in self-reported behavioral data. The goal is to measure misreporting and attribute the percentage of variance that is valid variance, systematic invalid variance, and random error. Systematic invalid variance is explained by many factors, but the focus of this study is the effect of respondent characteristics and social desirability bias. To examine this study's research questions, comparing the self-reported behaviors to the institution-reported behaviors among subsets of students is critical. These subsets differ by their placement on a continuous scale (i.e. grade point averages, number of CRF visits, etc.) and in categories (i.e. LLV participants, Greek status, gender, etc.). Described below is my methodology, which is modeled to replicate methods of prior studies that compare self-reported behaviors to institution-reported behaviors.

RQ 1: Validity Estimation

This section describes methods used to answer RQ1. RQ1 tests the retrieval and judgment components of the Model of the Response Process within the context of the summary statistic's validity as a substitute for target population means, sample means, and the actual values among respondents. The retrieval component describes recalling relevant information from memory with the help of a retrieval strategy. The judgment response component describes evaluating retrieved memories for accuracy and then imputing missing information to create an aggregated description of behaviors. I evaluated a respondent's ability to accurately report past behaviors within the context of dates and frequencies. Prior research recommends that I identify sources of variance in self-reported measures and then estimate the percentage of variance attributed to valid variance, systematic invalid variance, and random error (Kuncel et al., 2005). My focus is to estimate systematic invalid variance as observed as differences in self-reported and actual behavior that is explained by other measured respondent descriptors or survey processes.

The survey instrument requires respondents to retrieve memories about academic performance and CRF use from their memories and then use a strategy to impute missing information to self-report their behaviors. This description may or may not be accurate, and the accuracy is a function of ability and effort. A conclusive answer to RQ1 will include a correlation between the summary statistics and survey statistics along with evidence of systematic, nonrandom misreporting (or the lack thereof). A high correlation suggests the survey instrument provides an accurate description of behavior across the sample. Meanwhile, a low correlation suggests that self-reported data is a poor substitute for actual behavior. To identify self-reported validity differences among subpopulations, I calculated the correlation for each subpopulation; subpopulation examples include high GPA students, HSMs, and students putting varying levels of effort into completing the survey. Because the population means are known, I also evaluated the survey statistic as a valid descriptor of actual values across the target population, including for subpopulations of students described by demographics, behaviors, and ability. The methods used to answer RQ1 are described below.

Estimated validity. The correlation between self-reported responses and institution reported responses measures the validity of self-reported data as a substitute for the actual population values. For estimating the validity from one trial, Groves et al. (2009, p. 275) recommends

$$\text{Estimated Validity}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\widehat{\mu_i}-\bar{\bar{\mu}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\widehat{\mu_i}-\bar{\bar{\mu}})^2}}$$

where y_i is the response for the i^{th} respondent, $\bar{y}$ is the mean response for all respondents, $\hat{\mu}_i$ is the i^{th} respondent's actual value of record, and $\bar{\hat{\mu}}$ is the mean of the actual values across all respondents. The estimated validity, or correlation score, is an absolute value between 0.0 and 1.0; values approaching 0.0 suggest no validity, and values approaching 1.0 suggest perfect validity (Groves et al., 2009).

Means comparison. The leniency score is a means comparison between the self-reported and institution reported (actual) mean. The leniency score is represented as

$$\text{Leniency Score} = \bar{y} - \bar{\hat{\mu}}$$

where $\bar{y}$ is the mean of self-reported responses to a given question, and $\bar{\hat{\mu}}$ is the corresponding institutional reported mean. Statistically significant differences between the means are detected with a t-test. A statistically significant difference between the means provides evidence that students are not accurate self-reporters, and the survey statistic is not a valid estimator of actual behavior.

Intergroup comparisons. Intergroup comparisons are conducted using intercorrelations and between groups comparisons of leniency scores. An intercorrelation shows the correlations of every group of interest and draws attention to trends of interest as a function of explanatory variables. This method allows us to locate certain subpopulations who are more accurate and less accurate self-reporters. Figure 3.2 uses plausible but fake data; it would provide additional evidence of high GPA students being more accurate self-reporters. If this were actual data, the plot's downward slope would indicate that the true overall GPA explains the accuracy in self-reporting.

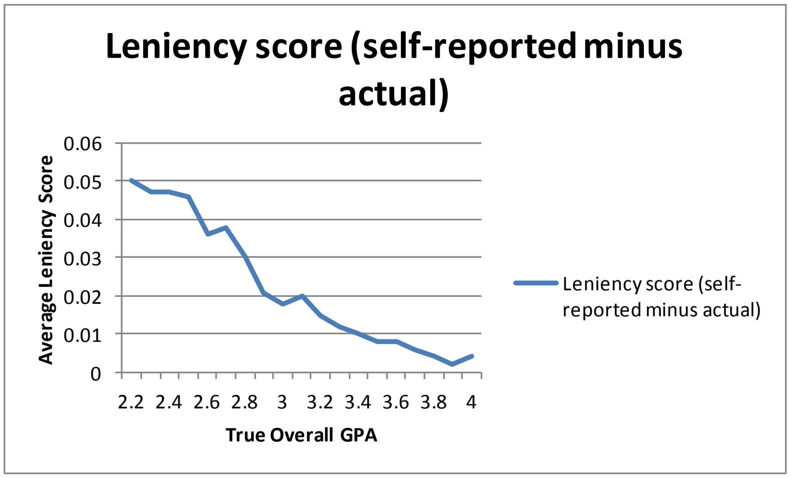

Figure 3.2. Leniency Score Plot.

Estimates of bias. In most studies of nonresponse bias, researchers possess at least one known survey statistic but otherwise rely on estimation methods and/or theory for measures of nonresponse propensity and actual survey variable values. In such cases, Groves (2006) suggests that the study of nonresponse bias embraces nonresponse models founded upon correlations between survey behaviors and unobserved nonresponse propensities. However, this study possesses known actual survey variable values, and the actual response decisions of an individual's nonresponse propensity is the only unknown value. As such, Grove's recommendations for nonresponse study are less applicable than measures of bias recommended by Olson (2006) in a study with similar validation data and design (Olson's publication summarizes a dissertation chaired by Robert Groves). Olson's formula to calculate nonresponse bias is:

$$bias_{NR} = \left| \frac{\bar{y}_{respondent\ record} - \bar{y}_{frame}}{\bar{y}_{frame}} \right|$$

Where $\bar{y}_{respondent\ record}$ is the unadjusted, actual mean of the variable of interest y for all respondents; then $\bar{y}_{frame}$ is the unadjusted, actual mean for all members of the sample frame (e.g., actual mean for 100% response rate). When $bias_{NR}$ is close to zero, nonresponse bias is a minimal concern because the actual mean of respondents is not substantively different from the actual mean of all potential respondents. However, as $bias_{NR}$ increases from zero, the nonresponse bias presents a greater threat as the validity of self-reported data diminishes and becomes a less generalizable estimate of the sample's behavior. This measure of nonresponse bias does not incorporate individual nonresponse propensities.

RQ2: Social Desirability Bias

RQ2 tests the retrieval and reporting components of the Model of Response Process (Tourangeau, 1984, 1987) within the context of social desirability bias. I hypothesize that the presence of systematically invalid variance is a predictable function of perceived social desirability of the behavior and the respondent's propensity for self-monitoring. Validity in the retrieval component involves recalling relevant information from memory with the help of a retrieval strategy. The reporting component involves providing an answer that is not edited for acceptability (or other criteria). In behavioral self-reports, socially desirable questions are a subset of sensitive questions that presuppose that "respondents believe there are norms defining desirable attitudes and behaviors and that they are concerned enough about these norms to distort their answers to avoid presenting themselves in an unfavorable light" (Tourangeau et al., 2000, p. 257). To estimate the social desirability bias, I evaluated a respondent's self-reporting accuracy with the assumption that earning high grades, co-curricular participation, attending all HESF courses, and exercising frequently conform to desirable social norms. The recommended method was to first establish the validity of the survey statistic. Then, if invalid, attribute the systematically invalid variance to social desirability bias in cases for which the question content or MCSDS score are predictive of the ratio of over reporting to underreporting in a socially desirable direction.

The completed survey instrument includes self-reported behavior about socially desirable academic performance and co-curricular participation as well as an MCSDS score that measures the likelihood to self-monitor. Assuming that the MCSDS is a valid measure of likelihood to self-monitor and holding all else constant, social desirability bias is present if the MCSDS is a significant predictor for misreporting behavior on socially sensitive topics or if only questions about socially desirable behaviors are systematically inflated (Groves et al., 2009). A conclusive answer to RQ2 will include (1) a correlation between over reporting socially desirable behavior and the MCSDS score or (2) a substantially higher ratio of over reporting than underreporting for questions about socially desirable behavior with a one-to-one ratio of over/under reporting for other questions. Evidence of social desirability bias suggests that the survey instrument would have provided an invalid description of behaviors with social desirability bias contributing to the invalidity. Invalid survey responses are a threat to the validity of research based on those questions.

To answer RQ2, I compared the survey instrument's self-reported data to corresponding institution-reported validation data to estimate impact of social desirability bias on self-reported behavioral data. The methods used to answer RQ2 are described below.

Frequency of error. Frequency of error describes the number of occurrences of over reporting and underreporting a behavior. I theorized that a socially desirable behavior would be over reported much more often than underreported. For example, a meta-analysis of college GPAs by Kuncel et al. (2005) found that only 54% of students correctly reported their collegiate GPAs; students over reported their GPAs at almost four times the rate that they were underreported (34.5% versus 8.8%), suggesting that college students are more likely to inflate self-reports of socially desirable behavior. Table 3.5 provides the Kuncel et al. (2005, p. 75) findings.

Table 3.4.

Over Reporting Ratios

	% Over	% Accurate	% Under
College GPA	34.5	54.3	8.8

In addition to self-reported GPA, I performed this analysis for all self-reported variables to identify high rates of over reporting. High over reporting ratios should be a function of socially desirable questions.

This analysis can be segmented by the MCSDS score, which is a widely recognized measure of likelihood to self-monitor image. The distributions of intergroup MCSDS score comparisons are the distribution of leniency scores as a function of the score. This method allows us to locate subpopulations of HSMs – who are theorized to inflate socially desirable behavior – and LSMs who are theorized to be more accurate self-reporters. Using plausible but fake data, I provide an example of a leniency score distribution graph which follows in Figure 3.3. If this analysis used actual data, there is strong evidence of the likelihood to self-monitor affecting accuracy in reporting. I might conclude that collegiate sensitive questions are likely to be invalid estimators of socially desirable behavior.

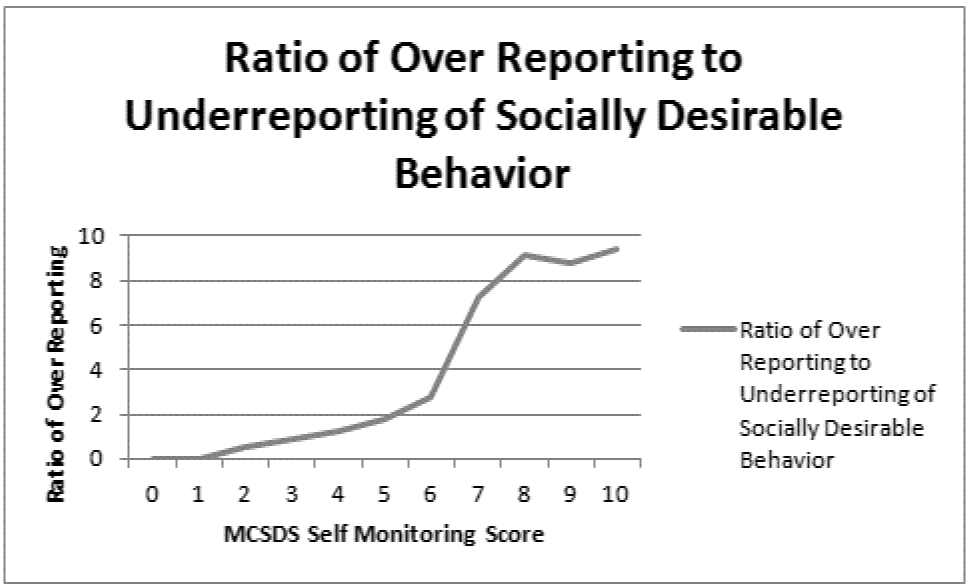

Figure 3.3. Ratio of reporting by MCSDS score.

Magnitude of bias. Mayer, et al. (2007) provide a measure of bias for comparing the severity of over reporting to underreporting calculated as the absolute value of the ratio of the difference between the self-reported measure and the institution-reported measure divided by the range of all possible scores. In the example of cumulative collegiate GPAs, the range is 4.0 grade points. For example, a student self-reports a 3.8 GPA, but their institution-reported GPA is 3.4. Their bias ratio is 0.10, or |(3.8-3.4)|/4.0. These bias ratios are separated by over reporting students versus underreporting students and then aggregated across the group for t-test comparison. The formula for a Welch's t-test for two groups of unequal size and unequal variance is

$$\mathrm{t} = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{X_1} - \overline{X_2}}}$$

where $s_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Social desirability is detected if the over reporting group has a statistically significant and larger bias ratio than the underreporting group.

Correlation comparison. The correlation between the self-reported and institution reported values is an indicator of validity (Kuncel et al, 2005). Questions about socially desirable topics are theorized to be less valid measures than those about neutral topics. Cohen (1991) provides a method for comparing correlation scores of two questions.

$$\mathrm{Z} = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{1/(N_1 - 3) + 1/(N_2 - 3)}}$$

where Z_{r_1} and Z_{r_2} are the correlations of differences to be tested, and N_1 and N_2 are the corresponding sample sizes. Theoretically, social desirability bias is present if there is a statistically significant and higher correlation for socially neutral questions.

Duncan tests. Duncan's (1955) new multiple range test is a multiple comparison procedures test that consists of an iterative series of pairwise means comparisons that test every independent variable mean against all other groupings' combinations of means for the given variable. First, the test ranks the means from largest to smallest. For each given mean, m_i, the test compares that m_i - m_j is not statistically significantly different from zero at the critical value. If m_i - m_j does not exceed the critical value, the test goes on to test that m_i minus the subset of (m_j, m_{j+1}) is not significantly different from zero; if this is not significantly different from zero, it goes on to test the combinations of the set (m_j to m_I) until a significant difference is detected or all combinations are tested without significant differences. After testing m_i, the Duncan test moves on to m_{i+1}, and then m_{i+2}, etc., until all mean values have been compared to all other means grouped by the independent variable value. The result is an iterative significance testing process that groups independent variables that are significantly different from the remaining values that have mean differences in both directions. For example, a Duncan Test to determine significant differences in cumulative GPAs between students in each major yields groupings of majors that have significantly different average GPAs from other groups. Table 3.6 illustrates a fictitious example.

Table 3.5.

Duncan Test Sample

Duncan Group	Mean GPA	N	Major Ranking
A	3.48	52	Business
B	3.31	4	History
B	3.14	8	English
B C	3.01	21	Engineering
C	2.89	4	Statistics

Means with different letters are significantly different from means in other letter groups.

In this example, the results show that students majoring in rigorous STEM majors like engineering and statistics have significantly lower cumulative GPAs than students in all other major combinations. This difference is signified by the “C” Duncan group designation where the combination of all majors with a “C” label have significantly different means than all groups not labeled “C.” Meanwhile, the same Duncan test finds that students in business majors have significantly higher GPAs than all other major groupings combined; this difference is designated by the “A” Duncan grouping. Finally, the “B” group, when combined, possesses significantly lower GPAs than business alone, and has higher GPAs than statistics alone. The Duncan test is a valid comparison of means but invites Type I errors that incorrectly reject true hypothesis (Berry & Hochberg, 1999).

Methods Summary

The literature review suggests that demographic variables, measures of ability, and actual levels of participation are explanatory in models predicting an individual’s ability to accurately retrieve and report facts (Kuncel et al., 2005). There is a need in the literature to predict self-reported validity based upon other respondent traits, including MCSDS score, Greek organization membership, residential LLV involvement, STEM majors, student major certainty, household size, family income, and participation in athletically oriented activities like club sports, intramural sports, and formal fitness programs. This section reviewed how the proposed methods provide answers to the research questions.

Limitations, Ethical Considerations, and Other Concerns

Limitations

This is a single institution study that tests the validity of a narrow segment of self-reported student behaviors on undergraduate students. This is an attempt to replicate typical survey questions used in studies on college student academic and social participation. However, questions on CRF use are usually limited to an uncommon niche of behavior based around CRF needs assessments and campus recreation program assessments. Many researchers study academic engagement and class skipping rates, but HESF courses are atypical of the academic experiences researchers are typically describing in more intellectually driven curricula. As such, this study will forward the field, but the methodology and data limit the generalizability to the academic disciplines and co-curricular behaviors commonly reported in national and local surveys.

The saliency of using HESF courses as a substitute for other academic subjects is debatable, but HESF courses remain the best available to measure class attendance across large groups at this institution. Further study may use computer login data for courses that meet in computer labs or building entry swipes to study class meetings among other disciplines housed in academic buildings.

Results come with limited generalizability to other settings because NCSU is a selective institution with students who have an unusually high concentration of STEM oriented majors and high rates of participation in the traditional residential college experience. The study of social and academic behavior is applicable to all types of students that NCSU cannot replicate in large numbers. Further study should expand this methodology to other institutions.

Ethical Considerations

Maintaining ethical interactions with subjects and protection of the students' personal data are my primary ethical considerations. All student interactions conformed to IRB standards and all subject interactions occurred after obtaining IRB approval. Participation was voluntary and unconditional; if the questions or methods made the respondents uncomfortable, they could decline to answer without cause for hesitation. The obfuscation of student identities in the data sets mitigates ethical concerns about privacy invasion. It is important that the data remain private because the validation data includes detailed descriptions of grades and time stamped information about a given student's whereabouts. Protection of personal information was assured by stripping data of personally identifiable information and by storing the data on secure, password-protected servers and devices.

Other Concerns

Student data provided by the institution is presumed accurate but may describe instances of non-students borrowing student identification cards to enter the CRF at rates that are unknown. The CRF student employees are required to adhere to CRF policies and carefully check the photo identification to the individual presenting the student identification card. However, student worker fidelity to this policy is unknown.

Prior research shows that students' ability to accurately self-report developmental gains vary by the topic (Bowman, 2011). It follows that students interested in some topics may more accurately self-report behaviors about that topic. This study does not access information about interest in exercise and recreation, so an explanatory variable may be omitted from the study. Similarly, findings from this study about CRF behaviors may not generalize to data on other categories of self-reported behavior.

Chapter Summary

This chapter described the research design, specified research questions intended to be answered, established the study sample creation, described the survey instrument and data requested for analysis, and concluded with a discussion of data concerns, opportunities, and limitations. This research design is the best available to help assess the validity of survey instruments among bachelor's level students at NCSU.

CHAPTER 4: RESULTS

The cost of conducting online surveys continues to decline, causing researchers to expand their reliance upon self-reported data for the study of college student behaviors. This reliance articulates an unmet need to investigate the validity of the self-reported data. This study tests the validity of self-reported behavioral data by seeking (1) response bias or systematically inaccurate descriptions of behavior and (2) nonresponse bias or systematic differences between respondents and nonrespondents. The presence of bias has implications for all college student research that uses self-reported data. My results suggest that self-reported data cannot be presumed a valid estimate of actual behavior. Chapter 4 describes the survey instrument and tests the survey statistic as a valid estimator of behavior across the respondents, sample, and target population by comparing self-reported data to the corresponding institution reported values. In this chapter, I describe the survey process and then evaluate each research question to test for the presence of bias and explore the social desirability of question topics as a potential source.

Sample and Respondent Summary

The Office of Institutional Research and Planning (OIRP) at North Carolina State University (NC State) provided a sample of 6,000 students drawn from the target population of 20,775 bachelor's degree level students. The sample was randomly selected from the population of students who were (1) enrolled on the tenth day of classes in spring 2016, (2) first enrolled at NC State after summer 2006, and (3) were not included in the 18 student survey question pilot study. On March 14, 2016, members of the sample received an email invitation to participate in an online survey available from March 14 to 24. Follow up reminders were emailed on March 21 and March 23. All responses – including incomplete and broken off surveys – were frozen on March 25. The online survey totaled 27 questions with 25 closed-ended questions and 2 open-ended questions. Overall, 25.5% of the sample partially or fully responded to the survey request.

The 6,000 student sample and 20,775 student target population averages are almost identical for demographics, academic success, co-curricular participation and college preparation. Means and standard deviations are nearly identical except that the sample contains slightly more ethnically white students (73.5% versus 73.1%). The typical student in the sample and population entered the university as a first time freshman, is a 21-year-old white male, and has been enrolled for 4.9 fall/spring semesters (i.e., a mid-year junior). The average student is academically prepared with a 1233 SAT score and also is academically successful with a 3.18 cumulative GPA. Most students in the study are not Pell Grant eligible.

The response rate substantially exceeded the number required to reach meaningful conclusions about the validity of survey responses. To obtain satisfactory confidence in the research conclusions, I required at least 299 responses to each question. This threshold was easily eclipsed since between 487 and 1,340 members of the sample responded to each question on the survey. Overall, 25.5% of the sample answered at least one question, and 19.8% of the sample answered every required question. Among closed-ended questions, the response rates to all questions exceeds 20.7%; 17.3% responded to the first open-ended question; and 9.2% responded to the second open-ended question. Compared to recent surveys conducted by researchers at this institution, this study's response rate is atypically high. Though the reason for the high response rate is unknown, the university's survey staff believe that my student-to-student email and personal appeal successfully motivated students to respond to my survey request. Email prompts successfully motivated responses because I received 46.1% of responses after the first email prompt, 27.2% after the second, and 14.3% after the third. The remaining 12.4% of responses were completed on a day without an email reminder. The survey request claimed that the survey could be completed in 7 minutes or less, and 80.0% did so. Another 13.3% of respondents did so in 7 to 20 minutes, and the remaining 6.7% of respondents broke off or edited their responses well after their initial engagement.

The 1,534 respondents and 4,566 demographic distributions were similar with the exception of gender, but average GPAs and SAT scores of respondents were higher than those of nonrespondents. Respondents were better prepared for college and academically more successful, earning higher average cumulative GPAs of 3.28 (versus 3.14) and SAT scores of 1258 (versus 1224). Females were more likely to respond as 28.5% of the women in the sample completed a survey versus 22.9% of men. Though not significant or substantively different, the respondents were more likely to be Whites or Hispanics, have entered the university as first-time freshmen and not have received a Pell Grant. The average age and number of semesters enrolled were identical.

Institution Reported Data Summary

The survey includes 14 questions for which I have corresponding institution reported data across all 20,775 members of the target population. By matching the self-reported value to the institution reported value, I assessed the accuracy of self-reported values by comparing differences. However, the precision of self-reported values collected via the survey instrument was impaired by forcing the complexity of all behaviors into limited response options to make the survey instrument manageable for students. The reduced precision complicated the validation process because self-reported values are simplified representations of behavior, and institution reported values remain direct measures. To accommodate the analysis, institution reported scales were modified to match the response options used in the survey instrument to enable valid comparisons of self-reported and institution reported values. A clarification of how the self-reported values compared to institution reported values follows.

Question 1A: Prior 7 Days of CRF Use

The prior 7 days of CRF use is a continuous value ranging from "0" to "20" for self-reports and institution reports. A correct report is an exact match between the self-reported and institution reported value. Over reporting is defined as a self-reported value greater than the institution reported value.

Question 1C: Prior Academic Year of CRF Use

The self-reported CRF use over the academic year is defined as a rate-based variable with five response options, and the institution reported value is the number of visits divided by number of weeks in the year. The response option "I never entered the Carmichael Recreation Complex" is accurate when the respondent never entered the CRF in the 2015-16 academic year; "Less than once per week" is accurate for a weekly rate greater than 0.00 and less than 0.50; "Once per week" is accurate for a weekly rate greater than or equal to 0.50 and less than 1.5; "Twice per week" is accurate for a weekly rate greater than or equal to 1.50

and less than 2.5; and “Three or more times per week” is accurate for a weekly rate greater than or equal to 2.50. Over reporting occurs when the self-reported range exceeds the institution-reported range.

Question 2: Cumulative GPA

The cumulative grade point average is a continuous value between 0.00 and 4.00 for self-reported and institution reported values. For some analysis, a correct report is a match rounded to the one hundredth of a grade point, and other analysis requires a match to the tenth of a grade point. Over reporting is defined as self-reported GPA higher than institution reported.

Question 3: Fall 2015 Semester GPA

The semester grade point average is a continuous value between 0.00 and 4.33 for self-reported and institution reported values. A correct report is a match between the self-reported and institution reported value. For some analysis, a correct report is a match rounded to the one hundredth of a grade point, and other analysis requires a match to the tenth of a grade point. Over reporting is defined as self-reported GPA higher than institution reported.

Question 4: Total D/F/Unsatisfactory Grades

The number of low grades is a continuous value ranging from “0” to “5 or more” for self-reported and institution reported value. A correct report is an exact match between the self-reported and institution reported numbers. Over reporting is defined as a self-reported value less than the institution reported value.

Question 5: Total A Grades Earned

The number of high grades is a continuous value ranging from "0" to "20 or more" for self-reported and institution reported values. A correct report is an exact match between the self-reported and institution reported value. Over reporting is defined as a self-reported value greater than the institution reported value.

Question 6: Total PE/HES Courses Taken

The number of courses taken is a continuous value ranging from "0" to "4 or more" for self-reported and institution reports. A correct report is an exact match between the self-reported and institution reported values. Over reporting is defined as a self-reported value greater than the institution reported value.

Question 7: PE/HESF 100 Level Course Taken

For self-reports with corresponding institution reported values, the PE/HESF 100 level course taken is a text value selected from a drop down list of all such courses as described by course catalog number and name. There are 14 response options in the self-reported and institution reported values. A correct report is an exact match between self-reports and institution reports. Misreporting is labeled as "other misreport" and is not defined as over or under reported value that implies socially desirable superiority among response options.

Question 8: Grade Earned in PE/HESF 100 Level Course

For the self-reported and institution reported values, the grade earned in the PE/HESF 100 level course is an ordinal value ranging from "F/Unsatisfactory" to "A-/A/A+" with an additional "Satisfactory" response option. Plus/minus letter grade response options combine into a single response option; for example, “A-“rolls into “A-/A/A+”. “Satisfactory” is a passing grade without grade point values used for other letter grades; the lack of grade points prevents a “Satisfactory” grade from implied socially desirable superiority to other passing letter grades from “C-“ to “A+”. For letter grades, a correct report is an exact match between self-reported and institution reported grades. Over reporting is defined as a self-reported value greater than the institution reported value. For the “Satisfactory” grade, the self-reported value is over reporting when the actual value is an “Unsatisfactory”, “F”, “D-“, “D”, or “D+”; in instances where the institution reported grade is a passing letter grade, it is considered “other misreport” because the relative seniority of the grades cannot be determined.

Question 9: PE/HESF Class Skips

The number of class skips is a continuous variable ranging from “0” to “7 or more.” A correct report is an exact match between the self-reported value and institution reported value. Over reporting is defined as fewer self-reported class skips than the institution reported value. Other misreporting describes a mismatch where the course is not correctly identified as a distance education or on campus section. To count class skips, this study’s methodology requires that a section occur inside the CRF with a standard meeting time. This condition is untrue of distance education sections.

Question 10A, 10B, 10C, and 10D: Co-Curricular Participation

The co-curricular participation rate response options exist on a four point ordinal scale ranging from “I have never participated” to “In all semesters.” A correct report is an exact match between the self-reported value and the institution reported value. Over reporting is defined as a participation rate greater than the institution reported rate.

For each respondent, the summed accuracy of their self-reports is described as the percentage of questions answered correctly, percentage over reported, and percentage underreported. Rather than count the number of questions, I used the percentage of questions because the numbers of gradable questions differ by the respondent. The number of usable questions differs when (1) their lack of enrollment in a PE/HESF 100 level course results in skipping past questions about course taking behavior; and/or (2) misstated self-reported values do not possess socially desirable seniority over institution reported values and cannot be classified as over or under reported; and/or (3) the respondent is randomly selected to answer the vaguely worded question 1B whose response options cannot be evaluated. Respondents must answer all offered behavioral questions to be assigned an accuracy percentage with partially completing respondents excluded from pertinent analysis.

Systematically and across most questions, the survey mean exceeds the institution reported mean in a socially desirable direction. For example, the survey yields an average fall 2015 semester GPA of 3.34, and the institution reported mean is significantly lower at 3.05. Similarly, respondents self-reported entering the CRF an average of 1.80 times over the prior 7 days, but the corresponding institution reported mean is only 0.59. The only exception to the pattern of biased self-reports in a socially desirable direction is the number of A-/A/A+ grades earned over a student's career at NC State and some co-curricular participation metrics. The differences between the self-reported mean to the institution-reported mean of respondents, sample, and target population justify further investigation to estimate and explain the differences.

I imputed survey completion data for temporally sensitive questions for which the survey completion date is required to collect validation data for survey non-completers that include nonrespondents and students outside the sample. For many questions, but especially question 1A, a survey completion time is particularly important because CRF use rates are affected by the university's spring holiday during which time most students do not enter, but a minority of users use free time for physical activity. Surveys completed between March 14 and March 20 include an atypical behavior because the reference period includes the spring university holiday. However, for survey non-completers, there is no survey completion date to define institution-reported computations. As such, I imputed a survey completion date for non-completers using a distribution of actual survey submission times. First, I computed the likelihood that a survey would be submitted at any given time using the distribution of actual submissions. Next, a random number generator assigned survey completion times to non-completers such that the distribution of imputed completion times mirrored that of the respondents' actual completion times. Finally, I used the imputed completion time to compute institution reported validation data for temporally sensitive questions. Among respondents, 53.7% of respondents completed their survey between March 14 and March 20; for 52.8% of nonrespondents and 53.3% of the population, the imputed survey completion dates were from this period. The imputed survey completion date allows an unbiased comparison of respondents and nonrespondents.

Review of Research Questions

This validation study compares self-reported behaviors to institution reported data to assess the accuracy of self-reports and the validity of self-reported college student behavior data using a series of questions about CRF use, academic performance, and co-curricular participation. The first research question analyzes sample and respondent trends to establish the widespread presence of both measurement and nonresponse errors in self-reported data. Error from RQ1 suggests that self-reported behavior requires scrutiny before the survey statistic can be accepted as valid. The most accurate self-reporters are high GPA students and those with less complex academic histories. The second research question explores the relationship between the topic of a survey question and both the social desirability bias and changes in nonresponse rates. Results from RQ2 suggest that potential respondents who actually possess admirable attributes are more likely to respond to a survey request in which the respondent can describe them. The MCSDS is theorized to predict the propensity for self-monitoring and helps researchers identify which respondents are more likely to inflate their self-reports in a socially desirable direction, but this study found no evidence that the MCSDS is predictive of differences in response behavior. This section will summarize data in the context of research questions to determine if self-reported means are a biased estimate of behavior across the sample and target population.

Research Question 1

To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?

Research question 1 reviews the validity of using the self-reported survey statistic as an estimate of actual behaviors across the pool of respondents, the sample, and the target population. This question is assessed first from the perspective of measurement error from the respondents by evaluating their reports and determining which respondent characteristics are associated accurately in reporting. Next, I determined how the proximity of behavior and other temporal factors affect self-reported behaviors. Finally, I assessed the role of nonresponse error and how response accuracy is affected by the response propensity of students in the sample. At this section's end, the person-based validity will show self-reported data as a biased estimator of actual behavior.

Research Question 1A

What student characteristics are associated with the most accurate and inaccurate self-reporters? Do demographic or background characteristics affect accuracy in self-reporting? Do students with high collegiate GPAs or standardized test scores more accurately self-report behaviors? Do participatory students provide more accurate self-reported behavioral data?

This research question evaluates the presence of response bias in self-reported data of college students. Response bias is a persistent threat to the validity of self-reported behavioral data collected in this study. Most self-reported means have statistically significant differences from their corresponding actual means, and the difference is often substantive. Table 4.1 shows that for the 13 questions with validation data, 11 have self-reported means that are significantly different from the corresponding institution reported mean at the .0001 level, and for 7 questions, the means differ by at least 20%. Questions about exercise – both informal and occurring within an academic class – are the most sensitive to measurement error with social desirability bias causing respondents to systematically over report physical activity. For example, the prior 7 days of CRF use is inflated by 170%. Respondents also systematically underreported the number of high/low grades earned and their co-curricular participation rates because they failed to recall all of the instances over their academic career. For example, respondents frequently failed to recall Outdoor Adventure participation because the one-day programs are easily forgotten. Finally, measurement error affected questions about co-curricular participation and were more complex than I initially believed due to colloquial terms that describe Campus Recreation activities that are frequently unfamiliar to respondents. For example, "club sports" describe inter institutional athletic competitions that respondents frequently confuse with intra institutional "intramural athletics." This research question explores unit-level attributes that explain a given respondent's likelihood to systematically misreport.

Respondent demographics. High GPA and newly enrolled students are more likely to provide accurate self-reported behaviors. Though other demographic and background measures are correlated with response accuracy, regression and correlation analysis reveal that measure is also collinear with academic career length. For example, respondents 19 and younger answer 68.7% of questions accurately compared to 60.5% for respondents aged 20 to 25. However, when an OLS regression controls for years of experience, the relationship is not significant. Subsequent analysis, however, finds that age is lowly correlated with response accuracy. Younger respondents are predominately first year students who more accurately describe their collegiate behavior because their reference period is shorter (e.g., the prior 8 months). Meanwhile, older students are more likely to be late-career students with reference periods spanning multiple academic years, and past behaviors more taxing to recall and articulate. First year respondents accurately respond to 71.8% of questions asked, and respondents enrolled four or more years are only 59.3% accurate. High GPA respondents accurately respond to 74.4% of questions, and low GPA respondents only respond with 56.7% accuracy.

Table 4.1.

T test on Unit Level Self-Reported Mean Equal to Institution Reported Mean.*

Question	N	SR Mean	IR Mean	Diff Mean	Pct. Bias	Std Dev	T Stat
Q1A: CRF Use - Prior 7 Days	487	1.81	0.67	1.138	170%	2.0736	12.11
Q1C: CRF Use - Weekly Rate	503	1.65	1.00	0.653	65%	1.0284	14.25
Q2: Cumulative GPA	1318	3.34	3.31	0.029	1%	0.0048	6.33
Q3: Semester GPA	1292	3.36	3.23	0.135	4%	0.5151	9.40
Q4: D/F/U Grades Earned	1340	0.60	1.02	-0.416	-41%	0.0440	-9.46
Q5: A Grades Earned	1307	7.42	10.05	-2.631	-26%	5.5189	-17.23
Q6: HESF Course Taken	1316	1.35	1.57	-0.219	-14%	1.0218	-7.77
Q8: HESF Grade	110	3.55	3.49	0.055	2%	0.3792	1.51
Q9: HESF Class Skipping	526	1.87	2.61	-0.736	-28%	1.9906	-8.48
Q10A: Intramural Sports Partic.	1253	0.73	0.68	0.0145	2%	0.5373	2.94
Q10B: Fitness Program Partic.	1256	0.33	0.34	-0.006	-2%	0.6027	-0.37
Q10C: Club Sports Partic.	1262	0.80	0.65	0.150	23%	0.7612	6.99
Q10D: Outdoor Adventures Partic.	1241	0.18	0.30	-0.116	-39%	0.7531	-5.43

†limited to instances where numeric self-reported and institution reported data is available
Note. Significant differences indicate respondents do not accurately self-report behaviors.
**p*<.05, ***p*<.01, ****p*<.001.

Demographics correlated to age, ethnicity, gender, socio-economic status, and other demographic measures possess a low correlation with response accuracy after controlling for GPA. For example, women have higher GPAs than men do, and there is not a statistically significant, gender-based difference in response accuracy after the relationship is moderated by GPA. Researchers can only turn to cumulative GPA and length of academic career when assessing the credibility of responses and must ignore other respondent attributes.

A final background measure is the complexity of a student's actual behaviors. Consistent with survey research theory, respondents describing simpler behaviors are more likely to describe their behaviors accurately because the survey request is less mentally taxing. For example, respondents who never participated in a CRF-based, co-curricular activity easily self-report their abstinence; whereas, participants must both recall their past behavior and frame it within response options. Interest in an activity of study may also motivate respondents to inflate their participation rate. Nonparticipants in CRF-based, co-curricular behaviors self-report with 67% accuracy; respondents actually participating in 1 such activity self-reported with 63% accuracy; respondents actually participating in 2 such activity self-reported with 57% accuracy; and respondents actually participating in 3 or more such activity self-reported with 52% accuracy. Meanwhile, the overall participatory nature of a student is unrelated to his or her ability to self-report accurately. Holding other factors constant, students involved in Greek Life and Living Learning Villages report their CRF use with similar accuracy as non-participating students. High intensity participants in co-curricular activities are less likely to accurately report their behaviors in that activity because they describe more complex behaviors.

Student cumulative grade point average. The characteristics of a successful student are highly aligned with those of an ideal survey respondent. As such, I hypothesize that a high GPA student is less likely to satisfice and more likely to provide accurate responses, which leads to unbiased self-reported behavior, even after controlling for ability. Conversely, low GPA students more frequently misestimate and satisfice, both of which lead to misreporting in a socially biased direction because thoughtless survey responses are more likely to be biased in a socially desirable direction. Other data supports the hypothesis that cumulative GPA is predictive of satisficing and the likelihood to provide an accurate response. Low GPA students, for example, complete the survey in 3.5 minutes or less, and high GPA students show diverse completion times between 3 and 7 minutes. A student's cumulative grade point average describes academic success in classroom assignments that measure a student's ability to exert effort to prove consistent with the instructor's intent.

For each question, Table 4.2 shows the average cumulative GPA for those who (1) underreported, (2) correctly reported, (3) over reported, and (4) misreported but without a clear relative position to the correct report. For each question, I performed a Duncan Test to determine significant differences in cumulative GPAs between the reporting accuracy groups; in cases where a significant difference was detected, the value is bolded, and the Duncan grouping is provided in the parenthesis. In cases where the Duncan grouping letter is present in more than one reporting accuracy group, those groups can be combined to show significant differences from all other groups. For example, for "Q1A: CRF Use – Prior 7 Days" combining underreports and correct reporters together into a single group results in a mean that is significantly different from that of over reports. Simultaneously, if the groups were reordered by combining correct reporters and over reporters into a single group, its mean would be significantly different from that of under reporters. For Q1A, Duncan groupings show that under reporters and over reporters possess significantly different GPAs from each other while correct reporters, alone, are not significantly different from any group.

High GPA students are more likely to correctly or underreport in questions about CRF use (Q1A and Q1C), overall academic performance (Q2, Q3, Q4, Q5, Q6, Q7, and Q8), and co-curricular participation (Q10A, Q10B, Q10C, and Q10D). Respondents correctly reporting have the highest GPA in seven questions and under reporters have the highest GPA on six questions. The over reporters have the highest GPA only in question 9, which describes the number of class meeting skipped. High GPA students are more accurate self-reporters of their cumulative GPA, as 72.2% of the top quintile of cumulative GPA respondents accurately described their cumulative GPA to the hundredth of a grade point, but only 18.1% in the bottom quintile did so. High GPA students are also more likely to recall and count other behaviors accurately and, when they misreport, underestimate their past behaviors and misreport in a socially undesirable direction. Meanwhile, low GPA students are more likely to misreport behaviors in a socially desirable direction. Researchers can look to high GPA students for more accurate and less biased self-reported data.

Table 4.3 shows the over reporting to underreporting ratios for each question split into categories for the cumulative GPA of respondents. All questions were oriented such that the over reporting is in a socially desirable direction; thus, the higher the ratio, the more systematically behavior is misreported. Unbiased self-reporting would be equal to one, and systematic underreporting would approach zero. In all but two questions, the highest GPA respondents have the lowest ratio, suggesting that high GPA students are less likely to satisfice and are better able to recall behavior accurately. Meanwhile, in seven questions, the low GPA have the highest ratios. The ability and effort combine to create more accurate and less biased responses. These ratios provide more evidence that high GPA students provide more accurate self-reported, behavioral data.

Table 4.2.

Cumulative GPA by Question and Reporting Accuracy.

	Under Reporters	Correct Reporters	Over Reporters	Other Misreports
Question	3.37	3.30	3.28	
Q1A: CRF Use - Prior 7 Days	**3.41(A)**	**3.28(A,B)**	**3.22(B)**	
Q1C: CRF Use - Weekly Rate	**3.28(B)**	**3.50(A)**	**3.05(C)**	
Q2: Cumulative GPA	**3.29(B)**	**3.57(A)**	**3.12(C)**	
Q3: Semester GPA	**2.77(C)**	**3.42(A)**	**2.98(B)**	
Q4: D/F/U Grades Earned	**3.30(B)**	**3.41(A)**	**3.13(C)**	
Q5: A Grades Earned	**3.29(A,B)**	**3.32(A)**	**3.17(B)**	
Q6: HESF Course Taken		3.31		3.23
Q8: HESF Grade	3.12	3.37	3.03	3.12
Q9: HESF Class Skipping	3.30	3.36	3.38	3.27
Q10A: Intramural Sports Partic.	3.40	3.29	3.32	
Q10B: Fitness Program Partic.	3.36	3.30	3.33	
Q10C: Club Sports Partic.	3.34	3.32	3.28	
Q10D: Outdoor Adventures Partic.	**3.36(A)**	**3.31(A)**	**3.18(B)**	

Note. Bold denotes Duncan Test intra-question statistically significant mean differences (P<.05). Means with different letters are significantly different from means in other letter groups. "Other Misreports" describe incorrect self-reports that lack a numeric relationship and cannot be classified as under or over institution reports

Figure 4.1 tracks an individual's response behavior across all of their responses to show how GPA, as a respondent attribute, affects the same person's response tendencies across all questions. Overall, the rate of underreporting behaviors is unaffected by GPA, though high GPA respondents are more likely to correctly report. Students with a GPA around 2.5 are the most likely to over report behaviors in a socially desirable direction, and each subsequent GPA category is less likely to over report than the one before. Consistent with the hypothesis that GPA is predictive of response accuracy, the inner quartile boxes narrow and shift downward as GPA increases. NC State's course repeat policy causes low GPA respondents to possess a more complex, harder to describe academic history.

Table 4.3.

Over Reporting to Underreporting Ratio by Respondents' Cumulative GPA.

Question	GPA 0.00 - 2.49	GPA 2.50 - 3.49	GPA 3.50 - 4.00
Q1A: CRF Use - Prior 7 Days	17.0	13.6	10.9
Q1C: CRF Use - Weekly Rate	33.0	7.4	4.8
Q2: Cumulative GPA	3.9	1.4	0.8
Q3: Semester GPA	3.8	2.0	1.4
Q4: D/F/U Grades Earned	7.4	7.8	3.5
Q5: A Grades Earned	1.9	0.3	0.4
Q6: HESF Course Taken	0.5	0.5	0.2
Q8: HESF Grade		8.3	
Q9: HESF Class Skipping	0.4	0.6	0.7
Q10A: Intramural Sports Partic.	10.0	10.4	6.5
Q10B: Fitness Program Partic.	1.0	1.7	1.0
Q10C: Club Sports Partic.	4.4	4.5	4.2
Q10D: Outdoor Adventures Partic.	1.3	0.5	0.4

Note. Unbiased reports have a ratio of 1.0; ratios above 1.0 suggest systematic misreporting in a socially desirable direction while ratios under 1.0 suggest systematic underreporting

Accordingly, I performed the same analysis only on questions about CRF use and co-curricular participation to find similar results with lower levels of statistical confidence. Simply put, better students are more accurate survey takers, and worse students are less accurate survey takers.

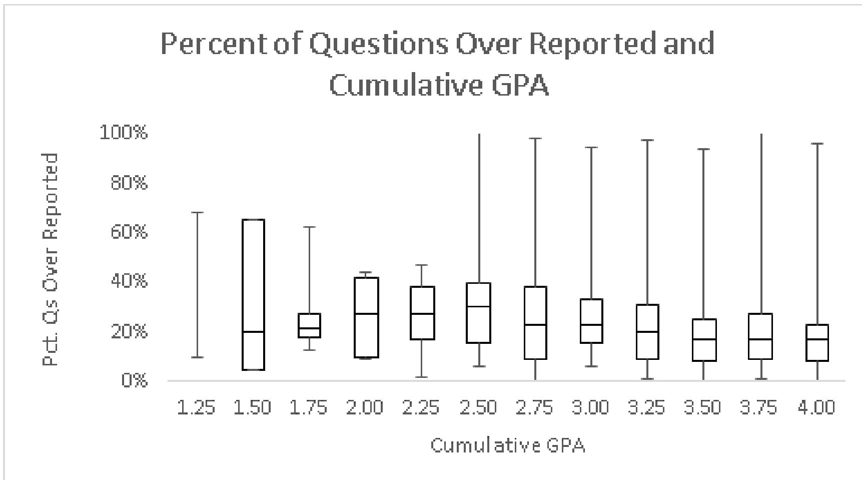

Figure 4.1. Percent of questions over reported and GPA.

I considered several hypothesizes for why high GPA students are more accurate self-reporters than low GPA students. First is the theory of social desirability bias where students inflate their behaviors to portray themselves as more successful than is true of an authentic response. Only low GPA students can systematically inflate their performance to mimic academically successful students who are stuck on the high point response option scale. This explanation is negated by similar trends observed with co-curricular participation behaviors that are not associated with academic success. A second explanation is that high GPA students, on average, are further into their academic careers because they persist longer with fewer discouraged by the consequence of poor collegiate performance and/or affected by suspension policies that remove them from the population before their career can lengthen. Recall is more taxing for later career students because their academic records are more complex and possess longer reference periods. A cross tabulation of student career length and self-reporting accuracy provides little evidence for this postulation; across all lengths of enrollment, high GPA respondents are more accurate self-reporters, and the correlation between career length and GPA is 0.09 and insignificant. A third explanation is an institution specific grade replacement policy that complicates GPA calculations. Students earning a low grade in a given course are eligible to apply for grade replacement, which drops the low grade from GPA calculation upon completion of a suitable replacement course in the following semester. This explanation is also negated because low GPA students offer less accurate responses for non-academic questions, as well. A final explanation is that high GPA students are also more accurate self-reporters. Exam grades describe a student's ability to comprehend questions, accurately recall facts, and organize the information logically, thus describing course content consistent with the instructor's intentions. Because the test taking and survey taking processes are similar, the characteristics of a successful exam taker are highly aligned with those of an ideal survey respondent with high grades, thus predicting higher likelihood of providing accurate survey responses. I note that GPA correlates with student descriptors like satisfaction with social networks, satisfaction with NCSU attendance, etc. that more directly measure the likelihood to provide accurate responses and future research should seek evidence of such relationships.

SAT score. Respondents with higher cognitive ability are expected to provide more accurate self-reported data because they are better equipped to recall past behaviors and organize the information into the survey's response options. Cognitive ability is also an influential input of a student's cumulative GPA, leading me to consider GPA as a measure of ability in addition to their probability of satisficing. A student's SAT score describes their academic ability, though predicted academic success neglects to measure the student's willingness to engage in the academic curriculum or exert sufficient effort. Cumulative GPA does describe the latter attributes, making it a distinctly different measure than SAT scores. Because cumulative GPA is predictive of response accuracy but SAT score is not, I argue that the cumulative GPA measures respondent attributes that are also predictive of satisficing behavior.

Table 4.4 shows the average SAT score for those who (1) underreported, (2) correctly reported, (3) over reported, and (4) misreported but without a clear relative position to a correct report. Unlike the corresponding cumulative GPA table, there is not a predominant relationship between SAT score and reporting accuracy. Of the 7 academic performance questions (Q2, Q3, Q4, Q5, Q6, Q7, and Q8), high SAT score students are more likely to correctly report in 4; meanwhile, there is no discernible relationship between SAT score and the likelihood to underreport, correctly report, or over report in non-academic performance questions. High SAT score students are more likely to be high GPA students who accurately self-reported academics, but the GPA-SAT score relationship does not extend to other topic areas. Outside of academic measures, there is not a systematic relationship between over reporting responses and lower SAT scores.

Among respondents with both a collegiate GPA and SAT score on record, Figure 4.2 shows the percentage of all survey questions correctly reported and sorted by cumulative GPA (left) and SAT score (right). A comparison of median point slopes shows that the GPA is more predictive of reporting accuracy; high GPA students are less likely to over report in a socially desirable direction. Meanwhile, there are no systematic differences in over reporting based upon SAT score. A significant relationship in Figure 4.2 should see the inner quartile boxes narrow and shift upward towards more survey questions accurately answered. This relationship is not observed for the SAT scores but is for the cumulative GPA. This relationship is also consistent among only CRF use and co-curricular participation question Table 4.4.

SAT Score by Question and Reporting Accuracy.

Question	Under Reporters	Correct Reporters	Over Reporters	Other Misreporters
Q1A: CRF Use - Prior 7 Days	1270	1261	1284	
Q1C: CRF Use - Weekly Rate	1237	1255	1246	
Q2: Cumulative GPA	**1244(B)**	**1288(A)**	**1230(B)**	
Q3: Semester GPA	**1242(B)**	**1296(A)**	**1243(B)**	
Q4: D/F/U Grades Earned	**1215(B)**	**1268(A)**	**1217(B)**	
Q5: A Grades Earned	**1259(A)**	**1275(A)**	**1230(B)**	
Q6: HESF Course Taken	1251	1261	1258	
Q8: HESF Grade	1200	1268	1234	1238
Q9: HESF Class Skipping	1259	1282	1264	1252
Q10A: Intramural Sports Partic.	**1215(B)**	**1255(A)**	**1275(A)**	
Q10B: Fitness Program Partic.	1275	1254	1283	
Q10C: Club Sports Partic.	1246	1261	1257	
Q10D: Outdoor Adventures Partic.	**1283(A)**	**1256(A,B)**	**1240(B)**	

Note. Bold denotes Duncan Test intra-question statistically significant mean differences (P<.05). Means with different letters are significantly different from means in other letter groups. "Other Misreports" describe incorrect self-reports that lack a numeric relationship and cannot be classified as under or over institution reports

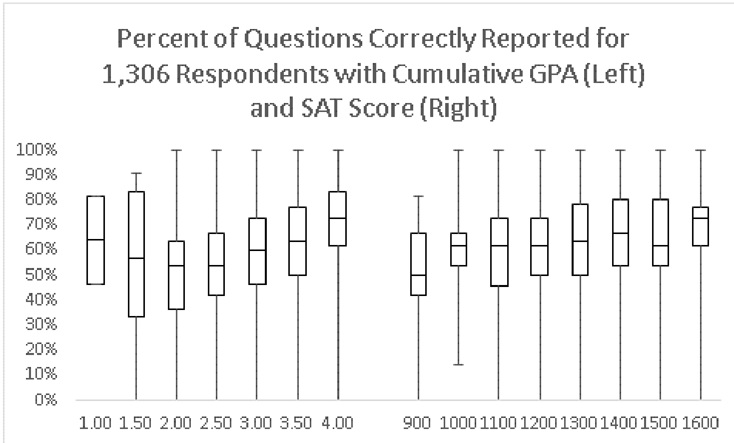

Figure 4.2. Percentage of questions correctly reported by GPA and SAT score.

topics. Though the SAT score captures the potential for academic success, it does not describe intangible qualities about how seriously a student takes academic responsibilities, how well they handle complex and ambiguous questions, or how well they can focus to complete tasks accurately. These are all measures that are also predictive of satisficing, which a collegiate GPA captures well, and the SAT scores does not. A respondent's cumulative GPA measures their credibility with attributes not described by ability, as measured by SAT score.

For the same respondents with both a collegiate GPA and SAT score on record, Table 4.5 regresses the percentage of all survey questions correctly reported on background variables to determine if GPA is a more powerful predictor of response accuracy across four models. First, I ran a regression analysis on SAT score only (Model 1), then cumulative GPA only (Model 2), next both cumulative GPA and SAT score (Model 3), and finally on SAT neither score or cumulative GPA (Model 4). Consistent with earlier conclusions, the cumulative GPA is a strong predictor of response accuracy; whereas, SAT score is less so. Without including the cumulative GPA (Model 1), SAT score is a significant predictor where a 100-point SAT score increase is associated with a 2.0 percentage point increase in overall response accuracy. Meanwhile, a 2.0 percentage point increase in GPA is associated with a .17 grade point increase. Model 2 has the highest adjusted R-squared and is 79% higher than Model 1 using SAT score only. As measured by adjusted R-squared, including both the SAT score and cumulative GPA (Model 3) reduces the strength of the model. For the SAT score only model (Model 1), gender becomes significant because women have significantly higher GPAs than men, and the collinear relationship between gender and GPA results in measurement error. Regression analysis confirms that cumulative GPA is a valid predictor of response accuracy, though it should be noted that the truncated variance of the SAT scores limit the generalizability of these findings to lower ranges. NCSU is a selective institution, and nearly all students possess SAT scores significantly higher than average. As such, a similar analysis among populations with heterogeneous SAT scores could yield different results.

Table 4.5.

Regression on Response Accuracy by SAT Score and Cumulative GPA

	Model 1: SAT Score Only	Model 2: Cum. GPA Only	Model 3: SAT & Cum. GPA	Model 4: No GPA or SAT
Intercept	**0.427***** **(0.1613)**	**0.273***** **(0.1373)**	**0.219***** **(0.1579)**	**0.803***** **(0.1313)**
Female	**0.034**** **(0.0122)**	0.015 (0.0117)	0.016 (0.0120)	**0.026*** **(0.0121)**
African American	-0.013 (0.0509)	-0.010 (0.0490)	-0.006 (0.0493)	-0.039 (0.0507)
Asian	-0.072 (0.0510)	-0.072 (0.0494)	-0.071 (0.0494)	-0.076 (0.0513)
Hispanic	0.039 (0.0506)	0.035 (0.0490)	0.036 (0.0491)	0.028 (0.0509)
International	-0.012 (0.0749)	-0.069 (0.0725)	-0.065 (0.0728)	-0.029 (0.0752)
Two Plus Races	-0.014 (0.0512)	-0.010 (0.0495)	-0.009 (0.0497)	-0.021 (0.0514)
White	0.012 (0.0434)	0.012 (0.0421)	0.013 (0.0421)	0.006 (0.0437)
Semesters Enrolled	-0.003 (0.0045)	-0.002 (0.0043)	-0.002 (0.0043)	-0.001 (0.0045)
Age	-0.001 (0.0070)	0.001 (0.0067)	0.002 (0.0068)	-0.007 (0.0069)
Number of Co-Curricular Activities	**-0.043***** **(0.0056)**	**-0.047***** **(0.0054)**	**-0.047***** **(0.0054)**	**-0.041***** **(0.0056)**
Cumulative GPA		**0.114***** **(0.0115)**	**0.111***** **(0.0123)**	
SAT Score	**0.000***** **(0.0001)**		0.000 (0.0001)	

Note. OLS Regression. N=1248. Adjusted R-squared for model 1 is 0.0724; for model 2 0.1297; for model 3 0.1293; and for model 4 0.0614.

**p*<.05, ***p*<.01, ****p*<.001.

Satisficing behaviors. Students with high cumulative GPAs provide more accurate and unbiased behavior reports than students with low GPAs. I argue that cumulative GPA is a measure, on average, of both ability and effort exerted when responding to this survey. This section establishes a relationship between cumulative GPA and two measures of satisficing: time taken to complete the survey and length of the open-ended question response. For most respondents, this analysis shows that a thorough and thoughtful response requires time to fully comprehend the question, work through their recall and estimation strategy, and then to consider the response options. However, students in the top quintile of cumulative GPA are able to provide the accurate responses using less time; high cognitive ability enables these respondents to quickly and accurately navigate the response process. Accordingly, time taken to complete the survey measures the likelihood that a respondent is providing a thorough and thoughtful response that must be moderated by cognitive ability. This survey concludes with two optional, open-ended questions, so respondents providing verbiage are likely to be exerting more effort to provide researchers with useful data for the open-ended questions and all of the prior closed-ended questions. As such, the word count of the open-ended responses are hypothesized to measure the likelihood that a given respondent provides a thorough and thoughtful response. Again, high GPA respondents are able to formulate an open-ended response in less time. If cumulative GPA is a measure of satisficing, then high GPA students will commit more time to survey completion and offer longer open-ended question responses.

The minutes taken to complete the survey measures the response accuracy and the likelihood to misreport in a socially desirable direction. Respondents who complete the survey in 3.5 minutes or less provide less accurate responses and systematically over report their behaviors on 10 of 13 questions. Meanwhile, correctly reporting respondents take significantly longer to complete the survey and, when they do misreport, are more likely to underreport socially desirable behavior. The ambiguity of "Q1A: CRF Use – Prior 7 Days" with the spring university holiday in the reference period is most prone to social desirability bias and misreporting. Respondents who provided accurate reports, on average, took 6.3 minutes to complete the survey, and over reporters averaged only 5.7 minutes. Respondents speeding through the survey are more likely to misreport, and to do so in a socially desirable direction, leading to the inferior accuracy of responses.

The number of words provided in optional, open-ended questions measure the likelihood to satisfice by identifying respondents who volunteer to provide additional information. Question 26 asks respondents to describe, "How do you believe physical activity is related to academic success?" using open-ended text typed into a word box. The content of the text is not being analyzed, and the submissions remain unread. Instead, the text is used to test the hypothesis that longer open-ended question word count measures satisficing; if correct, there will be a positive correlation between word count and accuracy. It follows that there is an inverse correlation between word count and likelihood to satisfice. Satisficing respondents avoid answering open-ended questions just as they avoid the work required to seriously contemplate responses to closed-ended questions, which leads to less accurate responses. For respondents who offered an open-ended response, the relationship between open-ended question word count and the accuracy of self-reported behaviors is not strong with a 0.11 correlation between word count and response accuracy, though the relationship between word count and satisficing is intuitive.

Table 4.6.

Satisficing Metrics and Response Accuracy by Cumulative GPA Quintile.

GPA Quintile	N	Minutes to Complete	Word Count	Pct. Correct Report	Over Report to Underreport
1st Quintile GPA (0.00 to 2.84)	303	5.26	5.89	55.0%	1.62
2nd Quintile GPA (2.85 to 3.21)	320	5.40	7.34	58.1%	1.38
3rd Quintile GPA (3.22 to 3.50)	313	5.73	6.99	63.4%	1.15
4th Quintile GPA (3.51 to 3.82)	278	6.02	8.93	62.5%	1.12
5th Quintile GPA (3.83 to 4.00)	271	5.80	9.08	71.7%	0.99

Note. Higher GPA respondents show lower levels of satisificing.

Responses from high GPA students have lower satisficing measures, are substantively more accurate, and are unbiased estimates of actual behavior. Aggregated by cumulative GPA quintiles for the target population, Table 4.6 shows the satisficing metrics and accuracy in self-reporting across all respondents who completed the survey in 20 minutes or less. A disproportionate number of high GPA students were omitted because they updated their survey responses in subsequent days, thus pushing their completion time above 20 minutes, eliminating their eligibility for Table 4.6. The lower GPA respondents used 10% less time to complete the survey and offered open-ended question responses with 54% fewer words than high GPA respondents. Consistent with my argument, survey responses from the low GPA students are 16.7 points less accurate than high GPA students. The over reporting to underreporting ratio for low GPA students is 0.63 points higher. Inaccurate self-reports for low GPA respondents are likely to be biased in a socially desirable direction; whereas, the 0.99 ratio for high GPA students shows that they provide unbiased behavioral estimates. The high GPA students also spent more time completing the survey and opted to provide longer open-ended question responses. Meanwhile, there is no relationship between SAT score and time to complete the survey, nor open-ended question word count. This section furthers the argument that GPA predicts response accuracy because high GPA respondents are more likely to provide responses that are more accurate and are less likely to satisfice.

Student participation in co-curricular activities. Respondents who do not participate in co-curricular activities – including Greek organizations, LLVs, intramural athletics, formal fitness programs, club sports, and Outdoor Adventures – are the most accurate self-reporters. This finding is consistent with the literature in two ways. First, nonparticipating students have a simpler history to describe because they have fewer relevant memories. Respondents employing the recall and count strategy are likely to respond accurately because they have no memories to find. Satisficing respondents can select the 'did not participate' response option as a straightforward behavioral description that demands minimal effort. Meanwhile, participants who must describe their participation are subjected to a more rigorous response process because they must first recall their degree of participation accurately and then navigate through multiple response options that may not explicitly describe their behavior. Second, for nonparticipants in an activity, their participation is unlikely to be tied to their identity, and the respondent is less motivated to exaggerate actual behavior to conform to self-impressions. Brenner and DeLamater (2014) found that respondents who possess an exercise identity systematically inflated self-reported CRF use, and those without an exercise identity provided an unbiased estimate of CRF use. Students who never participated in an activity are less likely to have incorporated that activity into their identity.

Table 4.7 shows how participation in campus activities is related to response accuracy. Participation in one activity is correlated with participation in other activities. For example, members of Greek organizations are 56% more likely to participate in Intramural Athletics; because increased participation leads to less accurate self-reporting, Greek members are less accurate reporters. LLV participants are also more involved, participating in 56.6% more co-curricular activities than non LLV participants. However, LLV participants are disproportionately high GPA students who are self-reporters and are more accurate. Among LLV participants, the presence of high GPA students interacts with response accuracy such that the aggregate differences in response accuracy are minimal.

For respondents who never participated in CRF-based co-curricular activities including Intramural Athletics, formal fitness programs, Club Sports, or Outdoor Adventures, 71.5% accurately self-reported their lack of participation. Meanwhile, only 60.8% of respondents who participated in one or more of these activities accurately described their behavior. Respondents are more likely to accurately recall co-curricular participation when the program requires engagement over weeks or months. For example, intramural seasons occur over many weeks, and 81.3% or respondents accurately recalled their participation in intramural athletics. Conversely, Outdoor Adventures outings occur over a single day or weekend and are easily forgotten during the survey response process. Therefore, only 73.5% correctly reported their participation with frequent under reporting. Formal fitness programs are a single hour and are easily conflated with informal workouts; though 84.3% respondents accurately describe their participation, the rate of over reporting their participation level is commonplace. The use of unfamiliar terms in the survey questions also contributed to

Table 4.7.

Percentage of Questions Answered Correctly by Student Description.

	N	Accuracy Pct.	Std Dev
All Students			
Total	1114	64.4%	0.1704
2015-16 CRF Use Rate			
Never entered	179	70.2%	0.1742
Less than once per week	301	66.8%	0.1739
Once per week	330	62.2%	0.1693
Three or more times per week	147	62.9%	0.1566
Twice per week	157	59.1%	0.1481
Peak Semester CRF Visits Over Career			
1st Quintile: 0 to 12	221	73.7%	0.1709
2nd Quintile: 13 to 25	227	63.9%	0.1653
3rd Quintile: 26 to 38	223	63.9%	0.1613
4th Quintile: 39 to 59	222	59.6%	0.1592
5th Quintile: 60 to 156	221	60.7%	0.1593
Greek Membership			
Greek Member	199	59.2%	0.1573
Independent	915	65.5%	0.1711
Learning Living Village Membership			
LLV Partic	625	64.0%	0.1629
No Village	489	64.9%	0.1795
CRF-Based Co-Curricular Activity Participation			
0 Activities	369	71.5%	0.1721
1 Activity	345	64.7%	0.1562
2 Activities	241	59.2%	0.1477
3 Activities	103	54.8%	0.1522
4 or More Activities	56	54.6%	0.1603

Note. Respondent accuracy not affected when survey variable unrelated to participation metric.

measurement error. Like Intramural Sports, Club sports occur over many weeks or months, but the term “club sport” is not a widely familiar term across non-participating students, which leads respondents to misinterpret informal recreation and/or intramural sports as club sport participation. Only 67.8% accurately reported their participation in club sports with common over reporting.

Research Question 1B

Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?

Describing behaviors that occur more recently and/or during a shorter college career is a less complex request that yields more accurate self-reported behavior. For example, asking a student to recall their prior 7 days of CRF use is a relatively easy task because of the prior week is more recent and because students rarely visit the CRF more frequently than daily, which limits the complexity of count. Meanwhile, a more complex request is CRF use over weeks or months because respondents must recall instances from a longer reference period, and recalled data requires more effort to communicate. The accuracy of self-reports is also related to the type of behavior requested. Activities that occur infrequently or which are easily conflated with similar behaviors are less accurately reported; whereas students can recall habitual activities with more accuracy. When respondents cannot recall behavior accurately, they systematically over report socially desirable behavior. This section reviews data to suggest that respondents only accurately recall complex behaviors from recent days and defer to generalities as time passes.

Table 4.8 illustrates this principle because respondents who are enrolled in fewer years provide more accurate responses because longer reference periods lead to less accurate self-reports. In cases where the passage of time causes more complexity, accurate self-reporters have a less complex and shorter history to recall; in cases where the passage of time does not cause complexity, respondents' failure to recall is correlated with the time's passage. In the two questions about CRF use, the under reporters had the shortest careers, but there is not a statistically significant relationship. Unlike other questions with significant differences between academic career length and response accuracy, the CRF use reference period and CRF use rates are unaffected by length of the student's career. Time of enrollment does not affect reporting accuracy when time's passage is not germane to the complexity of behavior or the reference period of the activity.

However, for behaviors where the passage of time is correlated with complexity of the behavior and the reference period, respondents with shorter careers provided responses that are more accurate. Among all eight of the questions about academics, respondents who correctly self-reported their behavior were enrolled, on average, fewer years than misreporting respondents. For academic history, the length of the academic career is highly correlated the temporal-related recall complexity. For example, asking a first year student to recall their fall semester GPA is a relatively easy task because they have very few semester GPAs to choose from, and the semester GPA is highly correlated with cumulative GPA. However, the same request is more complex for sixth year students because they must select

Table 4.8.

Years Enrolled by Question and Reporting Accuracy.

	Under Reporters	Correct Reporters	Over Reporters	Other Misreports
Q1A: CRF Use - Prior 7 Days	2.1	2.4	2.2	
Q1C: CRF Use - Weekly Rate	2.3	2.4	2.4	
Q2: Cumulative GPA	2.4	2.3	2.4	
Q3: Semester GPA	**2.4(A)**	**2.1(B)**	**2.5(A)**	
Q4: D/F/U Grades Earned	**3.2(A)**	**2.2(B)**	**2.4(B)**	
Q5: A Grades Earned	**2.7(A)**	**2.1(B)**	**2.0(B)**	
Q6: HESF Course Taken	**2.9(A)**	**2.3(B)**	**2.7(A)**	
Q7: HESF 100 Course Taken		**2.3(B)**		**3.1(A)**
Q8: HESF Grade	3.3	3.1	3.3	2.6
Q9: HESF Class Skipping	**3.0(A)**	**2.9(A)**	**3.1(A)**	**2.2(B)**
Q10A: Intramural Sports Partic.	**3.5(A)**	**2.4(B)**	**2.1(C)**	
Q10B: Fitness Program Partic.	**3.0(A)**	**2.3(B)**	**2.4(B)**	
Q10C: Club Sports Partic.	**2.9(A)**	**2.3(B)**	**2.3(B)**	
Q10D: Outdoor Adventures Partic.	**3.1(A)**	**2.3(B)**	**2.2(B)**	

Note. Bold denotes Duncan Test intra-question statistically significant mean differences ($P<.05$). Means with different letters are significantly different from means in other letter groups. "Other Misreports" describe incorrect self-reports that lack a numeric relationship and cannot be classified as under or over institution reports

from over a dozen semester GPAs, and a single semester's GPA is less correlated with cumulative GPA. As a student progresses into their academic career, the complexity of academic behavior increases because the repetitive nature of a classroom experience yields similar experiences, making it increasingly difficult to differentiate a given class from all the others. Question complexity influences response accuracy more than longer reference periods because cumulative GPA accuracy is independent of career length, but the accurate semester GPA reporters have statistically significantly shorter academic careers. For all students, the cumulative GPA was last calculated on the same date and, like the CRF use behavior, the reference period was unaffected by academic career length.

Finally, respondents systematically underreport behaviors where the passage of time was only correlated with the reference period and not the complexity of behavior. Among the four co-curricular participation questions, the under reporters always had statistically significantly longer academic careers than the correct and over reporters. The passage of time causes students to forget their past participation; it does not, however, increase the complexity of their participation portfolio in the context of these questions about co-curricular participation. A respondent must only recall if they frequently participated in, infrequently participated in, or never participated in a CRF-based co-curricular behavior. First year and fourth year students experience Intramural Athletic participation as an equally unambiguous experience. The passage of time increases the likelihood that a respondent cannot recall and then underreport.

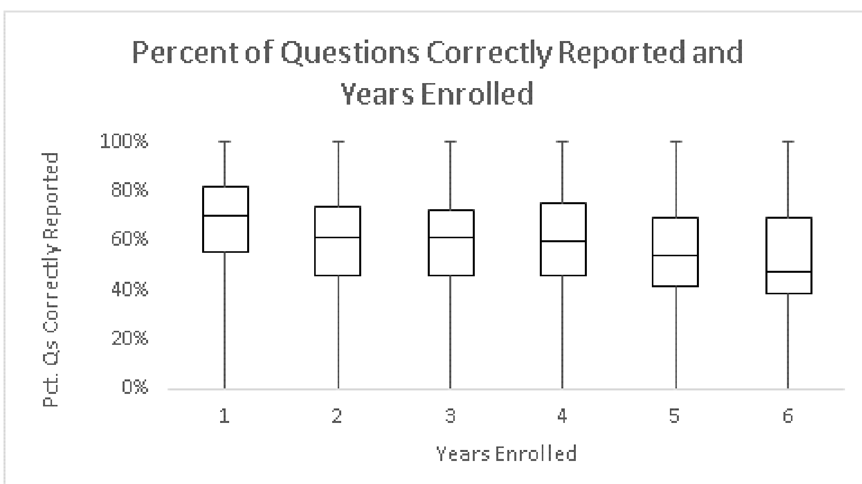

Figure 4.3. Percent of questions correctly reported by years enrolled.

Figure 4.3 tracks an individual's response behavior across all responses to show how the number of years in a respondent's academic career, as a respondent attribute, affects the same person's response accuracy across all questions. Overall, the rate of correct reporting decreases as a respondent progresses through their academic career; first year students report with 70.0% accuracy and sixth year respondents do so with 47.8% accuracy. As students' careers progress, they become more likely to underreport behaviors; first year students underreport 7.7% of their responses and sixth year students do so for 17.5% of their responses. There is no relationship between length of academic career and percent of questions over reported. As students progress through their academic career, they become less likely to report accurately and more likely to underreport their actual behaviors.

Reference period changes and CRF use. Shepherd (2003) suggests that the best method for collecting self-reported physical activity is to request survey takers to describe their prior 7 days of activity when that week is a typical period. This study confirms Shepherd's finding because "Q1A: CRF Use – Prior 7 Days" yielded the most accurate self-reported CRF use when the prior 7-day reference period is behaviorally typical (e.g., excludes the spring university holiday). Response accuracy decreases when the reference period is atypical (e.g., includes the spring university holiday) or requires respondents to describe their CRF use over an entire academic year. When the question is properly presented, there is a 0.71 correlation between the actual and self-reported prior 7 days of CRF use. Meanwhile the correlation is between self-reported and actual CRF use for the atypical 7 days prior is 0.50 and 0.69 for "Q1B: CRF Use - Prior Year (Likert)." When respondents were asked to describe CRF use with an atypical reference period or as a yearlong summary, I hypothesized that respondents would split into two categories of CRF use reporting styles: those providing accurate reports and those describing use at some point in the past. A well-constructed question makes respondents more likely to fall into the former category. When the response falls into the latter, Table 4.9 shows that the added complexity in question design leads to idealized self-reports that more closely correlate with the respondent's peak use over the student's career rather than the use over the intended reference period. The prior 7 days of use is also highly correlated to the responses to "Q1B: CRF Use - Prior Year (Likert)" (0.62) and "Q1C: CRF Use - Prior Year (Rate Based)" (0.69), suggesting that the prior 7 days of use captures the same information and collects more accurate self-reports. The positioning of the reference period relative to the spring university holiday has no significant impact on the accuracy of "Q1B: CRF Use - Prior Year (Likert)" and "Q1C: CRF Use - Prior Year (Rate Based)."

Table 4.9.

Correlation of Self-reported CRF Usage Metrics with Institution Reported

	Prior 7 Days	Prior 7 Days Without Spring Break	Average Weekly Visits (2015-16 year)	Max Semester Visits (Over Career)
Q1A: CRF Use - Prior 7 Days	0.504***	0.714***	0.478***	0.651***
Q1B: CRF Use - Prior Year (Likert)	0.387***	0.624***	0.528***	0.686***
Q1C: CRF Use - Prior Year (Rate Based)	0.439***	0.688***	0.535***	0.670***

*$p<.05$, **$p<.01$, ***$p<.001$.

Respondents offer significantly more accurate self-reports when the question's reference period does not cause confusion, as it does for respondents with spring break in the prior 7-day reference period. Respondents with spring break in the reference period offer systematically inflated self-reports compared to respondents without a spring break reference period. Surveys completed from March 14 to March 20 (e.g., with the March 4, 2016 to March 13 spring break in the 7-day reference period) have a 0.38 correlation with institution reported prior 7 days of use. Respondents who misreport to this question over report 124.0 times more often than they underreport. The corresponding ratio for respondents without the spring holiday in their reference period is 4.8. I hypothesize that some spring break respondents, knowing the reference period is atypical, assign a different meaning to question 1A and assume they should replace the question's literal interpretation with use over a typical 7-day stretch. Unfortunately, respondents substitute the typical period with an idealized period and systematically over report.

GPA self-reports. Memories from more recent events are recalled more accurately from those of periods further in the past. Respondents report grade point averages in two questions: cumulative GPA on the date of the survey request and their most recent semester GPA in fall 2015. Cumulative GPA should be easier to recall because it is a current metric that is used to assess academic accuracy, determine eligibility for university programs, and provides insight into a student's employability. A semester GPA, meanwhile, is a meaningful milestone towards computing a cumulative GPA, but the metrics are rarely revisited after the report card is published. As such, the accuracy of self-reported cumulative GPA measures shorter-term memories (i.e., that day) than the fall 2015 semester GPA (i.e., 3 months). The correlation between self-reported and institution reported cumulative GPAs is 0.95 with a self-reported mean of 3.33 that nearly matches the institution reported mean of 3.30. To the tenth of a grade point, 70.1% of responses matched institution reported data, and 46.7% providde the correct GPA to the hundredth of a grade point. Meanwhile, the correlation between self-reported and institution reported semester GPAs is 0.72 with a self-reported mean of 3.34, which is larger than the institution reported mean of 3.23. To the tenth of a grade point, 50.1% of responses matched institution reported data, and 33.4% provided the correct GPA to the hundredth of a GPA point. Table 4.10 shows cross tabulation of student career length, and self-reporting accuracy attributes most of the differences in cumulative GPA versus semester GPA to the time passed rather than to the complexity of request. The correlation between career length and reporting accuracy is .09, though respondents who first enrolled in fall 2015 offer substantially more accurate semester GPA reports, but their semester GPA usually equals their cumulative GPA. Proximate behaviors are more accurately recalled than those from further into the past, and when accurate values are not stated, the misreporting is likely to be biased in a socially desirable direction.

Recall of facts. When students are asked to recall and count behavior for socially neutral activities, the passage of time decreases the likelihood that a respondent accurately reports their behavior and frequently causes underreporting of past behaviors. For example, when asked to recall and count number of high/low grades received over their academic

Table 4.10.

Career Length and Self-Reported GPA Accuracy

	Cumulative GPA		Semester GPA	
Years Enrolled	Within .01	Within .1	Within .01	Within .1
1	55.6%	81.4%	57.4%	84.8%
2	49.7%	86.8%	28.1%	53.1%
3	48.2%	89.5%	30.6%	55.9%
4	53.4%	88.1%	34.0%	55.8%
5	48.9%	87.2%	30.2%	46.5%
6 or More	37.5%	62.5%	25.0%	37.5%

years, respondents underreported both the number of A grades earned and the number of D/F/Unsatisfactory grades. Students in their first two years accurately reported 89.7% of D/F/Unsatisfactory grades earned and 40.9% for A grades; these accuracy rates fall compared to 72.2% of D/F/Unsatisfactory grades and 34.2% of A grades for later career students. Misreports are over 3 times more likely to be underreported than over reported. Though grades earned may be subject to social desirability bias, a similar result is found for socially neutral topics. For example, among students enrolled 3 years or more and who self-reported taking 1 PE/HESF course, 75.9% accurately reported which course they enrolled in compared to 92.4% for students enrolled 2 years or less. Recall questions get responses that are more accurate when less time passes between the event and survey completion.

Even with a long delay between the referenced event and survey completion, habitual behaviors are more likely to be recalled than participation in standalone events. For example, 92.1% of all respondents correctly identified their PE/HESF 100 level course content from a drop down list. These courses were attended biweekly for nearly four months (e.g., an academic semester) and left an impression on participants. Similar trends are seen in the co-curricular participation questions. For most sports, Intramural Athletic participation should be easily recalled because athletic seasons require habitual behavior including competitions and informal gatherings like practices that span weeks or months. Meanwhile, formal fitness programs may be under counted because they occur over a single hour and are easily conflated with informal workouts. Some participants do attend formal fitness programs habitually, however. Finally, Outdoor Adventures outings should be underreported because they occur over a single day or weekend, and most participants do so only once. The data follows the hypothesized pattern: 81.3% of respondents correctly reported intramural participation, but underreporting participation rate is less common; 83.4% of respondents correctly recalled participating, but nearly all underreported their participation rate; finally, only 17.6% of participants accurately recalled their involvement in an Outdoor Adventures program. Recalling past events of socially neutral topics leads to systematic underreporting, though habitual events are more accurately recalled than behaviors of limited occurrence.

Social desirability bias. For behaviors subjected to social desirability bias, the proximity of the behavior does not guarantee a more accurate self-report because clarity motivates a respondent to intentionally misreport in cases where the respondent judges a past behavior as inapplicable to the question. For example, in question 9 the respondents who were recently enrolled in a PE/HESF course and had high GPAs are expected to provide accurate reports of their class skipping rate and, when they misreport, provide unbiased estimates. Instead, this group systematically underreported their class skipping. I attribute this to clear memories of why they skipped a class meeting and the group systematically rationalizing their behavior to omit it from their self-report. Across all respondents, those who took their PE/HESF 100 level course in the 2015-16 academic year were substantially more likely to underreport class skipping; underreporting occurs 4.7 times more often than over reporting. For students taking the course further into the past, the ratio is only 2.4. I theorize that respondents who recently enrolled are more likely to recall why they missed a class meeting and, if the cause is not recreational, they chose to omit it because it was for an honorable reason. For example, a respondent may remember that they missed one class to attend a funeral and may then decide to downward edit their class skipping count because skipping to attend a funeral is not a frivolous class skip. Respondents taking a class further into the past cannot recall why they skipped class, so they answer the questions as intended. An alternative explanation is that students with high GPAs are less likely to skip classes in general, and because they cannot recall details for class skipping, their estimation strategy is to describe their typical number skipped class meetings for any given course at NC State. Because higher GPA students skip fewer classes, they offer lower systematic estimates of class skipping. The actual number of classes skipped for high GPA students in PE/HESF 100 level courses is 2.3 compared to 3.1 for low GPA students.

Research Question 1C

Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?

To answer the survey's first question, each respondent was randomly selected to describe their CRF use with one of three response options. A comparison of accuracies across the three response option groupings helps assess how survey question design can determine the validity of a self-reported behavior susceptible to social desirability bias. The first response option was to recall and count the prior 7 days of CRF use; the second was to self-report academic year CRF use with vaguely worded, ordinal descriptors; and the third response option was a weekly, rate-based description of use over the academic year. Prior research finds that the first response option provides the most accurate responses because the request is the most precise and least mentally taxing. Response option 2 uses vague qualifiers like “Very Often” and “Somewhat Often” to describe CRF use with vagueness causing confusion and lessened validity of self-reports. The third response option is taxing because respondents must recall CRF behavior over 8 months and then transform use into a rate-based description. When the prior 7 day reference period is a typical week, self-reports are moderately correlated with both actual use and longer spans of CRF use. However, for all response options, respondents are likely to over report their CRF use in a socially desirable direction. This section is a comparison of response accuracy to questions 1 about CRF use that employs 3 sets of response options to describe behaviors.

CRF use by prior 7 day count response options. Question 1A asks respondents to describe their CRF use as the number of entries over the prior 7 days by using a continuous list of response options ranging from “0” to “20 or more”. For physical activity, Shepherd (2003) finds that the 7-day reference period yields the most accurate self-reports, though the correlation between self-reported and validation data sets rarely exceeds 0.5. This study’s results are similar with the caveat that the prior 7 days must be a period of typical behavior which was only met for 44.9% of question 1A’s respondents. For the remaining 55.1%, the reference period includes the spring university holiday, which leads to systematic over reporting as respondents struggle to consistently respond to a complexly implemented question. Respondents with spring break in their reference period are inaccurate self-reporters with a 0.38 correlation between self-reported and institution reported values and a 54.0% accuracy rate. Meanwhile, respondents without spring break in their reference period report with a 0.82 correlation and 68.1% accuracy rate. I hypothesize that the spring break respondents disregarded the stated instructions and described CRF use over a self-defined reference period and offer use rates that are often idealized rather than typical. It is also possible they reassigned the reference period to describe the most recent 7 days that classes met by appending a portion of days occurring before the spring university holiday to days occurring after.

Overall, 55.3% of respondents answered question 1A correctly with a 0.50 correlation between self-reported and institution reported values. Nearly half of respondents reported not visiting the CRF over the reference period and, of them, 99.6% actually did not. Of those reporting at least one visit, only 13.3% of self-reported visitors accurately reported the number of entries, and only 51.2% actually visited the CRF. Respondents who reported a visit without spring break in their reference period were more accurate with 56.3% correctly reporting their number of visits, and 88.2% accurately describing visiting, at all. The self-reported number of CRF entries for respondents with a spring break reference period is 1.77 with an institution reported value of 0.25; meanwhile, respondents without a spring break reference period self-reported an average of 1.85 visits with an institution reported value of 1.19. Misreporting respondents almost always over report their CRF use with a self-reported average of 1.80 compared to an institution reported average of 0.67 and 11.9 over reporters for every 1 under reporter. Measurement error threatens the validity of a seemingly easy question, and though respondents are substantively more accurate self-reporters, when the reference period is typical, the response mean is an inflated description of CRF use, as shown in Figures 4.4. and 4.5.

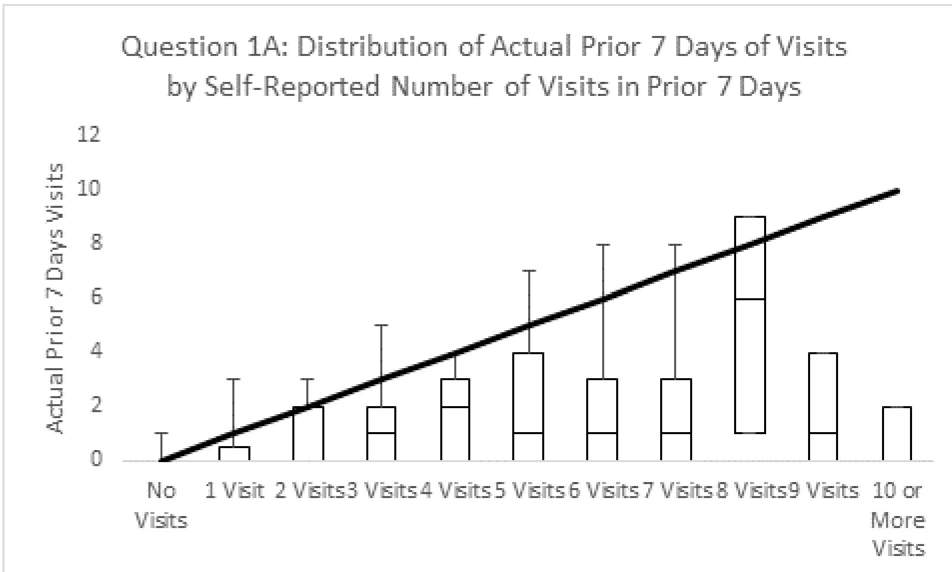

Figure 4.4. Question 1A: self-reported and institution-reported CRF use.

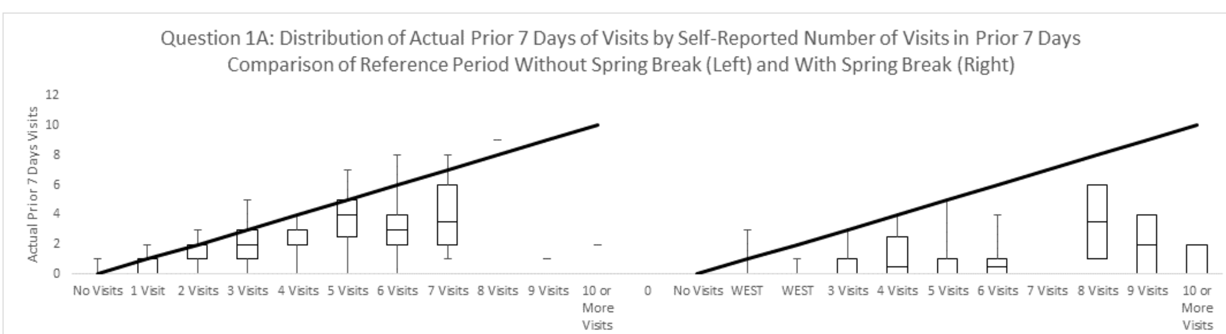

Figure 4.5. Question 1A by reference period.

Question 1A responses take a continuous scale, implying that researchers may use the results linearly to measure the intensity of physical activity, but validity issues make assumptions about data accuracy problematic. A Duncan test (see Table A.2) reveals that the respondents to question 1A are not sorting themselves into groups with distinct means or distributions. Respondents reporting above/below average levels of CRF use are accurately sorting themselves into comparatively accurate ordinal categories, but the categories do not accurately describe their actual use. Though infrequent users accurately describe their CRF use, there is no distinguishable relationship between the self-reported and institution reported values for high volume self-reporters. I hypothesize that a continuous scale provides too many response options, and reporting a smaller, ordinal scale is more appropriate. Nonresponse error inflates the prior 7 days of use mean from 0.52 to 0.67, and measurement error further inflates the survey mean from 0.67 to 1.81 with the survey means inflating the prior 7 days of CRF use by 270.2%.

CRF use by Likert scale. Question 1B asks respondents to describe their academic year (to survey date) CRF use on a vague Likert scale with response options of "Never," "Sometimes," "Often," and "Very Often." These response options are commonly used on national surveys, including NSSE, whose technical notes find respondents sort themselves into distinct and valid groupings (Kuh, 2004). Results from question 1B confirm this finding because the mean and variance of each response option group are distant and different from the others (see Table A.3). For the middle 50% of respondents, there is limited overlap between neighboring groups. Like questions 1A and 1C, respondents do divide themselves into ordinal ranked categories, but unlike questions 1A and 1C, these response options do not imply that the use level is proven inaccurate. The progression through response options approaches an exponential function that cannot be used linearly in quantitative analysis.

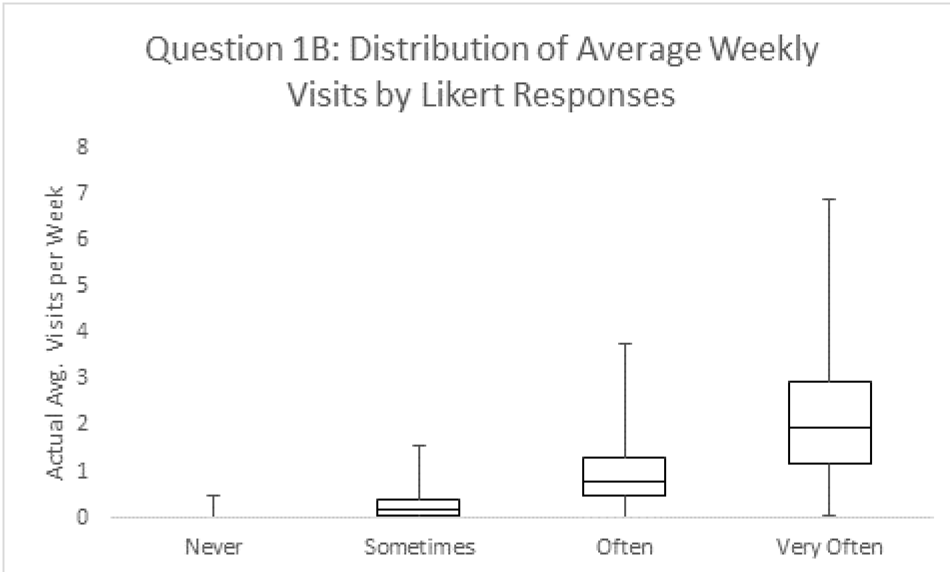

Figure 4.6. Distribution of question 1B responses.

Though the groupings are valid, unit-level responses to question 1B are not reliable descriptors of behavior. Figure 4.6 and Table A.3 show that there is no overlap between the groups' quarter 1 to quarter 3 ranges, but the range of each response option is fully a subset of the next highest frequency group. The middle 50% ranges do not overlap except for 1.18 visits per week to 1.32 visits per week among the "Often" and "Very Often" respondents. This area of overlap includes 10 "Often" and 8 "Very Often" respondents. "Never" respondents rarely use the CRF, but there is much intra-group variance for the top quarter of users in the "Sometimes" and "Often" response groups. The top quartile of respondents in these groups are not significantly different from the "Very Often" respondents. Of the 86.0% of all question 1B respondents who reported some level of CRF use, 94.4% of them did enter the CRF at least once. Conversely, of the 14.0% of respondents who reported never entering the CRF, only 80.6% never entered in the 2015-16 academic year (to survey date); the 19.4% who misreported entered an average of 0.09 times per week. Measurement error is difficult to evaluate, but evidence suggests that the respondents properly respond to ambiguous Likert scale response options. They may neglect to recall perfectly at the unit level but do align themselves accurately into groups with distinctive means and variance. The vague quantifier causes high frequency users to become marginally more likely to self-report themselves into lower frequency use groups.

CRF use by rate based estimation. Question 1C asks that respondents describe their CRF use in the 2015-16 academic year (to date) on a 5 point, rate based scale including response options “I have never entered,” “Less than once per week,” “Once per week,” Twice per week,” and “Three or more times per week.” The literature suggests that rate based estimation is an effective strategy for self-reporting for most behaviors but leads to over reporting of physical activity (Shepherd, 2003; Tourangeau et al., 2000). Results suggest that respondents do over report CRF use but accurately order themselves into ordinal ranked response groups with distinct means, as shown in Figure 4.7. Each response grouping has distinct mean, and the groupings possess different behaviors (see Table A.5) with less overlap in the extremes when compared to the question 1B response groupings.

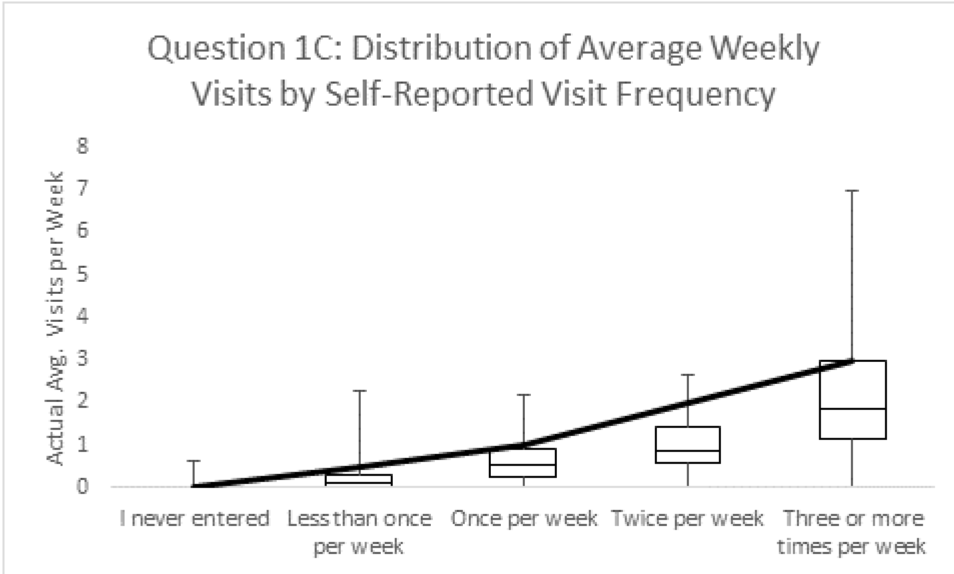

Figure 4.7. Distribution of question 1C results.

Across the groups, differences in the summary statistics for each rate-based response option are not dissimilar from the group differences for the Likert Scale groupings. Though not a substantial issue for any response option, high frequency CRF users responding to 1c more reliably place themselves into the highest user group because they are provided a reference point. For example, the top 10% of CRF users visit 2.2 times per week. In question 1C, only 2 of these high frequency users underreported their use; meanwhile, 9 of the high frequency users placed themselves into lower use response options. Systematic underreporting is not an issue for self-reported CRF use, but a lack of a benchmark rate adds ambiguity and increases the likelihood of underreporting. This phenomenon may be more observable in topics that are less prone to social desirability bias. Eliminating vagueness helps respondents orient themselves and marginally improves the validity of response option groupings.

Response option comparison. Only 47.9% of all respondents accurately self-report their use, and over reporting is frequent. Among self-reported non-users, 89.6% accurately reported their lack of CRF use. Among the self-reported CRF users, 92.4% did enter the CRF, but only 41.5% accurately reported their CRF use rate, and respondents over reported their use rate 8.8 times more often than they underreported. Overall, the question 1B Likert scale and question 1C rate based use scale provide comparable results that researchers could use identically. Among question 1B respondents, 13.3% of respondents reported having never entered the CRF over the course of the academic year, which is similar to the 14.0% rate reported by question 1C respondents. Overall, I conclude that the rate based estimation method communicates identical descriptive information as the Likert scale in Question 1B but that the specification of a numeric use rate is problematic because respondents do not accurately describe the rate of use, though researchers may believe that the values can be used in numeric comparisons. Question 1C's nonresponse error would inflate the average weekly visits from 0.78 to 1.00, and the measurement error would further inflate the surveyed mean from 1.00 to 1.65 with self-reported average weekly visits inflated 215.4% above the population mean

Across questions 1A, 1B, and 1C, I collected self-reported CRF use with varying response options and reference periods to determine which yields the most accurate self-reports. Question 1A describes the prior 7 days of CRF use on a continuous scale, question 1B describes CRF use over the academic year (to survey date) on a Likert Scale, and question 1C describes CRF use over the academic year (to survey date) on a rate-based scale. Shepherd (2003) suggests that question 1A is the best method because it requests the self-reported number of visits over the prior 7 days when the prior 7-day reference period is behaviorally typical, but ideal conditions yield only a moderate correlation between self-reported and actual physical activity. Results from questions 1A, 1B, and 1C confirm Shepherd's (2003) findings because question 1A yielded the highest correlation with physical activity when the reference period was affected by spring break. Respondents provided substantively inflated reports when the reference period included the atypical spring holiday (at NC State, atypical reference periods include final exams, university holidays, Valentine's Day, Halloween, adverse weather days, and high profile intercollegiate athletic competitions). For all three question formats, respondents over reported CRF use, and only 31.1% of those entering the CRF in 2015-16 provided accurate self-report of CRF use. Respondents do accurately sort themselves into ordinal categories that describe use relative to each other, but assigning a numeric value to use just leads to inaccurate reports. For all measures, the self-reported CRF use in March 2016 was more correlated with the respondent's peak use over their student career than the reference period. As such, a more likely interpretation of self-reported CRF use is to use it as a description of the highest volume period over their student career. I theorize that respondents are split into two categories of CRF use reporting styles: those providing accurate reports and those describing use at some point in the past. Though non-users usually fall into the former group, there are no clues to determine which one a CRF user falls into.

Table 4.11.

Correlation of Self-reported CRF Usage Metrics with Institution Reported.

	Prior 7 Days	Prior 7 Days Without Spring Break	Average Weekly Visits (2015-16 year)	Max Semester Visits (Over Career)
Q1A: CRF Use - Prior 7 Days	0.504***	0.714***	0.478***	0.651***
Q1B: CRF Use - Prior Year (Likert)	0.387***	0.624***	0.528***	0.686***
Q1C: CRF Use - Prior Year (Rate Based)	0.439***	0.688***	0.535***	0.670***

*$p<.05$, **$p<.01$, ***$p<.001$.

Table 4.11 shows the correlations between the measures of CRF use and self-reported responses to questions 1A, 1B, and 1C. For question 1A, there is a 0.50 correlation between self-reported and institution reported visits in the prior 7 days, though the correlation climbs to 0.71 when spring break is removed from the reference period. The prior 7 days of use is also highly correlated to the responses to question 1B (0.62) and 1C (0.69). The correlation between actual CRF use and question 1B responses are comparable to the correlations for other use metrics that possess articulated use rates; responses to question 1B are collecting similar information to those in questions 1A and 1C. The self-reported values for questions 1B and 1C have nearly identical correlations for average weekly visits over the academic year in question 1C (0.53 and 0.54, respectively).

Response time comparison. Respondents offer significantly more accurate and less biased self-reports when question design does not force respondents to react to interpret the question's actual intent. In the response option format for describing prior 7 days of use, half of the respondents had the spring university holiday included in their reference period; these respondents reacted to the question by wondering if they should describe their actual use over an atypical period or describe behavior during a typical 7 day stretch. Many respondents chose to describe a typical 7 day stretch resulting in systematically inflated self-reports. Self-reports are inflated because (1) many students did not enter the CRF of the spring university holiday, which lowered the actual behavior and (2) many students described a higher-than-average use. Easily understood questions are a requirement for valid self-reported behavior. For other response option formats, the question's ambiguity and/or complexity also forced interpretation.

Respondents offer the most accurate responses to easily understood questions about recent behavior; requesting behaviors with confusing questions encourages social desirability bias. In Table 5.1, wave 1 responses include surveys that were completed from March 14 to March 20 with the March 4, 2016 to March 13 spring university holiday in the 7-day reference period; whereas, wave 2 and wave 3 reference periods include a typical 7-day reference period where the entirety of the period takes place during the normal class schedule. For question 1A, which describes the prior 7 days of CRF use, wave 1 responses have a low 0.38 correlation between self-reported and institution reported values; whereas, waves 2 and 3 have a combined correlation of 0.73. Respondents who misreport their question 1A response over report 124.0 times more often than they underreport. Meanwhile, the correlation for wave 2 and 3 respondents is 0.73, and the ratio is only 4.8. Question 1C, which is less sensitive to the spring university holiday, shows little variation in correlation, accuracy, or over reporting and underreporting ratios between waves. Spring break respondents, knowing the reference period is atypical, assign a different meaning for question 1A and assume that they should replace the question's literal interpretation with use over a typical 7-day stretch. This confusion does not exist for questions 1B or 1C. Unfortunately, respondents react to confusion by substituting the typical period with an idealized period that systematically over reports CRF use.

As shown in Table 4.12, validation data shows that third wave responses to questions 1A and 1C have a higher correlation with institution reported data than those provided in earlier waves. This finding raises concerns for a commonly used validation method discussed by Porter and Whitcomb (2005). The time of response analysis assumes that response accuracy from later respondents is not substantively dissimilar from nonresponses and cannot be used to evaluate nonresponse bias. Data collected here suggests that the late respondents are more accurate self-reporters but do not act significantly different from early wave respondents, leading to the consideration of two arguments. First, nonrespondents might provide more accurate responses than respondents would if they did submit a survey. Prior research finds that interest in the survey topic is correlated with both higher response propensities and motivation to over report behaviors in a socially desirable direction (Groves & Peytcheva, 2008). This argument gives preference to responses from individuals with higher nonresponse propensities. Second, survey means from the first wave of responders are not significantly different from those responding towards the end of the survey window and confidence in the time of response analysis is diminished. The second argument is consistent with findings from a similar validation study by Olson and Kennedy (2006) using recent college graduates that determined that late respondents accurately responding is not dissimilar from early respondents. Neither points are conclusive nor are the findings generalizable to all settings or research objectives.

Table 4.12.

Comparison of Response Options to Survey Responses Period.

		Correlation To				Over to Under
	N	Prior 7 Days	Avg. Weekly Visits	Max Semester†	Answering Correctly	Over to Under Reporting Ratio
Question 1A: Prior 7 Days Response Options						
Total (3/14 to 3/24)	487	0.504***	0.651***	0.478***	55.0%	11.9
Wave 1 (3/14 to 3/20 - Spring Break)	272	0.378***	0.664***	0.441***	54.0%	124.0
Wave 2 (3/21 to 3/22 - No Break)	146	0.678***	0.620***	0.545***	50.7%	5.0
Wave 3 (3/23 to 3/24 - No Break)	69	0.820***	0.734***	0.576***	68.1%	4.5
Question 1B: Likert Response Options						
Total (3/14 to 3/24)	515	0.389***	0.686***	0.528***	N/A	N/A
Wave 1 (3/14 to 3/20 - Spring Break)	290	0.273***	0.668***	0.486***	N/A	N/A
Wave 2 (3/21 to 3/22 - No Break)	140	0.645***	0.692***	0.595***	N/A	N/A
Wave 3 (3/23 to 3/24 - No Break)	85	0.588***	0.712***	0.494***	N/A	N/A
Question 1C: Avg. Weekly Visits Response Options						
Total (3/14 to 3/24)	503	0.439***	0.700***	0.535***	47.7%	6.9
Wave 1 (3/14 to 3/20 - Spring Break)	246	0.197**	0.698***	0.540***	46.7%	7.1
Wave 2 (3/21 to 3/22 - No Break)	174	0.672***	0.675***	0.534***	48.3%	6.5
Wave 3 (3/23 to 3/24 - No Break)	83	0.729***	0.774***	0.438***	49.4%	7.4

† Max semester is the highest average weekly CRF visits over respondent's student career.
Note. Response wave 1 includes spring university holiday in the 7 day reference window; response waves 2 and 3 include no university holidays.

*$p<.05$, **$p<.01$, ***$p<.001$.

The actual CRF use rates for first wave respondents is significantly higher than those of later waves and nonrespondents leading to nonresponse bias. For example, for the actual weekly average of CRF use over the academic year, wave 1 respondents average 1.19 weekly visits; wave 2 respondents average 0.85; wave 3 respondents average 0.81; and nonrespondents average 0.78 visits. Lower frequency users are more likely to self-report accurately because their behaviors are easier to describe, and improved accuracy may be unrelated to response wave and instead a result of actual use rates. While the accuracy of self-reported behaviors across response waves may be similar, the actual behavior of respondents across waves is dissimilar, and the survey statistic may not generalize to the sample or population.

Research Question 1D

Are sample means of self-reported behaviors valid estimates of target population means?

This section estimates the impact of nonresponse error and then integrates the measurement error analysis from RQ 1A to evaluate the validity of the survey mean as an estimate of student behaviors across the sample and target population. Within the sample, a comparison of institution-reported means for respondents and nonrespondents detects nonresponse error that is statistically significant and substantive. Next, I estimate the overall impact of combined measurement and nonresponse errors by comparing means across three groups: (1) self-reported data for respondents, only; (2) institution-reported data on respondents, only; and (3) institution-reported data on the entire sample. Differences in sample means between group 1 and 2 estimate the effect of response bias; differences between group 2 and 3 estimate the effect of nonresponse bias; and differences between group 1 and 3 estimate the overall biases. Nonresponse error and measurement error combine to make the self-reported mean a poor estimator of behavior across the sample and target population.

Nonresponse error. Respondents are more likely to be academically successful, active CRF users, and involved in co-curricular activities, making nonresponse error a threat to the validity of self-reported behavioral data. Nonresponse error describes systematic differences between members of the sample who choose to respond to the survey request and those declining the invitation. There are systematic differences between the groups, and even if respondents accurately reported their behaviors, the survey mean would not represent the sample mean because the respondent behavior does not generalize to the sample nor to the target population. At the .01 level, respondents have significantly different behaviors than nonrespondents for 11 of 13 questions with 6 questions having mean differences exceeding 20%. Table 4.12 shows that only “Q2: Cumulative GPA” and “Q9: HESF Class Skipping” are free from nonresponse error, and nonresponse error is a threat for questions about CRF use, GPAs, and co-curricular participation.

As shown in Table 4.13 and Figure 4.8, the cumulative and semester GPAs of respondents are higher than nonrespondents and introduce the threat of nonresponse error. The institution reported that cumulative GPA for respondents is 3.33; though, the nonrespondents’ mean is only 3.13 (the target population mean is 3.17). Across the sample, those with cumulative GPAs below 2.00 had an 18.4% response rate. Among students with a GPA of 2.00 to 2.99, 16.5% responded. Among students with a GPA of 3.00 to 3.49, 22.9% responded, and 27.0% of students with a GPA above 3.50 responded. The average institution reported fall 2015 semester GPA for respondents is 3.34, a substantially higher mean than the 2.83 average for non-respondents and 3.05 overall population mean. Because high GPA students are over represented among survey respondents, results from this survey's data may only be generalizable to high GPA students.

Table 4.13.

T Test on Respondent Mean Minus Nonrespondent Mean is Not Equal to Zero†

	Respondents		Nonrespondents				
Question	N	Mean	N	Mean	Diff Mean	Pct. Bias	F Value
CRF Use - Prior 7 Days	1505	0.67	4495	0.47	0.20	29.6%	**1.52*** **
CRF Use - Weekly Rate	1505	1.02	4495	0.70	0.32	31.3%	**1.46*** **
Cumulative GPA	1317	3.31	4482	3.14	0.17	5.1%	1.08
Semester GPA	1292	3.22	4447	2.99	0.24	7.4%	**1.28*** **
D/F/U Grades Earned	1340	1.02	4660	1.56	-0.54	-53.3%	**1.54*** **
A Grades Earned	1307	10.05	4693	8.89	1.16	11.5%	**1.15** **
HESF Course Taken	1316	1.57	4684	1.42	0.15	9.5%	**1.28*** **
HESF Grade	119	3.43	426	3.24	0.19	5.5%	**1.73*** **
HESF Class Skipping	526	2.60	1644	2.77	-0.16	-6.2%	1.04
Intramural Sports Partic.	1253	0.68	4747	0.61	0.07	10.9%	**1.16*** **
Fitness Program Partic.	1256	0.34	4744	0.24	0.09	27.2%	**1.40*** **
Club Sports Partic.	1262	0.65	4738	0.52	0.13	19.5%	**1.17*** **
Outdoor Adventures Partic.	1241	0.30	4759	0.21	0.09	29.1%	**1.34*** **

†limited to instances where numeric data is available

Note. Significant differences in means indicates respondents do not generalize to nonrespondents.

*$p<.05$, **$p<.01$, ***$p<.001$.

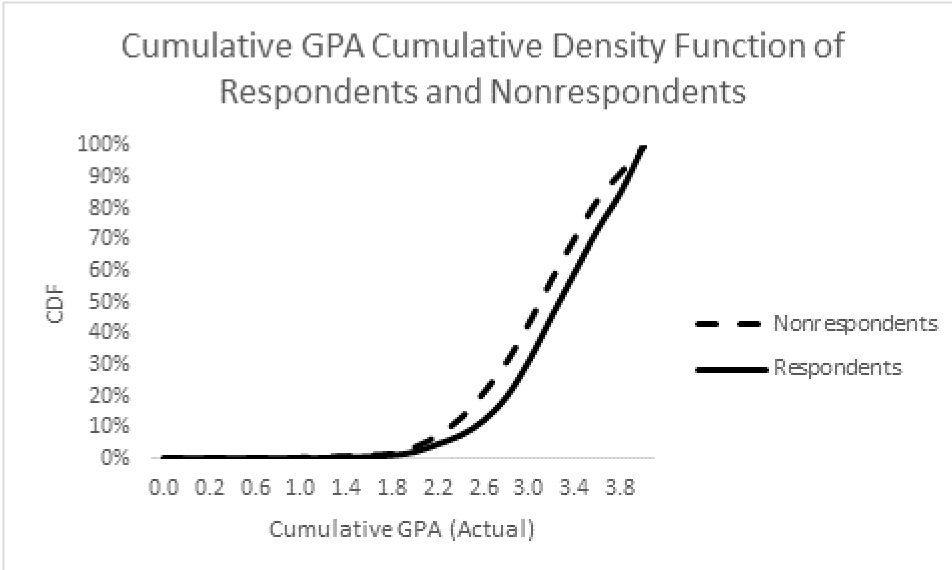

Figure 4.8. CDF by GPA and response status.

Nonresponse error is also a threat to questions about CRF use because respondents are more likely to use the CRF and visit 30% more often than nonrespondents. In the 7 days prior to survey completion, 26.9% of respondents visited the CRF compared to 21.5% of nonrespondents and 23.7% of the target population. Through the academic year, 82.5% of respondents visited the CRF at least once compared to 75.4% of non-respondents and 78.1% of the target population. When compared to nonrespondents, respondents are also higher frequency users. In the 7 days prior to survey completion, respondents visited 0.67 times versus 0.50 for nonrespondents and 0.56 for the target population. Across the academic year, respondents averaged 1.00 weekly visits, and nonrespondents averaged 0.78. Relying on self-reported data to estimate CRF use across the target population leads to inflated estimates.

Nonresponse error is not a substantial threat to the survey's validity for questions about PE/HESF class taking, class skipping, and grades earned. The respondents' means are slightly higher, but the impact of the difference is minimal. For example, 77.1% of respondents have taken at least one PE/HESF course over their career, and 74.3% of non-respondents and 75.3% of the population did so. PE/HESF 100 level grades of respondents have a nearly identical distribution as those of nonrespondents, as does class skipping distributions. For questions about PE/HESF 100 level course taking and academic engagement, nonresponse error is a minimal threat.

Respondents are also more active participants in co-curricular activities, including LLVs, intramural sports, formal fitness programs, club sports, and Outdoor Adventures. For students involved in LLVs for at least one semester, 29.9% responded to the survey request compared to only 23.5% of nonrespondents. For the CRF-based activities of Intramural Athletics, formal fitness programs, club sports, and Outdoor Adventures, participants are significantly more likely to respond than nonparticipants. In all cases, the more intense the student's engagement with the co-curricular activity, the more likely the student is to respond to the survey request. For example, Outdoor Adventures nonparticipants' rate is 19.7%; among those participating in only one semester, the response rate is 26.7%. Among those participating multiple semesters, the response rate is 27.0%, and among those participating in all of their semesters, the response rate is 29.2%. The lone exception to the trend is Greek Life where only 24.6% of Greeks responded versus 27.1% of non-Greek students in the sample. Relying on self-reported participation data will cause inflated rates of student participation.

Groves, et al. (2009) reviewed the causes for nonresponse bias and determined that its presence is concerning when the source of response propensity is correlated with the survey variables; nonresponse is detrimental when the survey variable drives response propensity. My study of response propensity and survey variables that determines the nonresponse bias is, at a minimum, concerning for CRF use, academic success, and co-curricular participation. The more physically active, academically successful, and/or participatory the student, the more likely they are to respond to my survey request. For example, high GPA students exert more effort in their academic pursuits, and this effort is often directed toward participation in formal and informal co-curricular programs. This participatory spirit also applies to my survey request. High GPA students are also more likely to respond because motivation is correlated with response propensity. A detrimental causal relationship between survey variables and response propensity is not immediately obvious because having a high GPA does not directly determine response propensity. Given these concerns, the magnitude of differences between respondent and nonrespondent means drives the assessment, and the observed differences justify concerns about nonresponse error in data.

The Leverage Salience Theory (Groves, et al., 2000) suggests that survey process choices may directly influence response propensity. The initial survey request described the survey as a study about academics and physical activity, but subsequent reminder emails described it as a survey about co-curricular involvement and class skipping. This strategy boosted the response rate but may have simultaneously increased nonresponse bias. It is impossible to determine how communication choices increased nonresponse error, but the survey topics mentioned in the invitations possess disproportionately high levels of nonresponse error when compared to topics not mentioned. The survey was promoted first as a gym users feedback survey that links CRF use to academic success; a subsequent survey prompt described the survey as collecting information on class skipping and academic success. A final survey prompt described it as collecting data about extracurricular participation. Accordingly, nonresponse error is present for CRF use, academic success, and co-curricular participation. PE/HESF course taking behavior was never explicitly stated in email invitations, yet this is the only topic area free from nonresponse error. Placing the survey topic in the email's personal plea for participation may have pushed more students with an interest in the topic to reply, thus unintentionally inviting nonresponse error. While my communications are consistent with similar studies, I am concerned that these choices boosted the response rate and simultaneously added to the magnitude of nonresponse error.

Measurement error and nonresponse error interaction. Though the presence of measurement error and/or nonresponse error are concerns for researchers, the presence of bias is not a prima facie invalidation of the self-reported behavioral data. Random differences between the survey statistic and actual population means are inevitable in the sampling process and can be corrected with statistical techniques. In some cases, the systematic bias may improve the survey statistic because respondents make adjustments that more accurately measure the construct as intended by the researcher (Kuncel, et al., 2005). In other cases, the magnitude and direction of misreporting is consistent across all respondents, resulting in a level shift of the survey statistic that does not affect the correlative relationship (Kuncel et al., 2005). In most cases, however, systematic bias is a threat to the validity of a study's findings, and the systematic biases in self-reporting must be identified and the degree of threat posed by the bias to the validity of the research conclusions determined. This section analyzes self-reported and institution-reported data to ascertain the effect of bias on the analysis of self-reported data.

For each question, Table 4.14 shows the self-reported means collected by the survey along with the institution-reported value for the entire sample. The respondent means are the self-reported values collected by the survey, and the sample means are the actual values if the respondents had offered accurate and unbiased responses. In 12 of 13 questions, the means are significantly different; the only question for which the self-reported mean does not significantly differ from that of the actual mean is "Q10D: Outdoor Adventures Participation." For 7 questions, the magnitude of the difference between means exceeds 20%. Questions about CRF possesses the most bias because the self-reported CRF use for the prior 7 days is 3.5 times the actual value and 2.1 times larger for the question about academic year use. Less affected, but also substantively different, is the cumulative GPA, which is inflated by 0.16 grade points, and the semester GPA, which is inflated by 0.32 grade points. In both cases, the respondents reported a B+ average though the actual value was a B average. Among other academically based success measures, the sample mean is systematically lower than the actual mean. Respondents are more likely to participate in CRF-based, co-curricular activities but underreport this use. Overall, measurement error and nonresponse error causes self-reported data to be a poor estimator of actual behavior across behaviors included in this study.

Table 4.14.

T Test for Differences Between Self-Reported Mean and Actual Sample Mean†

	Respondents		Sample				
Question	N	Mean	N	Mean	Diff Mean	Pct. Bias	F Value
Q1A: CRF Use - Prior 7 Days	486	1.81	6000	0.52	1.29	248.0%	**3.77*** **
Q1C: CRF Use - Weekly Rate	502	1.65	6000	0.78	0.87	111.5%	**1.80*** **
Q2: Cumulative GPA	1334	3.34	5799	3.18	0.16	5.1%	1.09
Q3: Semester GPA	1319	3.36	5739	3.04	0.32	10.5%	**1.87*** **
Q4: D/F/U Grades Earned	1339	0.60	6000	1.44	-0.84	-58.3%	**6.10*** **
Q5: A Grades Earned	1306	7.42	6000	9.14	-1.72	-18.8%	**1.79*** **
Q6: HESF Course Taken	1315	1.35	6000	1.45	-0.10	-7.1%	**1.76*** **
Q8: HESF Grade	185	3.55	545	3.28	0.27	8.2%	**1.34* **
Q9: HESF Class Skipping	680	1.87	2170	2.73	-0.86	-31.5%	**1.73*** **
Q10A: Intramural Sports Partic.	1252	0.73	6000	0.62	0.11	16.9%	**1.20*** **
Q10B: Fitness Program Partic.	1255	0.33	6000	0.26	0.07	26.5%	**1.49*** **
Q10C: Club Sports Partic.	1261	0.80	6000	0.55	0.25	46.2%	**1.32*** **
Q10D: Outdoor Adventures Partic.	1240	0.18	6000	0.23	-0.05	-21.3%	1.04

†limited to instances where numeric data is available

Note. Significant differences in means indicates respondents do not generalize to the sample or population.

*$p<.05$, **$p<.01$, ***$p<.001$.

For 9 of 13 questions in this survey, measurement and nonresponse error move in the same direction to compound the threat to self-reported data validity. For the 6 of 9 questions with positive correlation between the biases, the bias results in statistically significant differences between self-reported and sample means. For example, respondent estimates of CRF use are at least twice the actual rate of use across the sample, so relying upon self-reports is ill advised. Social desirability bias and nonresponse error cause the inflated of CRF use reports. GPAs are also positively correlated to social desirability bias and nonresponse error, but the inflation in self-reporting is not as severe. The upward constraint of the grade point scale limits respondent ability to exaggerate; however, for co-curricular participation and academic metrics that rely upon a recall and count strategy, the measurement and nonresponse errors are not consistently related, and the biases may interact to improve or further weaken the estimate of actual behavior. Respondents are more likely to be involved and be academically successful, but the recall and count estimation strategy does not reliably predict measurement error. For easily forgotten behaviors, the measurement and nonresponse error move in separate directions and may improve the estimate of population behavior. Habitual and memorable events, however, are more likely to positively correlate to cause the respondent mean to become a more inflated estimate of actual behavior. For these behaviors, the evidence confirms concerns that the self-reported behavior is invalid and is not generalizable to the sample from whence it came.

Research Question 2

Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?

Research question 1 established the influence of measurement error and nonresponse error on self-reported means of behavior. Research question 2 begins to explain the biases by exploring the impact of perceived social desirability of a behavior on the accuracy of its self-reported survey statistic. Social desirability bias describes misreporting that systematically pushes self-reported behavior to conform to social norms. This question is explored first from the perspective of question topic area and how more uniform agreement around a social norm causes altered response behavior, causing (1) a higher share of respondents to misreport in a socially desirable direction; (2) misreporting respondents to further exaggerate their behavior in a positive direction; and (3) alter the response likelihood as the topic area changes the respondents' response likelihood. The second research question explores the relationship between a respondent's propensity for self-monitoring and their likelihood to misreport their behaviors.

Research Question 2A

Do social norms cause students to systematically over report the incidence of socially desirable behaviors such as physical activity, class attendance, and high grades earned? Are self-reported respondent means of socially desirable behaviors less valid estimates of behaviors than those from more socially neutral behaviors?

Consistent with the established literature, this study identifies systematic over reporting of physical activity, GPAs, and class skipping that inflates respondent means in a socially desirable direction. Social desirability bias describes systematic, and sometimes intentional, misreporting with the respondents' inaccuracy creating a more favorable presentation than is actually true. Tourangeau et al. (1997) found that exercise is among the topics most susceptible to social desirability bias, and this study also finds that respondents inflate self-reported physical activity by at least 65%. Kuncel et al. (2005) found that respondents over reported their cumulative GPAs 3.9 times more often than they underreported them. This study only finds a ratio of 1.3, though low GPA students systematically over report. Schmitt et al. (2003) concluded that self-reported class absenteeism is subject to social desirability bias, and this study finds that respondents underreport PE/HESF class skipping by 28%. Aside from informal CRF use, no studies link co-curricular behavior and PE/HESF course taking to social desirability bias, and this study does not find evidence that they are systematically affected by social desirability bias. For topics studied here, the literature and this study are aligned on the specification of social desirability bias.

As measured by the over reporting to under reporting ratio, Table 4.15 shows that CRF use and GPAs are the most suceptible to social desirabilty bias, though other behaviors are more randomly distributed around unbiased self-reporting. CRF use questions (Q1A & Q1B), GPA questions (Q2, Q3, & Q8), and class skipping (Q9) all have over reporting to underreporting ratios in socially desirable direction (e.g., above 1.00); though most questions have ratios that exceed 6.00. These questions are also among the most frequently misreported. Meanwhile, the ratios for other questions are not free of bias but are randomly distributed around a socially neutral ratio of 1.00. Topics outside of CRF use and GPAs may also be subject to social desirability bias, though the ratios are lower because those behaviors possess a less uniform definition of being socially acceptable. For example, the number of A grades earned reflects academic success for some; though others worry about presenting themselves as 'know it all.' This question also faces underreporting pressure as respondents fail to recall each instance of behavior, but unlike the CRF use questions, the social norms are not as uniformly felt. Social desirability factors are a function unique to each respondent with each behavior yielding uneven influence across respondents. By almost all, more physical activity is perceived as more is better, so social desirability bias overpowers the other sources of response bias. The literature provides an accurate inventory of socially desirable behaviors, including CRF use and GPA.

Nonresponse error also has a theoretical relationship with social desirability bias since a common reaction to sensitive questions is to avoid responding to avoid the disclosure of embarrassing or illegal information (Groves, et al., 2009). However, Figure 4.9 confirms the theoretical relationship between item nonresponse and socially desirability behaviors but observes the relationship inversely. Questions with the most observed social desirability bias also have the highest response rates. The questions for which respondents are most likely to inflate their behavior are also the questions for which the students possess more socially desirability and are also more likely to respond to the question. For socially neutral questions, however, measurement error and nonresponse error often move in opposite directions to help cancel. As such, Groves, et al.'s (2009) finding is confirmed because respondents grasp the opportunity to brag about their positive behaviors, causing social desirability bias to reduce item nonresponse.

Respondents report less biased behavior when the question's content is less directly associated with the social norm. The theory of social desirability bias presumes that there are social norms that define desirable behavior, and many respondents answer questions in an attempt to present themselves as complying with social norms (Kreuter et al., 2008). When identifying socially desirable behaviors, a summative metric like overall physical activity or cumulative GPA more directly associate with the social norm that it represents. For example, physical fitness is widely perceived as an admirable attribute, so a summative question about CRF use is more likely to invoke social norms. Meanwhile, social desirability bias is less impactful when the question requests descriptions about components of the overall behavior like participation in exercise-based activities. Participation in Intramural Athletics requires physical activity and contributes to physical fitness, but the question is perceived as less tied to the summative metric. As such, respondents rely more on memory and the recall and count strategy when describing components, and the effect of social desirability bias is mitigated. Another example is cumulative GPA, which is systematically over reported, yet the number of A grades earned are systematically underreported because respondents rely upon the recall and count strategy rather than impute missing values with social desirability

Table 4.15.

Over Reporting to Underreporting Ratio by Question.

	Number of Respondents					Over Reporting to Underreporting Ratio
	Under-report	Correct Report	Over Report	Other Misreport	Percent Misreporting	
Q1A: CRF Use - Prior 7 Days	17	268	202	0	45.0%	11.88
Q1C: CRF Use - Weekly Rate	33	240	229	0	52.2%	6.94
Q2: Cumulative GPA	324	586	407	0	55.5%	1.26
Q3: Semester GPA	306	431	555	0	66.6%	1.81
Q4: D/F/U Grades Earned	205	1106	29	0	17.5%	0.14
Q5: A Grades Earned	576	481	250	0	63.2%	0.43
Q6: HESF Course Taken	147	1113	56	0	15.4%	0.38
Q7: HESF 100 Course Taken	N/A	1024	N/A	193	15.9%	N/A
Q8: HESF Grade	3	439	36	106	24.8%	12.00
Q9: HESF Class Skipping	267	171	126	163	76.5%	0.47
Q10A: Intramural Sports Partic.	96	1019	138	0	18.7%	1.44
Q10B: Fitness Program Partic.	112	1048	96	0	16.6%	0.86
Q10C: Club Sports Partic.	133	855	274	0	32.3%	2.06
Q10D: Outdoor Adventures Partic.	234	912	95	0	26.5%	0.41
Total	2453	9693	2493	462	35.8%	1.02

Note. Unbiased reports have a ratio of 1.0; ratios above 1.0 suggest systematic misreporting in a socially desirable direction while ratios under 1.0 suggest systematic underreporting. "Other Misreports" describe incorrect self-reports that lack a numeric value.

in mind. This finding differs from the discussion of Halo Effect error found by Pike (1999b), where respondents inflate components of behavior yields to conform with other summative claims of socially desirable behaviors.

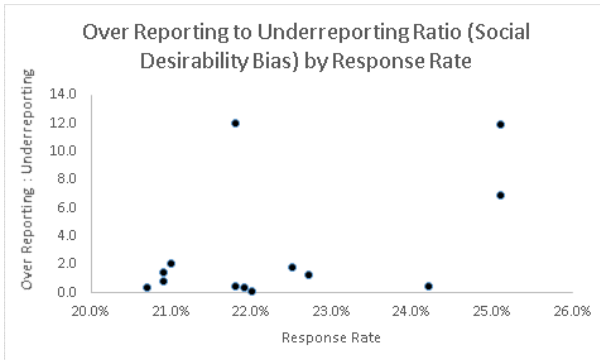

Figure 4.9. Over reporting to underreporting ratio by response rate.

From the perspective of social desirability bias, each question topic influences each respondent differently. For example, Brenner and DeLamater (2014) found that college students who score highly on an exercise identity survey are also more likely to inflate their recent CRF use; while low scoring respondents offer unbiased estimates. Identity theory finds that respondents who identify a behavior with their core personality will systematically misreport to better conform to their self-perceptions. Though this study does not contain an exercise identity inventory, I do have every respondent's historic CRF use and CRF-based co-curricular activity engagement. I hypothesize that historically high frequency CRF users are more likely to incorporate physical activity with their identity and systematically exaggerate their CRF use behaviors. Table 4.16 shows that the top quintile of respondents who, at any point in their academic career, have visited the CRF, most often also have the highest over reporting to undereporting ratio for present day CRF use questions. At 12.0, the ratio for top quintile users is almost double the 6.5 average among respondents from the lower quintiles. Respondents in the top quintile also have the highest percentage of respondents correcly describing their behavior, but this reflects their position on an extreme of the response option scale as much as ability to report accurately. For all questions, sitting on the endpoint of the scale improves response accuracy. For this question, the bottom quintile of users sitting on the opposite extreme are the second highest correct reporters. My measure of exercise identity is less theoretically proven than Brenner and DeLamater's, but I also find respondents who are likely to have incorporated exercise into their personality and have a higher propensity for over reporitng recent CRF use.

Table 4.16.

2015-16 CRF Use by Maximum Historical CRF Use Rate.

Quintile	Sample N	Response Rate	Pct. Correct Report	Over Report to Underreport Ratio
First (0 to 7 visits)	423	21.0%	14.2%	6.3
Second (8 to 21 visits)	387	20.9%	7.0%	6.6
Third (22 to 33 visits)	412	24.3%	10.9%	8.2
Fourth (34 to 52 visits)	387	29.2%	10.3%	5.1
Fifth (53 to 310 visits)	391	30.7%	17.4%	12.0

Note. Unbiased reports have a ratio of 1.0; ratios above 1.0 suggest systematic misreporting in a socially desirable direction while ratios under 1.0 suggest systematic underreporting.

Respondents react to complex questions with higher rates of misreporting in a socially desirable direction. Inflating self-reported behavior in a socially desirable direction is a common reaction to encountering a confusing question or when the behavior is difficult to recall. The unfamiliar term “club sports” likely caused confusion just as describing the prior 7 days of CRF is more confusing for the 56% of question 1A respondents for whom the spring university holiday was in their reference period. Meanwhile, describing CRF entries over an entire academic year is more complex than counting the prior 7 days of use. Unlike other questions about co-curricular participation, when respondents encountered the unfamiliar term “club sports” they over reported their participation at double the rate of other co-curricular participation questions. Respondents with spring university holiday in their reference period averaged an over reporting to underreporting ratio that exceeds 20 times the rate for those without the holiday. Respondents describing behavior over the prior seven days provided self-reports that are 32% more accurate than those describing the academic year-to-date. Respondents may approach the survey with the intention to accurately report their behavior, but when the question adds confusion, they systematically react by describing idealized version of behavior rather than work through the complexity to provide more accurate descriptions of behavior.

Research Question 2B

Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?

Questions 11 to 23 ask a series of true/false questions that combine to a shortened version the MCSDS that is theorized to measure the respondent’s propensity for self-monitoring. A higher MCSDS score indicates that the respondent is more likely to misreport their behavior to portray themselves in a more favorable light. Respondents earn one

MCSDS point per unlikely response to a true/false question with each point increasing the probability of the respondent being an HSM. For example, answering “True” to “No matter who I’m talking to, I’m always a good listener” earns an MCSDS point because it is highly unlikely that any college-aged individual is actually such a good listener and instead more likely that they are misreporting their responses for the sake of image management. The maximum MCSDS score is 13, and respondents scoring 13 are theorized to be the highest HSMs. However, this analysis finds that the MCSDS composite score is a poor predictor of the likelihood to misreport or over report behavior in a socially desirable direction.

The MCSDS scores are normally distributed around the test’s midpoint and are sufficiently robust for analysis, though no demographic groups have significantly higher MCSDS scores when compared to other respondents. In total, 1,246 of the 1,529 survey respondents answered all 13 of the MCSDS questions, and I only analyzed full MCSDS completers. MCSDS scores are normally distributed around a mean of 6.6 with a mode of 7.0 and a standard deviation of 2.78. Kurtosis scores of -0.16 and skewness score of -0.18 are consistent with a normal distribution. No demographic groups separate themselves with significantly higher or lower MCSDS score distributions. The highest scoring subgroup was Hispanic students with an average score of 7.50, and at 6.11, the lowest scoring subgroup was ethnicity “Unknown or Other Minority.” Gender and age differences were minimal. The MCSDS score was inversely correlated with cumulative GPA and SAT scores with the most academically successful students scoring lowest on the MCSDS scale.

The relationship between MCSDS score and misreporting is not consistent with the scale predicting self-monitoring behavior. Correct reporters score the lowest on the MCSDS, but misreporting in a socially desirable direction is unrelated to MCSDS score. The MCSDS score has a theoretically consistent relationship with inflated responses for only the cumulative GPA (Q2) and semester GPA (Q3) questions. I interpret this as a simultaneous correlation (i.e. multicollinearity) between GPA and reporting accuracy along with GPA and MCSDS score. A theoretically consistent relationship between social desirability bias and MCSDS is observed in one-third of the questions, which would be expected for values with a random distribution across three groups.

Question 1A, which requests prior 7 days of CRF use, provides the only evidence for the MCSDS's validity in this setting. Temporally based confusion for respondents whose reference period includes spring break caused multiple interpretations among respondents and resulted in exceptionally high rates of misreporting. The question's ambiguity could motivate those with higher propensity to self-monitor to over report more obviously. Among high MCSDS scoring respondents, the average score for correct reporters was 6.25, compared to 6.95 for over reporters (not significantly different with a t-stat of -0.68); meanwhile, correct reporters for respondents without spring break in the reference period averaged scores of 6.63 versus 6.73 for over reporters (also not significantly different). A temporal issue for question 1A invoked systematic inflating of physical activity, yet the high MCSDS scorers responded as low scorers

As shown in table 4.17, MCSDS score is also a poor predictor of systematic over reporting across all questions for a given individual respondent. Figure 4.10 is a box and whiskers plot describing the relationship between the MCSDS score and the percentage of all questions those respondents over reported upon sorted by the score on the MCSDS.

Table 4.17

Regression on MCSDS and Likelihood to Over Report

	All Respondents		First Wave Respondents Only†	
	Model 1: No MSCDS Score	Model 2: MCSDS Score	Model 3: No MSCDS Score	Model 4: MCSDS Score
Intercept	**0.450***** **(0.0386)**	**0.444***** **(0.0397)**	**0.502***** **(0.052)**	**0.496***** **(0.0544)**
Female	-0.015 (0.0071)	-0.015 (0.0071)	-0.013 (0.0093)	-0.013 (0.0093)
African American	-0.022 (0.0283)	-0.023 (0.0284)	-0.035 (0.0396)	-0.035 (0.0396)
Asian	0.007 (0.0292)	0.006 (0.0292)	-0.009 (0.0405)	-0.009 (0.0405)
Hispanic	-0.033 (0.0280)	-0.034 (0.0280)	-0.057 (0.0405)	-0.057 (0.0406)
International	-0.003 (0.0344)	-0.004 (0.0344)	-0.043 (0.0468)	-0.043 (0.0468)
Two Plus Races	0.024 (0.0285)	0.023 (0.0286)	0.012 (0.0391)	0.012 (0.0391)
White	-0.007 (0.0237)	-0.007 (0.0237)	-0.008 (0.0341)	-0.008 (0.0341)
Semesters Enrolled	**-0.008***** **(0.0017)**	**-0.008***** **(0.0017)**	**-0.008***** **(0.0023)**	**-0.008***** **(0.0023)**
Cumulative GPA	**-0.078***** **(0.0067)**	**-0.078***** **(0.0067)**	**-0.085***** **(0.0090)**	**-0.085***** **(0.0091)**
Age	0.001 (0.0010)	0.001 (0.0010)	-0.000 (0.0013)	-0.000 (0.0013)
Number of Co-Curricular Activities	**0.021***** **(0.0033)**	**0.021***** **(0.0033)**	**0.020***** **(0.0042)**	**0.020***** **(0.0042)**
MCSDS Score		0.001 (0.0013)		0.001 (0.0017)

†Response wave 1 includes spring university holiday in the 7 day reference window; response waves 2 and 3 include no university holidays.

Note. OLS Regression. *For model 1 and 2,* N=1233; for model 3 and 4 N=695. Adjusted R-squared for model 1 is 0.115; for model 2 0.114; for model 3 0.131; and for model 4 0.130.

**p<.05, **p*<.01, ****p*<.001.

Theoretically, the percentage of questions over reported in a socially desirable direction should increase with the MCSDS score, and the plot should thus possess an upward trend increasing from left to right. However, no such relationship exists, and the correlation between MCSDS score and percent of questions over reported upon is 0.04 and not significant at the 0.05 level. An OLS regression model using MCSDS as an independent variable that controls for cumulative GPA, gender, ethnicity, age, and number of co-curricular activities finds an MCSDS score parameter estimate of 0.001 that is also not statistically significant at the .05 level.

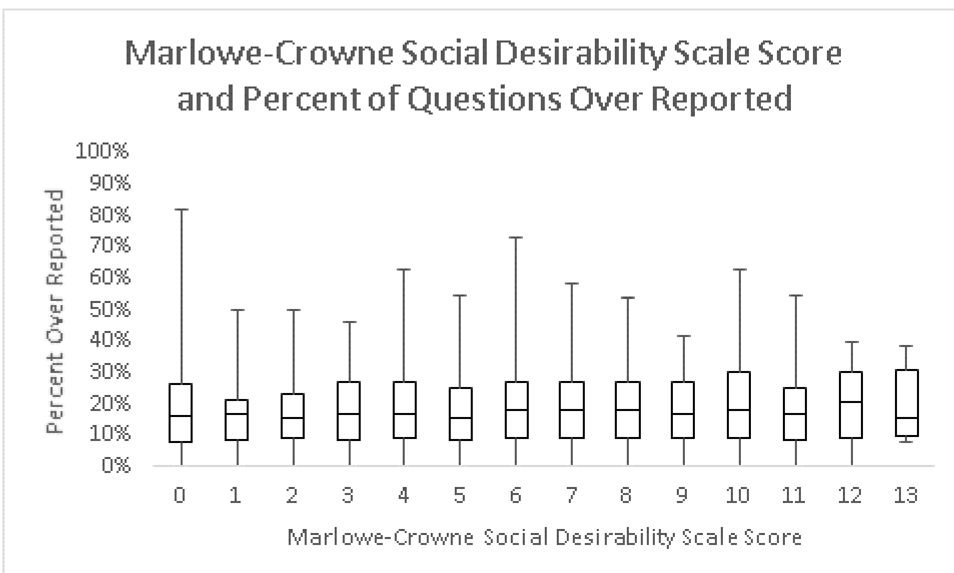

Figure 4.10. Over reporting rates by MCSDS score.

There is little evidence to support a statistically significant relationship between MCSDS score and systematic over reporting or misreporting. Even in question 1A which invites social desirability bias, the MCSDS score is not a statistically significant predictor of over reporting. There is no MCSDS-based evidence that misreporting is motivated by HSM image management as measured by the MCSDS. This finding does not necessarily generalize to other subject matter, studies, or settings for 4 reasons. First, my study uses Ballard's (1992) shortened scale, and this subset of questions may not be a valid substitute for the full array of MCSDS of questions. Second, there may be a mismatch between what I expect to be a socially desirable response and what the respondents view as socially desirable. For example, respondents systematically underreport the number of "A"s earned over their career that could result from a fear of being a 'know-it-all;' though I believed that HSMs would inflate the number of "A"s earned. Third, adept self-monitors could understand the intent of the MCSDS and strategically misreport on the MCSDS to avoid the HSM label, as described by Mueller-Hanson et al. (2003). Finally, social desirability bias may be less prevalent in online surveys because respondents have no contact with researchers and are less motivated to provide socially desirable responses because there is no interviewer to impress.

Conclusion

Researchers are increasingly reliant upon online surveys, which articulates a need to investigate the validity of the self-reported data. This study tests for the presence of measurement and nonresponse errors for a variety of topics that are commonly studied by higher education researchers. I find that bias threatens the validity of most topics studied. This chapter's review of the research questions confirmed the presence of measurement and nonresponse error in nearly all of the survey's questions. For most questions, the significant differences cause substantive deviations of survey means from actual means. A question-by-question review of responses, measurement error, and nonresponse error is located in the Appendix. For CRF use questions, the Appendix also includes a discussion of response accuracy across the three response option types including (1) number of visits in the prior 7 days, (2) vague use across Likert Scale options inspired by NSSE, and (3) year-to-date, rate based use.

Researchers must accept the risk of bias and cannot presume that their self-reported data is valid. The topic areas of CRF use and GPAs are systematically over reported, though respondents show little ability to accurately self-report other topics including PE/HESF course engagement, grades earned, and co-curricular participation. High cumulative GPA students are the most accurate self-reporters, and other background and behavioral measures are only loosely correlated with response accuracy. The MCSDS score does little to explain systematic over reporting. Instead, over reporting is systematic when respondents encounter complex questions or display satisificing behaviors. Overall, my results confirm my concerns that the validity of research conclusions are dependent upon survey data.

CHAPTER 5: IMPLICATIONS AND CONCLUSION

Survey methods are an indispensable tool in the study of higher education because sampling and survey data collection are cost effective methods to estimate the occurrence of behaviors across a target population (Pike, 1999a). Customized survey instruments enable the collection of targeted behavioral measures to specifically address a researcher's question (Kuh et al., 2001). For these reasons, survey research will continue to thrive and expand. However, this study finds that evaluating self-reported behavior as an unbiased and accurate measure of actual behavior is a requirement before accepting research conclusions based upon survey data. An unmet need remains for validation studies that evaluate the survey statistic as an estimate of actual target population behaviors (Porter, 2011). This study is relevant to higher education research because it meets the need for validation studies and justifies continued research on the validity of self-reported data as a measure of actual behaviors among respondents and target populations.

How students participate and engage in collegiate curricular and co-curricular programs and how participation affects student outcomes are a highly researched topic in the field of higher education. Because researchers rarely have access and/or ability to collect institution reported participation records for such experiences, reliance has been placed upon self-reported survey data. For a sample of North Carolina State University (NCSU) undergraduates, this study tests the validity of self-reported student behaviors using institution-reported validation data on CRF use, co-curricular participation, PE/HESF 100 class attendance, and academic performance. This study questions the presumed validity of self-reported behavior by testing the hypothesis that respondents inaccurately self-report socially desirable behaviors and more accurately describe recent and noteworthy events. Results help researchers evaluate the validity of self-reported data by topic area and evaluate influence of nonresponse error.

Measurement errors is important to researchers who study college student behavior. Measurement error refers to differences between the information that the researcher desires to collect and what is actually collected and attributed to the survey methods chosen (Groves et al, 2009). Measurement error is expected and unavoidable, but randomly distributed errors rarely bias the survey statistic because over reporting and/or underreporting tend to offset. However, the presence of systematic, non-random reporting errors is concerning because misreporting in one direction is not offset by corresponding misreporting in the opposite direction, which leads to biased self-reported means. Systematic, non-random errors that are committed across large portions of the respondent pool cause measurement error in the survey statistic. Common causes for the response bias include an inability to retrieve the requested information properly and intentional misreporting of the information to present one's self in a more favorable light (Groves et al., 2009). This study detects systematic response bias in all topic areas studied.

Meanwhile, nonresponse error refers to behavioral differences between survey respondents and those of nonrespondents that are attributed to systematic patterns in which members of the sample respond to a survey request. Nonresponse error is unrelated to the survey instrument but is attributable to differences between the respondents and target population (Groves et al., 2009). The information that the researcher collects, even if reported accurately, is not true of the target population when nonresponse error is present. In some cases, nonresponse error and measurement error are uncorrelated or may counteract the effect of the other; in such cases, bias may not threaten the validity of research conclusions. However, where the cause of nonresponse error is related to the data collected and/or the biases are positively correlated, bias is a threat to the validity of the survey statistic and research conclusions.

Theory, Previous Research, and Implications

The Model of Response Process (Tourangeau, 1984, 1987) is a four part, theoretical framework for assessing the validity of self-reported data. A valid survey response requires the respondent to (1) comprehend the survey question as intended; (2) correctly retrieve the information requested; (3) make good judgments when imputing and reporting missing memories; and (4) report their response accurately and truthfully. Participants are known to inaccurately respond to survey questions for several reasons; some reasons are intentional and others accidental. Many respondents are unable to accurately retrieve memories and provide behavioral descriptions using circumstantial evidence rather than actual memory (Tourangeau, et al., 2000). Other respondents respond to questions about sensitive subjects with intentional fabrications or by inflating their socially acceptable behaviors (e.g., social desirability bias). Within the context of the Model of Response Process (Tourangeau, 1984, 1987), this study finds that the validity of self-reported data is compromised by satisficing, the failure to recall past behavior, and by social desirability bias. The integrity of self-reported data is further threatened by nonresponse bias.

Research Question 1

To what extent can students accurately self-report past co-curricular and academic behaviors to researchers of higher education? Can higher education researchers presume students provide accurate self-reported behavioral information?

Research Question 1A

What student characteristics are associated with the most accurate and inaccurate self-reporters? Do students with high standardized test scores or collegiate GPAs more accurately self-report behaviors? Do participatory students provide more accurate self-reported behavioral data? Do demographic or background characteristics affect accuracy in self-reporting?

This validation study uniquely forwards the evaluation of survey data as a valid estimate of actual behavior because the institution-reported validation data can describe a diverse array of behaviors that are tracked at the unit level across 14 questions. This study's behaviors are not new to validation study as Brenner and DeLamater (2014) studied CRF use, Kuncel, et al. (2005) studied academic performance, Schmitt, et al. (2003) studied class skipping, and Pike (1999a) studied co-curricular participation. However, each of the aforementioned validation studies are limited to a single behavior or by unit level validation data that limits analysis to group comparisons without shedding light upon individual response patterns. For example, Tourangeau, et al. (1997) studied an array of illicit behaviors but lacked unit-level validation data, so estimates of social desirability bias are based upon mean comparisons of an experimental and control group. My validation study finds systematic over reporting correlated with respondent characteristics. Specifically, social desirability bias affects the validity of self-reported behaviors of GPA and CRF use because respondents are disproportionally motivated to over report behavior in a socially desirable direction.

Meanwhile, respondent characteristics affect the systematic underreporting of earning A grades, earning D/F/Unsatisfactory grades, class skipping, and participation in co-curricular activities. A respondent's cumulative GPA predicts whether a respondent will provide more accurate estimates of past behavior and be less susceptible to social desirability bias. The complexity of a student's career predicts response accuracy because simpler events are easier to report correctly. High GPA students are more accurate respondents because GPA assesses a student's ability to comprehend questions, accurately recall facts, and then organize information to effectively meet the requester's expectations. Students who do not participate in co-curricular activities accurately describe their lack of participation, either because recollections are accurate or because they know themselves to dislike such co-curricular activities. Finally, students recently enrolled in college have a less complex student career to describe because their experiences are less robust and because behaviors are more recent and easier to recall.

Respondents systematically over report composite measures of CRF use and GPAs and underreport the components of these measures. For example, A grades earned are systematically underreported, and cumulative and semester GPAs are systemically over reported. Similarly, co-curricular participation measures in this study require participants to enter the CRF, yet CRF use is systematically over reported, and co-curricular participation is systematically underreported. For co-curricular participation, high GPA students are under reporters, and low GPA students are accurate reporters. I hypothesize that high GPA respondents rely more on memory and the recall and count strategy; whereas, low GPA students use imputation strategies. Because one-time co-curricular participation like Outdoor Adventures is not noteworthy and easily forgotten, it is frequently underreported. To a lesser extent, this is also true of grades earned and class skipping. Because high GPA students are over represented among respondents, the results of their estimation strategy are over represented, and the components of GPA and CRF use are predominately underreported. Had more low GPA students responded to the survey request, the component measures may have been more accurate measures of actual behavior.

Implications. Responses from high GPA students should receive more weight and be presumed more accurate than those from low GPA students, especially for behaviors occurring in the prior academic year. However, low GPA respondents are likely to inflate their GPAs, so the high/low GPA distinction must be derived from institution reported GPAs. Fortunately, cumulative GPA is readily accessible by institutional researchers and registrars' offices who also readily provide email address lists to contact students with survey requests. When the target populations are entirely composed of inherently academically successful students with high GPAs, then the survey results are more accurate. For example, graduating senior surveys only contact students who successfully navigated the academic experience. Studies that can tolerate GPA homogeneity should over sample from populations of high GPA respondents and also provide their responses more weight. However, this implication is useful for survey research of at-risk students that require self-reported data from populations with robust GPAs.

Respondents are more likely to accurately respond when the response options are limited to Likert scales. For example, this study's respondents accurately described behavioral extremes like "none," "below average," "average," and "significantly above average." Non-participatory students accurately report their abstention, and participatory students somewhat accurately report participation intensity, though they cannot accurately report more nuanced participation behaviors like number of CRF entries. Self-reported data is not fit for analysis on a continuous scale because the numbers are not accurate descriptions of a linear progression through a scale. Respondents can accurately sort themselves into categories relative to the behaviors of neighboring response options. For example, those reporting that they entered the CRF twice per week did enter the CRF more often than those reporting once per week, but they did not enter twice as often. Providing numeric scales of use may be helpful for the respondents to orient their responses, but researchers cannot interpret the scales literally. See the Appendix for a further comparison of response options for CRF use.

There is no evidence that demographic groups answer survey questions with significantly different accuracy rates, and when there are observed differences, they are more directly explained by GPA or complexity of the student's career. For example, the oldest and youngest respondents most accurately describe their behavior, but age is not the best explanation for differences in recall ability. Instead, old and young students possess less complex academic careers that are easier to describe. First-year students possess shorter careers, and older students are nontraditional students who are less likely to have residential college experience and are less involved in campus life, making their nonparticipation easier to describe. Demographic differences, however, are important to consider in the survey design phase. International and other ESL students report with equal accuracy but require more time to complete the survey and offer shorter open-ended question responses. Target populations containing disproportionate numbers of non-native language speakers should provide more time for survey completion and carefully craft questions to meet their language needs. Older respondents require more time to complete the survey and need shorter open-ended question responses because they are less comfortable with technology and may struggle with smart phone survey applications that require the use of mobile phone keyboards. Standish, Joines, Young, and Gallagher (submitted for publication) find that all respondents offer longer open-ended responses when they use full sized, desktop/laptop computer keyboards.

Students with shorter academic careers recall their behavior more accurately, so I recommend that survey researchers collect self-reported data during a student's first year and then archive the survey results for later study. Student outcomes are not known until several years pass. However, the passage of time also reduces the validity of self-reported data. Archiving survey data possesses other benefits because it provides a more robust set of behavioral data and is collected from both students who will ultimately withdraw and those who persist. First year students have not yet withdrawn from the university, meaning first year data collection will capture a more complete description of behavior that better informs the relationship between behavior and outcome. Conducting multiple surveys and archiving self-reported data for later use will improve the quality of data.

Research Question 1B

Does the passage of time affect the accuracy of self-reported behaviors? Are behaviors that occurred in recent days more accurately described than behaviors occurring months or years in the past?

Recall is a key component of the Model of Response Process with the passage of time affecting the ability to accurately report behaviors. Factors affecting recall ability are both unit-level (i.e., memory) and situational. Holding all else constant, recently occurring events are easier to recall and describe compared to events further into the past. Occurrences in recent days can be retrieved with more ease than those in past weeks or months. Researchers can expect more accurate self-reports when a question is well written, asked in a context that the respondent understands, and the question topic is not inherently subject to social desirability bias. For example, respondents often fail to recall participation in Outdoor Adventures and formal fitness programs; however, more recent participants more accurately describe their participation in these programs. Question complexity affects response quality. When a reference period includes a university holiday, respondents may realize that the question is temporally awkward and re-interpret the question as “describe a typical seven days of CRF use.” However, they describe an idealized seven days rather than a typical seven days. Meanwhile, when the reference period contains a normal seven-day period, the responses are accurate. Respondents’ CRF use spanning months or years into the past should be viewed with skepticism.

Implications. Researchers can neither ignore the correlation between recall ability and the noteworthiness of the behavior nor the passage of time. Whenever possible, surveys should collect reports of more recent behaviors and noteworthy events. In the context of college student behavior, noteworthy events include long-standing, habitual behaviors like semester-long classes and season-long intramural athletic sports. Response options should be described as binary yes/no participation or on a limited scale because the nuance of participation rates are quickly forgotten. Standalone events like a weekend camping trip with Outdoors Adventures or yoga formal fitness class are not well recalled months or years later, and such questions should be avoided.

Researchers should collect behavioral data using shorter reference periods that are more frequently collected. The same cost reductions that drive the expansion of surveys can be redirected to implement frequently occurring, simple surveys. Examples include activity journals, one question text message surveys, and composite measures that summarize long-term behavior. Porter (2011) recommends activity journals to collect behavioral data using daily reference periods that can be compiled to describe behavior over long periods. Shepherd (2003) recommends one question, text message surveys that enable the creation of longitudinal behavioral profiles even when individuals only respond to a fraction of the requests. Composite measures are updated frequently and easily track behavior over longer spans; examples include cumulative GPA, student loan balances, and participation-based activity rankings.

One potential exception is that socially undesirable behaviors may be reported less accurately when memory is clearer and, instead, the events are intentionally misreported. A clear recollection of the events motivates some respondents to place qualifiers on behaviors that lead them to be discounted as inapplicable. For all other questions, high GPA students reply more accurately and should be better at recalling absences than low GPA students. Respondents with high GPAs and who took a PE/HESF 100 level course in the prior semester underreport class skipping; whereas, the corresponding low GPA respondents do not, nor do high GPA respondents describe class skipping further into the past. I interpret this as respondents recollecting the reason for skipping a class and then omitting it from their report after judging it irrelevant. For example, if a respondent skipped class for an illness, he or she may not associate this with researcher's intent and fail to report the absence; meanwhile, this same student would report a class skip for recreational purposes.

High GPA respondents may be overly confident in their recall ability, leading to less accurate self-reports of easily forgotten events. High GPA respondents discount the possibility that their memories are incomplete and fail to incorporate imputation as a recall strategy and systematically underreport. Meanwhile, respondents with lower recall ability more accurately describe easily forgotten behaviors because they are quick to rely on imputed values. For example, formal fitness programs are easily forgotten behaviors that low GPA respondents more accurately report when they think, "that's something I would do" (i.e. imputed memory) rather than high GPA respondents who think, "I don't recall doing that, and I trust my memory."

Overall, assessing the validity of self-reported data is a complex balancing act of rules, tendencies and assessments of social norms. I recommend that researchers attempt to obtain institution reported data whenever possible. Surveys should be reserved for instances where institution reported data is impossible to obtain.

Research Question 1C

Do vague qualifiers or excessively complex response options affect the validity of self-reported behavior?

When the reference period is typical and unambiguous, the prior 7 days of CRF use yield the most accurate self-reports of the three response option formats offered in question 1. In addition to a high correlation with institution reported values, the self-reported prior 7 days of CRF use maintains a high correlation with use periods further into the past. This finding confirms Shepherd's (2003) finding that the prior 7 days of use is the best metric of overall physical activity when measured by CRF visits. However, all three response option formats for CRF use have a 0.65 or higher correlation between self-reported use and the maximum semester visits over a given student's career; this correlation is comparable to the correlation between actual prior 7 days of use and self-reported prior 7 days of use (0.69). When response options introduce ambiguity, respondent descriptions of physical activity are more aligned with an idealized period of use rather than behaviors specified by the question. This finding is consistent with the theory of social desirability bias and identity theory described by Brenner and DeLamater (2014).

Implications. An analysis of the validity for self-reported CRF use across three types of response option formats finds that no alteration to the question's response option design substantively counters the effect of bias. The response option format requesting prior 7 days of CRF use yields marginally more accurate self-reports because it is the least vague and mentally taxing. For all response option formats, respondents are prone to systematic over reporting of CRF use in a socially desirable direction. For all response option formats, respondents sort themselves into groups possessing statistically significant different means that are suitable for comparative analysis across that format's response options. However, individuals do not validly or reliably respond to any response option formats. Accordingly, respondents who choose an identical response option display heterogeneity in actual CRF use such that researchers cannot presume that an individual accurately describes his or her use or that the response option's mean is an accurate description of actual use. Respondents self-report CRF use as a function of actual use over the reference period and an idealized use rate that conforms to the respondent's identity. Further, nonresponse bias and response bias are positively correlated, so the survey statistic is not generalizable to the target population. Selecting the optimal response option format requires situational considerations that marginally improve self-reported validity, but the survey statistic is threatened by bias.

Vague qualifiers have limited effect on overall response quality but negatively affect the validity of high frequency CRF users who minimize their behaviors. Question 1A precisely requests a count of CRF entries over the prior 7 days; question 1C somewhat precisely requests use with a rate-based visits per week scale; and question 1B employs vague, Likert-based descriptions of behavior. For all three response option formats, only respondents reporting no CRF use provide accurate self-reports, and moderate users systematically inflate their use. Meanwhile, high frequency users reliably respond to precise response options that quantify the use but inconsistently label their behaviors using the vague response option format. For example, over 10% of high volume users describe their CRF use outside the highest use response option despite being in the top quintile of users; yet when using precise response option formats, only 2% of the top users failed to place themselves into the highest option. High frequency CRF users create social networks of similar users, and the peer group they create causes its members to underestimate CRF use relative to all students. Precise response options help orient respondents.

Complex response option formats increase the likelihood that respondents will inflate socially desirable behavior. Recalling and counting the prior 7 days of use is the least complex of the response option formats and yields the most accurate self-reports as measured by correlation but with actual use and given the caveat that the prior 7 days is a typical week. Asking respondents to recall further into the past and then transform for a rate-based scale is more complex. For respondents describing a typical week, there is a 0.71 correlation between institution and self-reported use, but the correlation drops to 0.39 when the prior 7 days includes the spring university holiday. Meanwhile, for the rate-based and vague response option formats, the atypical week does not affect response behavior. I hypothesize that an atypical reference period causes confusion about the question's intent, and respondents do not reliably interpret the question. Some take the question literally and describe the atypical use, and others believe that the question is flawed and provide information that they believe the researchers desire (e.g., CRF use in a typical week). However, respondents employing the latter approach systematically inflate CRF use to levels that are correlated such that they align with peak use rates rather than a typical week. Complex questions force respondents to employ more types of estimation methods and, for socially desirable behaviors, respondents have more opportunity to systematically mis-estimate in a socially desirable direction.

Regardless of response option format, the correlation between self-reported and institution reported CRF use metrics does not vary much, so precision and complexity of response option formats only affects the accuracy on the margins. For CRF use, long term and short-term behaviors correlate enough that capturing one period also describes behaviors during the other period. Among respondents, the correlation between actual March 18 and 24, 2016 visits and the actual average weekly visits is 0.70; the correlation between actual March 18 and 24, 2016 visits and the peak career use is 0.50; and the correlation between actual average weekly visits and peak career use is 0.72. Capturing one of these behaviors explains the majority of variance for the others. Meanwhile, changing the reference period from prior 7 days to an academic year only marginally increases the correlation between self-reported and institution reported use. Requesting the prior 7 days yields a 0.71 correlation between self-reported and institution reported over the prior 7 days and 0.65 between self-reported and average weekly visits. Meanwhile, requesting the average weekly use rate yields a 0.53 correlation between self-reported and average weekly rates and 0.48 for prior 7 days. Because the response option format is least vague and complex, I recommend that survey researchers request very specific, short-term information that will not overburden respondents.

However, short term behaviors are affected by temporal factors that can significantly bias the extrapolation into longer term behavior when the prior 7 days of use is typical; the results are most accurate, but the logistics of finding a typical week are not straightforward.

An analysis of CRF use shows that atypical use periods are elusive and cannot be identified with sufficient lead times for best practice survey implementation. Some atypical CRF use periods are predictable; for example, university holidays, final exams, major holidays like Thanksgiving and Easter, and lower profile holidays like Valentine's Day and Halloween reduce CRF use. Other periods of atypical use cannot be predicted with enough lead time to affect IRB approval and pre-notification letters; for example, inclement weather and high profile, revenue sport NCAA competitions also reduce CRF use. Meanwhile, other periods possess atypically high CRF use but may not be intuitive; for example, use is higher during the middle of the fall semester and shortly after New Year's resolutions are made. Typical use periods are rare and unpredictable, so though collecting short-term behavior is recommended, survey researchers must proceed with caution.

CRF use is a topic that is highly susceptible to social desirability bias, and researchers must view self-reported data with skepticism. No response option definitely overcomes the structural obstacles of collecting such behaviors with a survey instrument. Researchers can make valid comparisons between groups of respondents to make general conclusions about behaviors between groups. However, researchers cannot assume that respondents offer accurate descriptions of behavior or that they reliably interpret survey questions unless the respondent self-reports no CRF use. Though no set of response options separates itself as superior to the others, a precise and less taxing question that asks respondents to recall and count their prior 7 days of CRF use yields the most accurate self-reports. From a typical period, short-term behavior is sufficiently correlated with other longer-term behaviors to answer a variety of research questions. However, finding a typical 7 day period from which to collect data is not intuitive or even predictable.

Research Question 1D

Are sample means of self-reported behaviors valid estimates of target population means?

For topics surveyed in this study, researchers cannot presume that sample means are valid estimates of population means. Sample members possess unequal probabilities of completing the survey with academically successful and participatory students disproportionately represented among respondents. If respondents self-reported accurately, the sample means collected in this study could be presumed to exceed population means. Because GPAs and CRF use self-reports are inflated, the respondent means further extend past population means. Researchers must assume that self-reported means are not representative of target populations, especially for socially desirable topics where the nonresponse error and measurement error move in the same direction.

Implications. Results from this study show that nonresponse error is a significant issue that survey researchers should address. Results from this study provide evidence that bias should be presumed present and "the tacit agreement in postsecondary research … that validity is assumed until proven otherwise" (Porter, 2011, p. 73) should be discontinued. In 11 of 13 questions, nonresponse bias caused significant differences between respondents and nonrespondents among 6 of 13 questions, and the magnitude of the differences exceeds 20%. Researchers would be ill advised to use results as if they were generalizable to the target population. These results are generated from a random sample and would be amplified by other sampling strategies. For example, convenience samples made up of subjects invited to participate because they are present in the CRF would further increase the effect of nonresponse error.

Research Question 2

Does the perceived social desirability of a behavior being reported lead to systematically invalid bias from over reporting positive behaviors?

Research Question 2A

Do social norms cause students to systematically over report the incidence of socially desirable behaviors such as physical activity, class attendance, and high grades earned? Are self-reported respondent means of socially desirable behaviors less valid estimates of behaviors than those from more socially neutral behaviors?

Though present in many question topics, social desirability bias is not the only source of bias leading to invalid self-reports. For some questions, social desirability bias is responsible for most of the inaccurate self-reporting but is a minor influence for others. The question's topic does not necessarily predict the presence of social desirability bias because questions about components of a behavior are differently affected than a summative measure of the same behavior. For example, students are prone to exaggerate their cumulative GPAs, and respondents in this study provide higher GPAs than their corresponding institution mean. Meanwhile, the lack of recall ability results in systematic underreporting for the number of A grades earned, which should imply underestimated GPAs. The challenge for researchers is to predict both the influence of social desirability bias and how the bias will impact the accuracy of self-reported behaviors for a given question topic and/or response option format. Of the 13 questions, 8 questions have self-reported means significantly different from institution means, but only 5 questions differ from the means in a socially desirable direction. In addition to GPAs, CRF use is also drastically over reported, though other behaviors like class skipping and co-curricular participation are not. Intuitively, there is no obvious difference in social norms for these topics, yet respondents systematically and reliably perceive these social norms. Given the relative scarcity of validation studies, particularly validation studies about rarified college student behaviors, researchers face an unknown risk of bias for each self-reported behavior they collect. The literature can identify affected topics, but each question format requires further study.

For questions about aggregate CRF use, social desirability bias is a debilitating factor with self-reported means at least two thirds higher than institution reported means. Researchers may employ 3 strategies to minimize bias. First, they may sidestep some bias by asking clear questions with simple answers. For example, the temporal issue of the spring university holiday increased measurement error for the question about the prior 7 days of CRF for respondents with a clear reference period, though it was reported accurately among respondents with a typical reference period. Second, researchers may ask about components of the summative behavior in ways that are less likely to invoke socially motivated misreporting. However, asking respondents for overly specific descriptions of component behaviors leads to systematic underreporting because respondents fail to retrieve memories from the distant past. Finally, the researchers can provide Likert Scale response options that vaguely describe the rate of use. When asked to describe CRF use on a scale of “Never” to “Very Often,” respondents successfully sort themselves into CRF use groups whose use levels are ordinally appropriate and dissimilar from the abutting groups of respondents.

However, there is not a linear relationship through the ordinal groups, so data cannot be computed as if on a continuous scale. Further study may clarify the issues surrounding self-reported physical activity, but self-reported descriptions cannot be depended upon as valid estimates of actual behavior.

Ineffective question design increases the impact of social desirability bias. Descriptions of CRF use over the prior 7 days are far more inflated when the question is asked on a temporally inappropriate date. When the prior 7 days are typical (e.g., do not include a university holiday), requesting the prior 7 days of CRF use is uniformly understood and more accurately answered. However, when the prior days include a university holiday, respondents choose from three interpretations of the question: (1) respond to the question literally, (2) describe a typical week outside the specified reference period, or (3) provide a summative measure of general physical activity, irrespective of actual use. In this study, respondents of vague questions evenly distribute themselves into the three categories which results in inflated respondent mean. If the question is designed properly, respondents are much more likely to sort themselves into the first decision group and provide accurate self-report.

Respondents who self-report abstention from a socially desirable behavior are the most likely to offer an accurate description of actual behavior. Intuitively, reporting a socially undesirable behavior signals no motivation to misreport. Lack of participation is easier to recall than the degree of participation, and nonuse offers the most straightforward response option. Most respondents who report participation actually have engaged in the behavior, but the intensity of their participation is often difficult to recall accurately and then describe within the constraints of available response options. When respondents are confused, they are more likely to offer responses affected by social desirability bias. As such, self-reported abstention for socially desirable behaviors is more likely to be accurate than other self-reported behaviors.

Groves, et al. (2009) found that nonresponse error is most threatening when the likelihood to respond is correlated with the survey topic. Results from this study are concerning because GPA and CRF use topic areas are correlated with the likelihood to respond and are also the primary topic of the survey. Even worse, research reveals that these topics suffer from significant social desirability bias. This study confirms prior research that social desirability bias also causes nonresponse errors among low GPA students (Kuncel, et al., 2005) and physically inactive individuals (Shepherd, 2003). This study adds further concern about the validity of self-reported means on the topics of high GPA students and high frequency CRF users because those characteristics are also correlated with response likelihood rates.

Implications. Determining which behaviors are subject to social desirability bias is a challenge. Tourangeau, et al. (1997) used a false pipeline study to test a variety of behaviors. They found that topics susceptible to social desirability bias include (listed in descending order from most susceptible to least): drinking more than average, exercising four or more times per week, using amphetamines, using other drugs, drinking more than you should, often having oral sex, using cocaine, and drinking and driving. There is no apparent pattern to the most and least desirable of the aforementioned behaviors. The lack of predictability presents problems for researchers attempting to control for social desirability bias. To be problematic, socially desirable behavior must affect large swaths of a population because a bias for a small share of the population does not meaningfully alter sample averages. Cocaine use, for example, is likely underreported due to social desirability bias, but because a small percentage of the population are habitual cocaine users, the bias is not widespread and minimally affects the survey statistic. Next, social desirability bias is problematic when respondents feel that the misrepresentation is harmless and exaggeration is without consequence. For example, career center GPA is reported more accurately than GPA for an optional survey among the same respondents (Krueter, et al., 2008). Perceived accountability for misrepresentations leads to more accurate self-reports of legal behaviors (prior research suggests accountability leads to underreporting of illegal behaviors [Groves, et al., 2009]). Problematic social desirability bias must also be a behavior that people widely agree is desirable or it will not cause widespread misreporting. For example, Brenner and DeLamater (2014) found exercise identity present in the majority of their study participants, and the presence of exercise identity respondents leads to more exaggerated reporting. The misreports and lack of passion from non-exercise identity respondents lead to more accurate reporting. Finally, social desirability bias is problematic when a social norm is agreed upon by nearly all members of a target population. Data from this study finds that 97.8% of the respondents believe that physical activity is important to academic success, so widespread over reporting is predictable. Further validation studies are required to expand the universe of behaviors known to be threatened by bias.

Further study is required for topics where likelihood to respond is negatively correlated with the direction of social desirability bias and the direction of nonresponse error.

CRF users and high GPA students who possess socially desirable behavior are more likely to respond because they seek out the opportunity to describe their good behavior. Conclusions may differ in cases where more of the behavior is considered undesirable and abstinence is positively correlated with response likelihood. For example, these findings may not apply to studies about drug addiction where addicts are presumed to be underrepresented among respondents. Addicts may not respond to avoid discussing their vice and also because their addiction may decrease their motivation to comply with a survey request. A more thorough analysis of response accuracy for ‘less is better’ behaviors is a recommended next step.

Even if social desirability bias is a known threat, further analysis is required to determine if the presence of bias is a prima fascia invalidation of subsequent research conclusions. If CRF use is uniformly inflated, it may cause a level shift that maintains the integrity of research conclusions even while providing misleading rates. For example, it may be true that those self-reporting 3 CRF visits per week actually average 3 visits, but this may be inconsequential if those self-reporting 3 visits actually visit 1 time, and so on. The motivation to misreport CRF use and GPAs in a socially desirable direction is self-evident, but the magnitude of the social desirability bias in these questions is unexpectedly significant. The next step is to compare research conclusions from self-reported data to those from institution reported data.

A more reasonable approach to minimizing bias is to identify members of the target population likely to provide less biased descriptions and overweight their responses. High GPA students and students with less complex collegiate histories are more accurate self-reporters and identifiable from institution reported data. Researchers can weight responses from these students more heavily or only collect self-reported data from members of the sample frame with such characteristics.

A final strategy to minimize social desirable bias is to offer the survey in a manner that minimizes motivation to misreport or provide the appearance that responses are evaluated for accuracy (Groves, et al., 2009). For example, surveys generate less biased self-reported GPAs when respondents believe their self-reported GPA accuracy is connected to hiring status at the college career center (Dobbins, et al., 1993). In a face-to-face interview, Tourangeau, et al. (1997) increased the perceived accountability by leading respondents to believe that he could detect respondent lies. Online surveys help minimize the severity of misreporting on socially desirable questions because the mode's lack of interpersonal interaction lowers the stakes for accurately reporting socially undesirable behaviors, thus minimizing the threat of judgment by interviewers (Kreuter, et al., 2008). Self-reported behaviors that are collected online are likely to be more accurate and less affected by social desirability bias.

Research Question 2B

Does the MCSDS score improve predictions for the probability that a student over reports socially desirable behavior and/or underreports a socially undesirable behavior?

There is little evidence to support a statistically significant relationship between MCSDS score and systematic over reporting or misreporting. Even for questions that contain widespread misreporting, the MCSDS score is not predictive of social desirability bias. There is no reliable relationship between the MCSDS score and misreporting for prior 7 days of CRF use, even when the reference period includes the university holiday. Neither is there a relationship among questions with systematic underreporting. This finding is a departure from the study of social desirability bias but is consistent with findings that online surveys are less susceptible to social desirability bias than in face-to-face interviews (Groves et al., 2009).

Implications. The MCSDS score does not predict misreporting in a socially desirable direction. I discuss two possible explanations: (1) HSM respondents ascertained the scale's intent and strategically responded to avoid the self-monitor label or (2) the shortened scale does not work as intended. Psychometric scales are increasingly commonplace as modern college students frequently encounter them in other studies, on social media, and with their physicians. As such, they are increasingly more likely to infer the MCSDS's intent than for students involved in validation testing at its introduction in 1960. Initial and subsequent validation studies of the MCSDS's validity as a predictor of HSM behavior may have used students less familiar with personality testing than modern college students. I hypothesize that respondents quickly note the presence of the MCSDS because it abruptly changes the survey's flow, and questions possess a different written style, subject matter, and response options. Once noticed, HSM respondents can ascertain the scale's intent and alter their response behavior to evade the HSM label. High GPA respondents with increased cognitive ability are more likely to detect the scale and understand its intent. Consistent with this hypothesis, high GPA students have the lowest MCSDS scores. An alternative explanation is that the scale does not work as intended and high achieving, high GPA students are less likely to self-monitor because they portray the characteristics low GPA students emulate.

Discussion, Implications, and Future Research

Findings from this study help future researchers navigate the impact of social desirability bias on self-reported data and evaluate the appropriateness of using survey data to generalize population behaviors. Findings also help researchers evaluate the validity of past research conclusions that rely upon self-reported behaviors. I next discuss potential research that may build upon this study's findings. I first present the limitations and advantages of this study and then continue with proposals for future research that embrace this study's advantages and address its limitations.

Limitations and Advantages

This validation study extended an invitation to complete an online survey about CRF use, co-curricular participation, and academic success to 6,000 randomly selected undergraduate students. Self-reported behaviors collected by the survey were compared to institution reported data values that use transactional data collected to record CRF-based and grade-based behaviors. Self-reported and institution-reported data were compared to assess the accuracy of self-reported data. This research was conducted within the context of the Model of Response Process (Tourangeau, 1984, 1987) that explains how respondents formulate and report responses to surveys. I begin with a discussion of the study's limitations and then continue to its advantages.

Limitations. This study's most significant limitation stems from being a single institution study at a selective, public, and research-oriented university composed primarily of traditional college students who are over represented in STEM disciplines. Validation data is solely based upon academic performance and behaviors that occur within the CRF or in CRF-based activities. This university has a rarefied grade repeat policy where students can eliminate the effect of selected low grades from their cumulative GPA but not remove them from their academic record. Further study would benefit from expanding the scope of participants and participant behavior topics. Because all students are required to take two PE/HES courses, the effect of self-selection is minimized. However, using behaviors outside the CRF strengthens the findings because including behavioral data from other areas yields different results. For example, class skipping behaviors could differ for other disciplines that offer a more traditional classroom experience. This policy creates complexity because students earning low grades may be confused by how to describe academic performance within the policy. CRF use data is used to create an institution reported data set on co-curricular participation, class skipping, and overall CRF use.

Results come with limited generalizability to other settings because NCSU is a selective institution with students who have an unusually high concentration of STEM oriented majors and high rates of participation in the traditional residential college experience. The study of social and academic behavior is applicable to all types of students that are not typically enrolled at NCSU in large numbers. NCSU students have significantly higher than average SAT scores and high school GPAs. Most analysis in this study, especially analysis of ability as measured by SAT scores, are conducted only on high ability students. As such, conclusions about the effect of SAT scores on the likelihood of responding are limited to students on the above average end of the SAT score range. Further study should expand this methodology to other institutions with more diverse SAT scores.

Finally, the substantive influence of the spring university holiday on self-reported behavioral accuracy in question 1A was not anticipated prior to data collection and, now knowing this, I would have collected data differently. All members of the sample received an invitation to complete the survey on March 14, the first day after the university holiday concluded. Just over half of the responses came in after the initial request, and these respondents grossly over reported the prior seven days of CRF use. For those not responding to the first request, follow up invitations were emailed March 21 and March 23. Just under half of the responses were received following these prompts, and those respondents also inflated CRF use but to a much lesser extent. Because the first wave of respondents immediately complied with the survey request and subsequent respondents required further prompting, the later respondents may be fundamentally different in ways that affect reporting accuracy. Future study should randomly assign students into control and experimental groups where the impact of the university holiday on reporting accuracy can be conducted without the prospect of selection bias.

Advantages. This study assesses the validity of self-reporting at the unit level, tracking response behavior from a single respondent over 14 questions that span a variety of topic areas. This approach allowed me to analyze response accuracy by student characteristics and over many question formats. Prior validation studies with unit-level validation data are limited to a single topic area, and studies that span many topic areas rely on aggregated validation datasets like population averages and mean comparisons of experimental and control groups. Also unlike prior research, this study includes a 6,000-person sample with a response rate that exceeds 25%; prior research usually draws conclusions from dozens of study participants or lower response rates. This may be the most comprehensive validation study in all of higher education research.

This study's participants received no clues of its true purpose and remain unaware of their participation in a validation study. Krueter, et al., (2008) found that response behavior differs when respondents believe the survey administrator has access to true values, making it vital to hide the study's true intent from participants. As such, participants did not alter their response behavior because they knew responses would be judged for accuracy. Instead, respondents were led to believe this is a typical study of higher education that connects physical activity to academic performance. Most of the existing validation studies provide contextual clues or outright statements that the purpose of the study is to evaluate response accuracy. This knowledge causes participants to monitor their behavior with unusual vigor because the upcoming study provides motivation to give their response atypical consideration, thus yielding atypically precise response quality. Validation studies that successfully hide their intent rely on costly data collection techniques that limit the number of respondents. For example, Klesges, et al. (1990) gathered validation data from research assistants secretly spying on participants. A relative advantage of this study is the self-reported data collection for large numbers of students.

Implications for Online College Student Survey Research

Self-reported behavioral data from online student surveys cannot be presumed valid, and the onus for its validity in research should shift from the presumption of validity toward requiring justification for its use as a valid measure of behavior. Evidence shows that respondents desire to provide an accurate description of their behavior to benefit the researcher in their studies, but the typical respondents will give no more than a cursory effort to complete a survey. High GPA respondents more adeptly move through the components of the Model of Response Process (Tourangeau, 1984, 1987) and provide a reasonably accurate and less biased estimate of behavior. Low GPA students are less able to do so. Respondents who are unable to provide low effort descriptions of their behavior are likely to fall back on satisificing response behavior that yields overly idealized, frequently biased descriptions of behavior. Despite this finding, there is no generalizable conclusion to adequately explain how or when respondents will provide accurate responses to surveys or how to obtain unbiased, accurate self-reported behaviors.

This study establishes a significant correlation between a student's GPA and his or her likelihood to respond to the survey request as well as a correlation between a students' GPA and their likelihood to visit the CRF or engage in co-curricular activities. This simultaneous correlation represents a serious threat to the validity of college student self-reported data and warrants further investigation in other settings and for other behaviors. At the same institution, Adams and Umbach (2012) also detected nonresponse error where an academically successful student is more likely to complete an online student evaluation of teaching survey. The positive correlation between GPA and participation rate is also established because high GPA students tend to be more engaged in campus activities. This is partially a preference and partially a consequence of low GPAs because some co-curricular activities require a minimum GPA to participate. Because of the relationship between GPA and survey response rate and GPA and participation, researchers should presume that nonresponse error is present in all self-reports about co-curricular participation.

The primary implication is that survey responses cannot be presumed adequate substitutes for institution reported data or as a self-evident substitute for population behaviors. In some cases, the mismatch between self-reported and institution-reported data is incidental and of little consequence to research conclusions. In other cases, differences are detrimental to the research and may invalidate the conclusions. Unfortunately, there is not a prima fascia test to assess the validity of any given survey question.

Future Research

This study added to the existing body of validation studies by providing a more comprehensive study than has been previously attempted for college students. Unit-level response accuracy is tracked for the same student across a variety of behaviors allowing the analysis of response accuracy for individuals and topics. Results justify and act as a starting point toward more widespread comparisons of self-reported college student behaviors to institution reported behaviors. The findings will improve survey design. However, future research should compare varying response option behaviors because physical activity is extremely biased and the results may not generalize.

Researchers should use the recent expansion of large, transactional data sets that capture other behaviors to enable future study. In today's higher education, students leave a broad electronic footprint that allows the study of more behaviors. Possibilities include replicating the study design with transactional data describing tutorial center visits, NCAA game attendance, dining habits, and any other campus interactions. Students' expansion of their electronic footprints is not limited to institutional data, so I encourage validation study using Fitbit bio data, financial tracking software that categorizes expenditures, video games systems that record active game play time, and geospatial cellular phone location tracking.

A natural next step is a comparison of regression analyses on how physical activity is related to cumulative GPAs. The purpose of the analysis is not to assess the impact of CRF use on academic success but rather to assess how using self-reported data would impact research conclusions on such a study. The first regression model uses self-reported data to estimate the effect of CRF use and co-curricular participation on GPA. A subsequent model uses an identical functional form but substitutes institution reported values for self-reported values to estimate the impact of response bias on the analysis. A final model maintains the function form while using institution reported data for the sample rather than respondents and estimates the effect of nonresponse error on the analysis. A comparison of parameter estimates would show the impact of response bias and nonresponse bias on research conclusions. Parameter estimate differences may be inconsequential; for example, systematic inflation of CRF use may cause a level shift that maintains the integrity of the findings while distorting the relationship's magnitude. It is also possible that the parameter estimates differ in ways that reveal entirely different conclusions, and any researcher using self-reported data must discuss implications that are potentially detrimental to the validity of research conclusions based upon self-reported data. For example, bias could alter the significance of parameter estimates because nonresponse bias omits CRF nonusers. The purpose of this regression analysis is not to establish a relationship between independent variables and academic performance, but rather to investigate how biases that are associated with self-reported data alter research conclusions.

Relative to other institutions of higher education and the general population, the college student population at NC State is homogeneous. Students tend to be traditionally aged, high performing in high school, and STEM focused. Almost all bachelor's level students are living the residential college experience; because the complexity of the collegiate career is found to impact response accuracy, it is necessary to replicate this study with different populations who experience college differently. These results may not generalize to institutions whose curricula are designed for nontraditional students, and it would be unfortunate for these institutions to alter their needs assessment surveys if my findings do not apply.

Future studies should also consider study on non-college student populations who are even more heterogeneous in their ability use technology like computers and smart phones. Study populations that vary by age and ability are likely to include individuals with low ability to use smart phones. This low ability level allows the analysis of two factors that could affect response accuracy. The first factor is a focus on the comparison of similar respondents' online surveys answered on a laptop/desktop computer versus those using smart phones because the effect of the device may alter survey response behaviors in ways not be fully understand. The second factor explores the impact of forcing low ability technology users to provide responses on a smart phone with which they may have little comfort. Because studies concentrate on 18 to 22-year-old student populations, future studies should compare differences in behavior among more diverse respondents using smart phones and those using traditional modes of survey response including desktop/laptop computers.

Compared to recent survey requests from faculty, students, and staff at NC State, this survey generated an atypically high response rate. The high response rate may be attributed to the survey request that described survey topics in the email subject. Specifically, respondents were provided the opportunity to discuss CRF use and co-curricular participation and how it leads to academic success. This marketing strategy increased the response rate but may have unintentionally fostered nonresponse error if the physically active students who would have otherwise not responded were motivated to do so by the survey's description. Future research should replicate this process with more neutral survey recruitment materials.

This study finds that the MCSDS score is not predictive of response accuracy or likelihood to misreport in a socially desirable direction. This finding is potentially situational, but further study should be conducted to determine the test's validity among modern college students. To determine if the entire scale provides better results, I recommend that future studies employ the full questionnaire rather than the subset of questions that I used. To determine if details of this study are not appropriate for the MCSDS, I recommend that future research take a similar format but for other behaviors with in person data collection. If these changes continue to find the MCSDS is not predictive of self-monitoring behaviors, then a new wave of study that assesses its application to modern college students is recommended.

Finally, I did not anticipate the magnitude of reporting behavior differences for respondents of question 1A when a university holiday is included in the reference period. The presence of the spring university holiday in the reference period led to grossly exaggerate self-reports of CRF use. The research design introduced a non-negligible chance that the response behavior of individuals completing the survey immediately after the first request differs from those completing the survey after the second or third request. As such, future study should replicate the comparison of survey response accuracy with similar reference periods but randomly send survey requests. Some should sample members the day after the university holiday and others a week after. Random assignment will help understand how students tend to inflate responses when they find the question confusing.

Conclusions

This validation study is intended to research the accuracy of self-reported, college student behavioral data collected by online surveys. Validation data is derived from institution reported CRF transactional data, class schedules, and academic performance. A randomly selected sample of NC State students was solicited to complete a survey that describes behaviors that are compared to the validation data for the purpose of evaluating the suitability of using their self-reports as an estimate for actual values across the pool of respondents and the target population. Student level and question topic components are analyzed to determine what respondent characteristics and topic areas are subject to response and/or nonresponse biases. If successful, this study will describe the validity of self-reported behavior across a wide swath of common college student behaviors.

Response bias is prevalent in 11 of the 13 questions asked and evaluated. Response bias is most prevalent for socially desirable behaviors like CRF use and GPAs. Systematic underreporting is also prevalent in easily forgotten behaviors like Outdoor Adventure participation and formal fitness program participation. Survey estimates for high GPA students are more likely to be accurate and unbiased, and low GPA students are more likely to offer inaccurate responses that are inflated in a socially desirable direction. However, no student subgroup systematically offered accurate answers, and the evidence suggests that researchers should presume the presence of measurement error rather than presume its absence. Simple questions that are unambiguous and offer a small number of response options generate the most accurate responses. The implications of response bias on research conclusions is not yet explored, but it should concern the community of higher education researchers.

Nonresponse error is prevalent in 11 of 13 questions asked and evaluated. High GPA and highly involved students are more likely to respond to a survey request. The behavior of these respondents is significantly different than that of nonrespondents and, in cases where the nonresponse error is correlated with the survey topic, the survey mean cannot be generalized to the target population. The presence of nonresponse error is significant despite a relatively high response rate that exceeds 25%. The topics of CRF use and cumulative GPA are most susceptible to response bias and nonresponse bias, making the self-reported values for these topics unsuitable for describing the target population. For other topic areas, nonresponse error is less prevalent. Evidence suggests that researchers should presume the presence of nonresponse error rather than presume its absence.

This research met its objective and has sufficiently demonstrated that measurement error and nonresponse error are present in the self-reported data collected for this study. For nearly all questions, the self-reported mean is significantly different than the institution reported mean, and the behavior of respondents is significantly different than nonrespondents. For the majority of questions, the magnitude of the differences exceeds

20%. Without using validation data, there is no rule of thumb to help survey researchers locate systematic bias, though higher GPA students and students with less complex academic histories are more likely to report accurately and without bias. I conclude that self-reported data cannot be accepted as self-evident and that the onus of responsibility should shift from the survey researcher presuming the self-reported data is valid to the survey researcher presuming self-reported data is invalid until proven otherwise.

REFERENCES

Abelson, R. P., Loftus, E. F. & Greenwald, A. G. (1992). Attempts to improve the accuracy of self-reports of voting. In *Questions About Questions*, ed. J. M. Tanur, p. 138-153. New York: Russesll Sage.

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Herbert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology, 161 (4),* 389-398.

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. Research in Higher Education, 53(5), 576-591

The American Association for Public Opinion Research (AAPOR) (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition.* AAPOR.

Artinger, L., Clapham, L., Hunt, C., Meigs, C., Milford, N., Sampson, B., & Forrester, S. A. (2006). The social benefits of intramural sports. *NASPA Journal, 43(1)*, 69-86.

Astin, A. (1993). *What Matters in College? Four Critical Years Revisited.* San Francisco: Jossey-Bass.

Astin, A., & Antonio, A. L. (2012). *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education.* Lanham, MD: Roman and Littlefield Publishers, Inc

Aviv, A. L., Zelenski, J. M., Rallo, L., & Larsen, R. J. (2002). Who comes when: Personality differences in early and later participation in a university subject pool. *Personality and Individual Differences, 33,* 487-496.

Ballard, R. (1992). Short forms of the Marlowe-Crowne Social Desirability Scale. *Psychological Reports, 71*(4), 1155-1160.

Barge, S. & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53(2)*, 182-200.

Belch, H. A., Gebel, M., and Maas, G. M. (2001). Relationship between student recreation complex use, academic performance, and persistence of first-time freshmen. *NASPA Journal, 38 (2),* 14-22.

Berry, D. A., & Hochberge, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference, 82(1)*, 215-227.

Bowman, N. A. (2010). Can 1st-year college students accurately report their learning and development? *American Education Research Journal, 47*(2), 466-496.

Bowman, N. A. (2011). Examining systematic errors in predictors of college student self-reported gains. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data.* New Directions for Institutional Research. San Francisco: Jossey Bass.

Bowman, N. A. (2014). The meaning and interpretation of college student self-reporting gains. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data. New Directions for Institutional Research*. San Francisco: Jossey Bass.

Bowman, N. A., and Herzog, S. (2011). Reconciling (seemingly) discrepant findings: Implications for practice and future research. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data. New Directions for Institutional Research*. San Francisco: Jossey Bass.

Bowman, N. A., and Hill, P. L. (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data. New Directions for Institutional Research*. San Francisco: Jossey Bass.

Bowman, N. A., Schuldt, J. P. (2014). Effects of item order and response options in college student surveys. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data. New Directions for Institutional Research*. San Francisco: Jossey Bass.

Brenner, P. S., & DeLamater, J. D. (2014). Social desirability bias in self-reports of physical activity: Is an exercise identity the culprit? *Social Indicators Research (117)*, 489-504.

Bryant, J. A., & Bradley, J. L. (1993). Enhancing academic productivity, student development, and employee potential. *Recreational Sports Journal, 42-44.*

Chase, D. R., & Godbey, C. G. (1983). The accuracy of self-reported participation rates. *Leisure Studies, 2(2)*, 231-235.

Chen, P. D. (2011). Finding quality responses: The problem of low-quality survey responses and its impact on accountability measures. *Research in Higher Education, 52*, 659-674

Clarkberg, M., Robertson, D., & Einarson, M. (2008, May 26). *Engagement and student surveys: Nonresponse and implications for reporting survey data*. Paper presented at the Annual Forum of the Association for Institutional Research, Seattle, WA. Ithaca, NY: Office of Institutional Research and Planning at Cornell University.

Cohen, J. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cole, J. S., and Gonyea, R. M. (2010). Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education, 51*(4), 305-319.

College Board, 2009. ACT and SAT concordance tables. Retrieved from http://research.collegeboard.org/publications/content/2012/05/act-and-sat-concordance-tables

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, *19(2),* 146-162.

Crowne, D. P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354.

Dey, E. L. (1997). Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education, 44,* 1-19.

Dillman, D. A., Tortora, R. D., and Bowker, D. (1998). *Principles for constructing web surveys.* Retrieved June 30, 2015, from Don A. Dillman Available Papers: http://www.sesrc.wsu.edu/dillman/papersweb/1998.html.

Dobbins, G. H., Farh, J. and Werbel, J. D. (1993). The influence of self-monitoring on inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology*, *23*, 321-334.

Dougherty, K. J., Natow, R. S., Bork, R. H., Jones, M. J., Vega, B. E. (2013). Accounting for higher education accountability: Political origins of state performance funding for higher education. *Teachers College Record, 115,* 1-50.

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrica, 11,* 1-42.

Durante, R., & Ainsworth, B. E. (1996). The recall of physical activity: Using a cognitive model of the question-answering process. *Medicine and Science in Sports and Exercise, 28(10)*, 1282-1291.

Eaton, J. S. (2012). An overview of U.S. accreditation. Viewed on September 6, 2015. http://files.eric.ed.gov/fulltext/ED544355.pdf

Ellingson, J. E., Sackett, P. R., and Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155-166.

Ewell, P. T., and Jones, D. P. (1993). Actions matter: The case for indirect measures in assessing higher education’s progress on the national educational goals. *Journal of General Education*, *42*(2), 123-148.

Garry, M., Sharman, S. J., Felman, J., Marlatt, G. A., and Loftus, E. F. (2002). Examining memory for heterosexual college students’ sexual experiences using an electronic mail diary. *Health Psychology, 21*(6), 629-634.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. In P. D. Umbach (ed.), *Survey Research: Emerging Issues*. New Directions for Institutional Research, no. 127. San Francisco: Jossey-Bass, 2005.

Gonyea, R. M., and Miller, A. (2011). Clearing the AIR about the use of self-reported gains in institutional research. In S. Herzog and N. A Bowman (Eds.), *New Directions for Institutional Research: No 150. Validity and limitations of college student self-reported data* (pp. 99-111). San Francisco, CA: Jossey-Bass.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*, 646-675.

Groves, R. M., Fowler, Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72(2),* 167-189.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly, 64(3),* 299-308.

Haines, D. J., (2001). Undergraduate student benefits from university recreation. *Recreational Sports Journal, 32(1),* 52-61.

Hall, D. (2006). Participation in a campus recreation program and its effect on retention. *Recreational Sports Journal, 30*, 40-45.

Herzog, S. (2011). Gauging academic growth of bachelor degree recipients: Longitudinal vs. self-reported gains in general education. In S. Herzog and N. A. Bowman (Eds), *The* *validity of college student self-reported data.* New Directions for Institutional Research. San Francisco: Jossey Bass.

Huesman, R. L., Brown, A. K., Lee, G., Kellogg, J.P., & Radcliffe, P. M. (2007). Modeling student academic success: Does use of campus recreation facilities make a difference? Presented at the 2007 National Symposium on Student Retention, Milwaukee, WI, September 24-26, 2007. The Symposium was sponsored by the Consortium for Student Data Exchange (CSRDE).

Inkelas, K. K., & Soldner, M. (2011). Undergraduate living-learning programs and student outcomes. *Higher Education: Handbook of Theory, 26*. Springer Science and Business Media, CT.

Johnson, T. P., & Wislar, J. S., (2012). Response and nonresponse errors in surveys. *Journal of the American Medical Association, 307(17),* 1085-1086.

Klesges, R. C., Eck, L. H., Mellon, M. W., Fulliton, W., Somes, G. W., & Hanson, C. L. (1990). The accuracy of self-reports of physical activity. *Medicine and Science in Sports and Exercise, 22(5),* 690-697.

Klesges, L. M., Baranowski, T., Beech, B., Cullen, K., Murray, D. M., Rochon, J., & Pratt, C. (2004). Social desirability bias in self-reported dietary, physical activity, and weight concern measures in 8-to-10 year old African-American girls: Results from the Girls Health Enrichment Multisite Studies (GEMS). *Preventative Medicine 38*, S78-S87.

Koss, M. P., Wilgus, J. K., & Williamsen, K. M. (2014). Campus sexual misconduct: Restorative justice approaches to enhance compliance with Title IX guidance. *Trauma, Violence, Abuse, 15(July)*, 242-257.

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys. *Public Opinion Quarterly*, *72*(5), 847-865.

Kronsnick, J. A., and Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 52*(2), 201-219

Kronsnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50,* 537-567.

Kuh, G. D. (2003). *The National Survey of Student Engagement: Conceptual Framework and Overview of Psychometric Properties*, Indiana University Center for Postsecondary Research, Bloomington, IN.

Kuh, G. D. (2004). The national survey of student engagement: Conceptual framework and overview of psychometric properties .Available at: http://nsse.iub.edu/pdf/conceptual_framework_2003.pdf (Accessed January 3, 2015).

Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., and Kennedy, J. (2001). *NSSE Technical and Norms Report,* Indiana University Center of Postsecondary Research and Planning, Bloomington, IN.

Kuncel, N. R., Crede, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Education Research, 75(1)*, 63-82.

Leite, W. L., and Beretvas, S. N. (2003). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement, 65*, 140-154.

Lindsey, R., & Sessoms, E. (2006 May). Assessment of a campus recreation program on student recruitment, retention, and frequency of participation across certain demographic variables. *Recreational Sports Journal, 30(1)*, 30-39.

Mahoney, S. A. (2011). *Students Involvement in Campus Recreational Sports and Outcomes.* Bloomington, IN: Indiana University (dissertation).

Mayer, R. E., Stull, A. T., Campbell, J., Almeroth, K., Bimber, B., Chun, D., & Knight, A. (2007). Overestimation bias in self-reported SAT scores. *Educational Psychology Review, 19,* 443-454.

Miller, K. H., Noland, M., Rayens, M. K., & Staten, R., (2008). Characteristics of users and nonusesrs of a campus recreation center. *Recreation Sports Journal, (32)*, 87-96.

Moinester, M., and Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. *The Quantitative Methods for Psychology, 10*(2), 124-130.

Motl, R. W., McAuley, E., and DiStefano, C. (2004). Is social desirability associated with self-reported physical activity? *Preventative Medicine 40(6)*, 735-739.

Mueller-Hanson, R. A., Heggestad, E. D., and Thorton, G. C. III. (2003). Individual processes differences in impression management: An exploration of the psychological processes underlying faking. *Journal of Applied Psychology, 88*, 348-355.

OIRP (2013). http://oirp.ncsu.edu/sites/oirp.ncsu.edu/files/soph13.exec_.summary.pdf (accessed on July 22, 2015).

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

Olson, K., & Kennedy, C. (2006). Examination of the relationship between nonresponse and measurement error in a validation study of alumni. *Sociology Department, Faculty Publications.* Paper 146.

Pike, G. R. (1995). The relationship between self reports of college experiences and achievement scores. *Research in Higher Education*, *36,* 1-21.

Pike, G. R. (1999a). The effects of residential learning communities and traditional residential living arrangements on educational gains during the first year of college. *Journal of College Student Development, 40*(3), 269-284.

Pike, G. R. (1999b). The constant error of the halo in educational outcomes research. *Research in Higher Education, 40*(1), 61-86.

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data.* New Directions for Institutional Research. San Francisco: Jossey Bass.

Pike, G. R., Schroeder, C. C., & Berry, T. R. (1997). Enhancing the education impact of residence halls: The relationship between residential learning communities and first year college experiences and persistence. *Journal of College Student Development, 38*(6), 609-21.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education, 35(1)*, 45-76.

Porter, S. R., Rumann, C., and Pontius, J. (2011). The validity of student engagement survey questions: Can we accurately measure academic challenge? In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data.* New Directions for Institutional Research. San Francisco: Jossey Bass.

Porter, S. R., & Umbach, P. D., (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47(2)*, 229-247.

Porter, S. R., & Whitcomb, M. E., (2005). Non-response in student surveys: The role of demographics, engagement, and personality. *Research in Higher Education, 46(2),* 127-152.

Rothwell, E., and Theodore, P. (2006). Intramurals and college student development: The role of intramurals on values clarification. *Recreational Sports Journal, 30(1),* 46-52

Sallis, J. F., & Saelens, B. E. (2000). Assessment of physical activity by self-reports: Status, limitations, and future directions. *Research Quarterly for Exercise and Sport, 71*(2 Supplement), 1-14.

Sarrah, S. and Tukibayeva, M. (2014). Survey page length and progress indicators: What are their relationships to nonresponse? In N. A. Bowman and S. Herzog (eds.), *Special Issue: Methodological Advances and Issues in Studying College Impact.*. New Directions for Institutional Research, no. 161. San Francisco: Jossey-Bass, 2014.

Sax, L. J., Gilmartin, S. K., and Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44(4)*, 409-432.

Sax, L. J., Gilmartin, S. K., Lee, J. J., & Hagedorn, L. S. (2008). Using web surveys to reach community college students: An analysis of response rates and response bias. *Community College Journal of Research and Practice, 32*, 712-729.

Schmitt, N., and Kuncel, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55*, 569-588.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., and Yoo, T.-Y. (2003). Impact of elaboration on socially desirable responding and the validity of bio-data measures. *Journal of Applied Psychology, 88*, 979-988.

Seifert, T. A., and Asel, A. M. (2011). The tie that binds: The role of self-reported high school gains in self-reported college gains. In S. Herzog and N. A. Bowman (Eds), *The validity of college student self-reported data.* New Directions for Institutional Research. San Francisco: Jossey Bass

Shepherd, R. J. (2003). Limits to the measurement of habitual physical activity by questionnaires. *British Journal of Sports Medicine, 37(3)*, 197-206.

Snyder, M. (1987). *Public appearances, private realities: The psychology of self monitoring.* New York: W. H. Freeman/Times Books/Henry Holt and Co.

Standish, T., Young, K., Joines, J. and Gallagher, V. (Manuscript submitted for publication). A practitioner's guide to improve student evaluation of instruction responses.

Stassen, M. L. A. (2003). Student outcomes: The impact of varying living-learning community models. *Research in Higher Education, 44*(5), 581-613.

Stryker, S. (1980). Symbolic interactionism: A social structural version. Caldwell, NJ: The Blackburn Press.

Sudman, S. and Bradburn, N. (1982). *Asking Questions: A Practical Guide to Questionnaire Design*, San Francisco: Jossey-Bass.

Supiano, B. (2008). Student aid is up, but the rise in college costs outpaces family incomes. *Chronicle of Higher Education, 55*(11), A23.

Todd, M., Czyszcson, G., Wallace Carr, J., and Pratt, C. (2009). Comparisons of health and academic incides between campus recreation facility users and nonusers. *Recreational Sports Journal,* 33(1).

Torres, V. (2011). Using student development theories to explain student outcomes. *Higher Education: Handbook of Theory, 26*. P 425-448.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, and R Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (p. 73-100). Washington DC: National Academy Press

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In H. Hippler, N. Schwartz, and S. Sudman (Eds.), *Social information processing and survey methodology* (p. 149-162). New York: Spring-Verlag.

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013). *The Science of Web Surveys.* New York: Oxford University Press.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Tourangeau, R., Smith, T. W., and Rasinski, K. (1997). Motivation to report sensitive behaviors in surveys: Evidence from a bogus pipeline experiment. *Journal of Applied Social Psychology, 27*(3), 209-222.

Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-93.

Umbach, P. D. (2005). Getting back to the basics of survey research. In P. D. Umbach (ed.), *Survey Research: Emerging Issues*. New Directions for Institutional Research, no. 127. San Francisco: Jossey-Bass, 2005.

Umbach, P. D., and Kuh, G. D. (2006). Student experiences with diversity at liberal arts colleges: Anther claim for distinctiveness. *The Journal of Higher Education, 77*(1), 169-192.

Upcraft, M. L., and Schuh, J. H. (1996). *Assessment in Student Affairs*. San Francisco: Jossey-Bass.

Villar, A., Collegaro, M., and Yang, Y. (2013). Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. *Social Science Computer Review, 31*(6), 744-762.

Warnecke, R. B., Johnson, T. P., Chavez, N., Sudman, S., O'Rourke, D. P., Lacey, L., et al. (1997). Improving the question wording in surveys of culturally diverse populations. *Annals of Epidemiology*, 334-342.

Young, S. J., Ross, C. M., and Barcelona, R. J. (2003). Perceived constraints by college students to participation in campus recreational sports programs. *Recreational Sports Journal*, *27(2)*, 47-62.

Zimmerman, M. A., Caldwell, C. H., and Bernat, D. H. (2002). Discrepancy between self-report and school-record grade point average: Correlates with psychosocial outcomes among African Americans adolescents. *Journal of Applied Social Psychology, 32*, 86-109.

Zizzi, S., Ayers, S. F., Watson II, J. C., and Keller, L. A. (2004, Summer). Assessing the impact of new campus recreation centers. *NASPA Journal, 4*, 588-630.

APPENDICES

Appendix A

Question 1: CRF Use

Table A.1.

Summary of Question 1A Results.

		Actual Visits																	
Respondent Category	N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Respondents	1505	1100(73.1%)	135(9%)	114(7.6%)	55(3.7%)	45(3%)	24(1.6%)	14(0.9%)	4(0.3%)	9(0.6%)	5(0.3%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Non Respondents	4495	3533(78.6%)	401(8.9%)	235(5.2%)	141(3.1%)	91(2%)	42(0.9%)	24(0.5%)	15(0.3%)	8(0.2%)	1(0%)	3(0.1%)	0(0%)	0(0%)	1(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Sample	6000	4633(77.2%)	536(8.9%)	349(5.8%)	196(3.3%)	136(2.3%)	66(1.1%)	38(0.6%)	19(0.3%)	17(0.3%)	6(0.1%)	3(0.1%)	0(0%)	0(0%)	1(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Population	20773	15823(76.2%)	1902(9.2%)	1318(6.3%)	675(3.2%)	431(2.1%)	261(1.3%)	174(0.8%)	92(0.4%)	48(0.2%)	29(0.1%)	8(0%)	2(0%)	4(0%)	5(0%)	0(0%)	0(0%)	0(0%)	1(0%)
		Actual Visits																	
Respondent Category	N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0 Self-Reported Visits	237	236(99.6%)	1(0.4%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
1 Self-Reported Visits	48	36(75%)	9(18.8%)	2(4.2%)	1(2.1%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
2 Self-Reported Visits	42	24(57.1%)	5(11.9%)	8(19%)	4(9.5%)	1(2.4%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
3 Self-Reported Visits	58	24(41.4%)	7(12.1%)	16(27.6%)	6(10.3%)	2(3.4%)	3(5.2%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
4 Self-Reported Visits	31	8(25.8%)	3(9.7%)	8(25.8%)	7(22.6%)	4(12.9%)	1(3.2%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
5 Self-Reported Visits	35	15(42.9%)	4(11.4%)	3(8.6%)	2(5.7%)	4(11.4%)	2(5.7%)	3(8.6%)	2(5.7%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
6 Self-Reported Visits	15	6(40%)	3(20%)	1(6.7%)	2(13.3%)	2(13.3%)	0(0%)	0(0%)	0(0%)	1(6.7%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
7 Self-Reported Visits	9	4(44.4%)	2(22.2%)	0(0%)	1(11.1%)	0(0%)	0(0%)	1(11.1%)	0(0%)	1(11.1%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
8 Self-Reported Visits	3	0(0%)	1(33.3%)	0(0%)	0(0%)	0(0%)	0(0%)	1(33.3%)	0(0%)	0(0%)	1(33.3%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
9 Self-Reported Visits	3	1(33.3%)	1(33.3%)	0(0%)	0(0%)	1(33.3%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
10 Self-Reported Visits	3	2(66.7%)	0(0%)	1(33.3%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
11 Self-Reported Visits	1	1(100%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
12 Self-Reported Visits	0																		
13 Self-Reported Visits	0																		
14 Self-Reported Visits	1	0(0%)	0(0%)	1(100%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
15 Self-Reported Visits	1	1(100%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
16 Self-Reported Visits	0																		
17 Self-Reported Visits	0																		
Item Non Respondents	90	84(93.3%)	2(2.2%)	1(1.1%)	1(1.1%)	1(1.1%)	1(1.1%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Unit Non Respondents	4405	3449(78.3%)	399(9.1%)	234(5.3%)	140(3.2%)	90(2%)	41(0.9%)	24(0.5%)	15(0.3%)	8(0.2%)	1(0%)	3(0.1%)	0(0%)	0(0%)	1(0%)	0(0%)	0(0%)	0(0%)	0(0%)

Table A.2.

Duncan Groupings by Question 1A Response.

Duncan Grouping	Inst Report Mean	N	Question 1A Category
D E F	0.67	6	10+ Self Reported Visits
B C D	1.67	3	9 Self Reported Visits
A	5.33	3	8 Self Reported Visits
B C	1.89	9	7 Self Reported Visits
B C D	1.73	15	6 Self Reported Visits
B	1.97	35	5 Self Reported Visits
B C	1.87	31	4 Self Reported Visits
B C D E	1.38	58	3 Self Reported Visits
D E F	0.81	42	2 Self Reported Visits
E F	0.31	48	1 Self Reported Visits
F	0.00	237	0 Self Reported Visits

Note. Means with different letters are significantly different from means in other letter groups (P<.05).

Table A.3.

Duncan Groupings by Question 1B Response.

Duncan Grouping	Inst Report Mean	N	Question 1B Category
A	2.18	160	Very Often
B	0.97	120	Often
C	0.30	163	Sometimes
D	0.02	72	Never

Note. Means with different letters are significantly different from means in other letter groups (P<.05).

Table A.4.

Summary of Question 1C Results.

		Actual Value				
Response Category	N	Never Visited	Less Than Once per Week	Once Per Week	Twice Per Week	Three or More Times per Week
Respondents	1505	255 (16.9%)	403 (26.8%)	428 (28.4%)	211 (14%)	208 (13.8%)
Non Respondents	4495	1045 (23.2%)	1511 (33.6%)	1179 (26.2%)	414 (9.2%)	346 (7.7%)
Sample total	6000	1300 (21.7%)	1914 (31.9%)	1607 (26.8%)	625 (10.4%)	554 (9.2%)
Population Total	20775	4382 (21.1%)	6557 (31.6%)	5727 (27.6%)	2191 (10.5%)	1918 (9.2%)
		Actual Value				
Response Category	N	Never Visited	Less Than Once per Week	Once Per Week	Twice Per Week	Three or More Times per Week
I never entered	67	60 (89.6%)	6 (9%)	1 (1.5%)	0 (0%)	0 (0%)
Less than once per week	129	26 (20.2%)	78 (60.5%)	22 (17.1%)	2 (1.6%)	1 (0.8%)
Once per week	53	1 (1.9%)	23 (43.4%)	26 (49.1%)	3 (5.7%)	0 (0%)
Twice per week	95	2 (2.1%)	15 (15.8%)	53 (55.8%)	24 (25.3%)	1 (1.1%)
Three or more times per week	159	2 (1.3%)	4 (2.5%)	45 (28.3%)	46 (28.9%)	62 (39%)
Item Nonresponse	90	18 (20%)	40 (44.4%)	15 (16.7%)	10 (11.1%)	7 (7.8%)
Unit Nonresponse	4405	1027 (23.3%)	1471 (33.4%)	1164 (26.4%)	404 (9.2%)	339 (7.7%)

Table A.5.

Duncan Groupings by Question 1C Response.

Duncan Grouping	Inst Report Mean	N	Question 1C Category
A	2.12	159	Three Times Per Week
B	0.99	95	Twice Per Week
C	0.64	53	Once Per Week
D	0.29	129	Less Than Once Per Week
E	0.02	67	Never Entered

Note. Means with different letters are significantly different from means in other letter groups ($P<.05$).

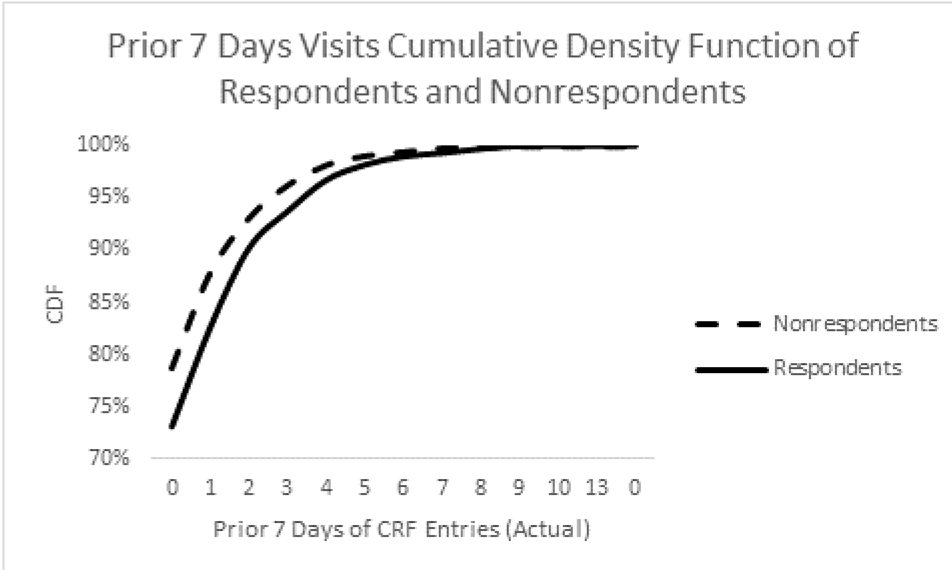

Figure A.1. CDF of question 1A.

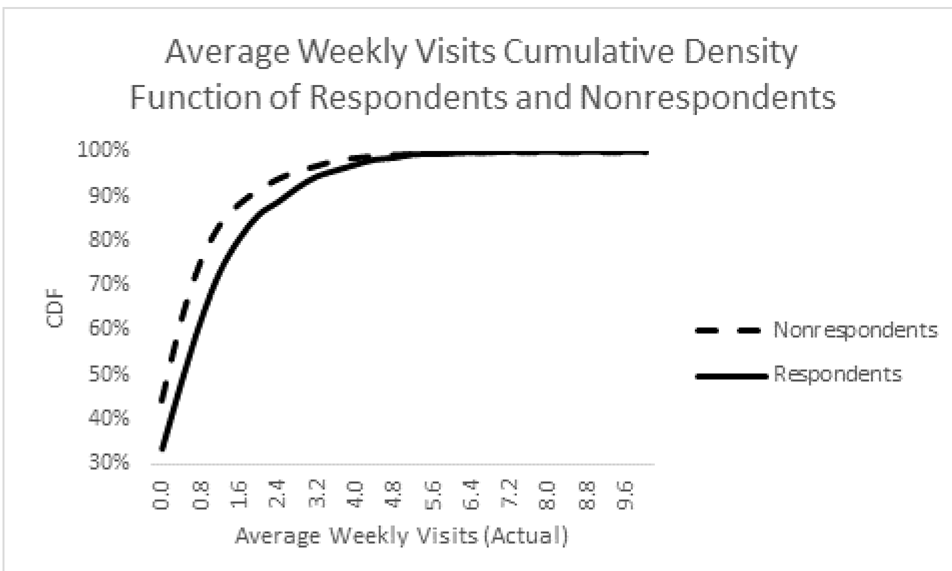

Figure A.2. CDF of question 1C.

Appendix B

Question 2: Cumulative GPA

Table B.1.

Summary of Question 2 Results.

		Actual Values								
Response Category	N	No Cum GPA	0.00 to 0.49	0.50 to 0.99	1.00 to 1.49	1.50 to 1.99	2.00 to 2.49	2.50 to 2.99	3.00 to 3.49	3.50 & Above
Respondents	1335	18 (1.3%)	1 (0.1%)	0 (0%)	4 (0.3%)	18 (1.3%)	76 (5.7%)	231 (17.3%)	454 (34%)	533 (39.9%)
Non Respondents	4665	183 (3.9%)	2 (0%)	0 (0%)	21 (0.5%)	74 (1.6%)	481 (10.3%)	1073 (23%)	1529 (32.8%)	1297 (27.8%)
Sample total	6000	201 (3.4%)	3 (0.1%)	0 (0%)	25 (0.4%)	92 (1.5%)	557 (9.3%)	1304 (21.7%)	1983 (33.1%)	1830 (30.5%)
Population Total	20775	676 (3.3%)	10 (0%)	0 (0%)	60 (0.3%)	298 (1.4%)	1984 (9.5%)	4527 (21.8%)	6859 (33%)	6340 (30.5%)
		Actual Values								
Response Category	N	No Cum GPA	0.00 to 0.49	0.50 to 0.99	1.00 to 1.49	1.50 to 1.99	2.00 to 2.49	2.50 to 2.99	3.00 to 3.49	3.50 & Above
No Cum GPA	0	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)
Cum GPA 0.00 to 0.49	4	2 (50%)	1 (25%)	0 (0%)	(0%)	(0%)	(0%)	1 (25%)	(0%)	(0%)
Cum GPA 0.50 to 0.99	0	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)
Cum GPA 1.00 to 1.49	3	(0%)	(0%)	0 (0%)	3 (100%)	(0%)	(0%)	(0%)	(0%)	(0%)
Cum GPA 1.50 to 1.99	7	(0%)	(0%)	0 (0%)	(0%)	7 (100%)	(0%)	(0%)	(0%)	(0%)
Cum GPA 2.00 to 2.49	61	1 (1.6%)	(0%)	0 (0%)	1 (1.6%)	7 (11.5%)	48 (78.7%)	2 (3.3%)	2 (3.3%)	(0%)
Cum GPA 2.50 to 2.99	205	1 (0.5%)	(0%)	0 (0%)	(0%)	3 (1.5%)	20 (9.8%)	180 (87.8%)	1 (0.5%)	(0%)
Cum GPA 3.00 to 3.49	498	7 (1.4%)	(0%)	0 (0%)	(0%)	1 (0.2%)	7 (1.4%)	46 (9.2%)	423 (84.9%)	14 (2.8%)
Cum GPA 3.50 & Above	557	7 (1.3%)	(0%)	0 (0%)	(0%)	(0%)	1 (0.2%)	2 (0.4%)	28 (5%)	519 (93.2%)

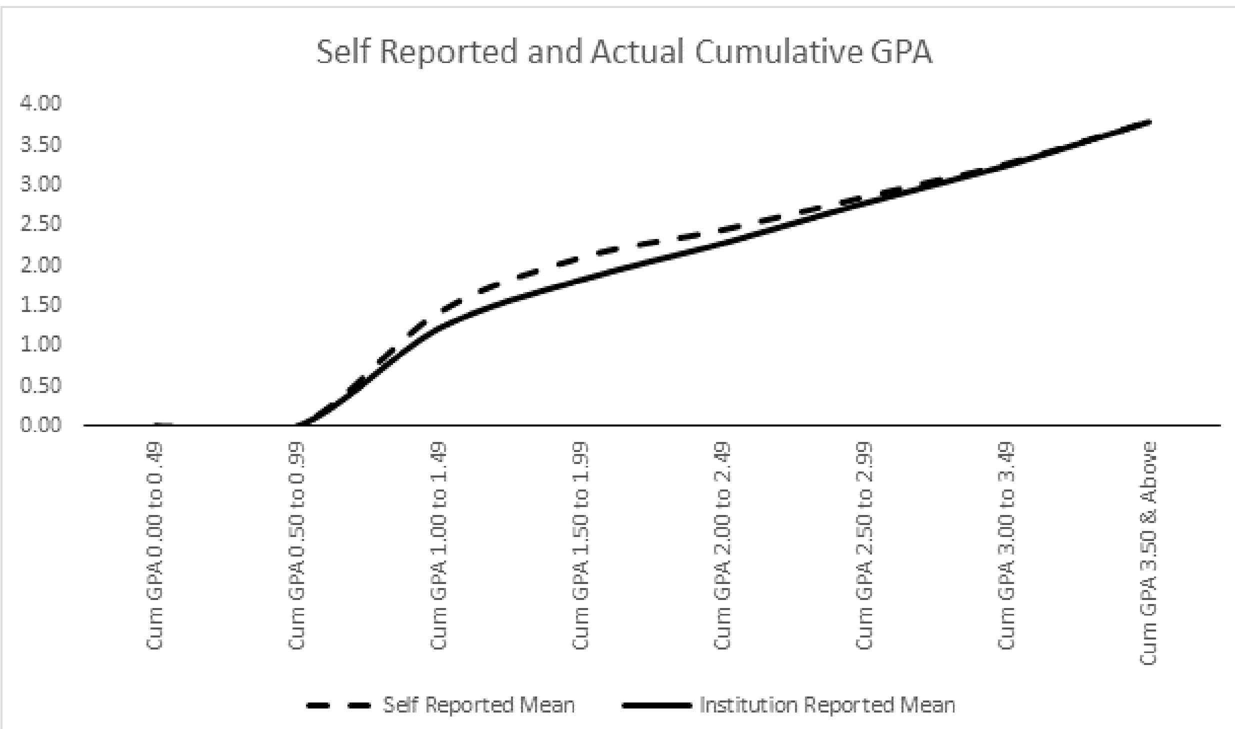

Figure B.1. Self-reported and institution-reported cumulative GPA.

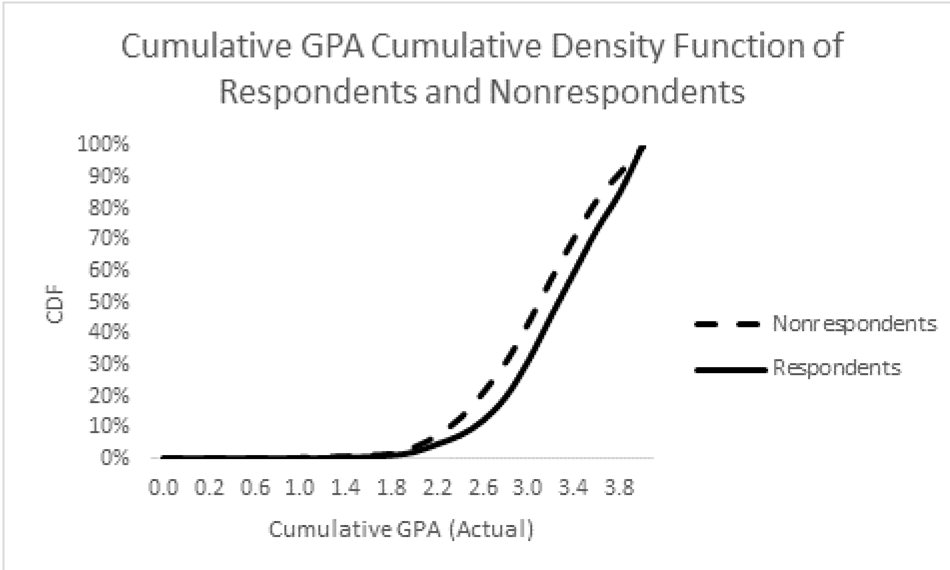

Figure B.2. CDF of question 2.

Appendix C

Question 3: Fall 2015 Semester GPA

Table C.1.

Summary of Question 3 Results.

		Actual Values									
Response Category	N	No Sem GPA	0.00 to 0.49	0.50 to 0.99	1.00 to 1.49	1.50 to 1.99	2.00 to 2.49	2.50 to 2.99	3.00 to 3.49	3.50 to 3.99	4.00 & Above
Respondents	1320	28 (2.1%)	21 (1.6%)	3 (0.2%)	20 (1.5%)	34 (2.6%)	101 (7.7%)	183 (13.9%)	365 (27.7%)	410 (31.1%)	155 (11.7%)
Non Respondents	4680	233 (5%)	117 (2.5%)	31 (0.7%)	114 (2.4%)	212 (4.5%)	494 (10.6%)	837 (17.9%)	1249 (26.7%)	1081 (23.1%)	312 (6.7%)
Sample total	6000	261 (4.4%)	138 (2.3%)	34 (0.6%)	134 (2.2%)	246 (4.1%)	595 (9.9%)	1020 (17%)	1614 (26.9%)	1491 (24.9%)	467 (7.8%)
Population Total	20775	830 (4%)	440 (2.1%)	154 (0.7%)	398 (1.9%)	900 (4.3%)	2040 (9.8%)	3551 (17.1%)	5577 (26.8%)	5148 (24.8%)	1737 (8.4%)
		Actual Values									
Response Category	N	No Sem GPA	0.00 to 0.49	0.50 to 0.99	1.00 to 1.49	1.50 to 1.99	2.00 to 2.49	2.50 to 2.99	3.00 to 3.49	3.50 to 3.99	4.00 & Above
No Sem GPA	0	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)
Sem GPA 0.00 to 0.49	10	7 (70%)	1 (10%)	(0%)	(0%)	(0%)	1 (10%)	(0%)	1 (10%)	(0%)	(0%)
Sem GPA 0.50 to 0.99	0	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)
Sem GPA 1.00 to 1.49	6	(0%)	1 (16.7%)	(0%)	5 (83.3%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
Sem GPA 1.50 to 1.99	18	(0%)	(0%)	2 (11.1%)	6 (33.3%)	10 (55.6%)	(0%)	(0%)	(0%)	(0%)	(0%)
Sem GPA 2.00 to 2.49	57	1 (1.8%)	1 (1.8%)	1 (1.8%)	8 (14%)	8 (14%)	33 (57.9%)	5 (8.8%)	(0%)	(0%)	(0%)
Sem GPA 2.50 to 2.99	178	2 (1.1%)	(0%)	(0%)	1 (0.6%)	10 (5.6%)	47 (26.4%)	104 (58.4%)	12 (6.7%)	2 (1.1%)	(0%)
Sem GPA 3.00 to 3.49	410	10 (2.4%)	6 (1.5%)	(0%)	(0%)	5 (1.2%)	17 (4.1%)	63 (15.4%)	290 (70.7%)	18 (4.4%)	1 (0.2%)
Sem GPA 3.50 to 3.99	458	4 (0.9%)	4 (0.9%)	(0%)	(0%)	1 (0.2%)	3 (0.7%)	11 (2.4%)	59 (12.9%)	374 (81.7%)	2 (0.4%)
Sem GPA 4.00 & Above	183	4 (2.2%)	8 (4.4%)	(0%)	(0%)	(0%)	(0%)	(0%)	3 (1.6%)	16 (8.7%)	152 (83.1%)

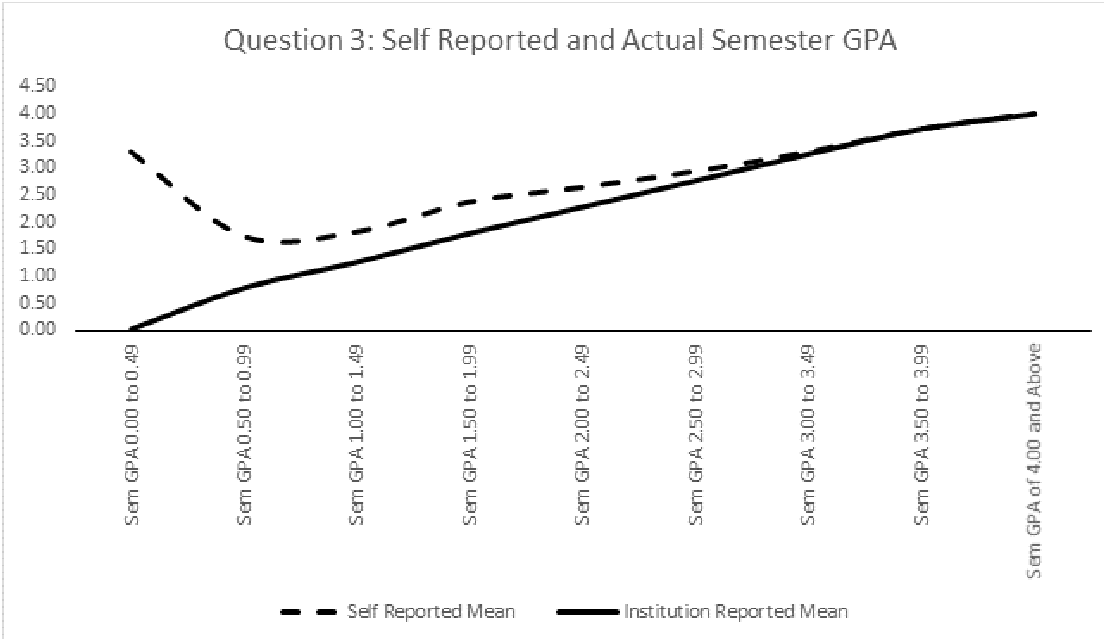

Figure C.1. Self-reported and institution-reported semester GPA.

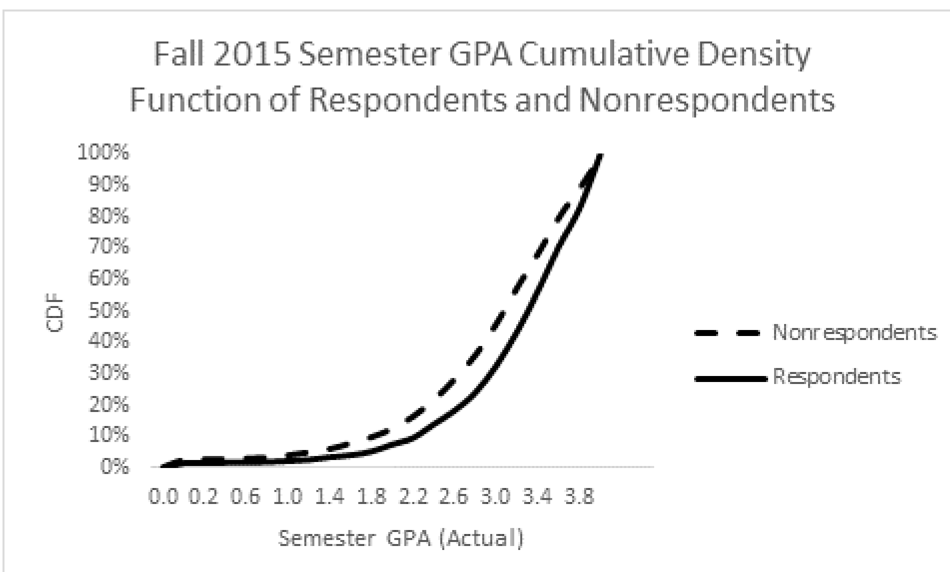

Figure C.2. CDF of question 3.

Appendix D

Question 4: D/F/Unsatisfactory Grades Earned

Table D.1.

Summary of Question 4 Results.

		Actual Ds/Fs/Unsatisfactory Grades Earned					
Respondent Status	N	0	1	2	3	4	5 or More
Respondents	1340	886 (66.1%)	188 (14%)	95 (7.1%)	44 (3.3%)	39 (2.9%)	88 (6.6%)
Non-Respondents	4660	2568 (55.1%)	719 (15.4%)	409 (8.8%)	269 (5.8%)	187 (4%)	508 (10.9%)
Sample	6000	3454 (57.6%)	907 (15.1%)	504 (8.4%)	313 (5.2%)	226 (3.8%)	596 (9.9%)
Population	20775	11793 (56.8%)	3195 (15.4%)	1740 (8.4%)	1119 (5.4%)	737 (3.5%)	2191 (10.5%)
Self Reported Data	N	0	1	2	3	4	5 or More
0 Self-Reported Ds/Fs/Unsatisfactory Grades Earned	928	874 (94.2%)	40 (4.3%)	7 (0.8%)	2 (0.2%)	3 (0.3%)	2 (0.2%)
1 Self-Reported Ds/Fs/Unsatisfactory Grades Earned	198	8 (4%)	137 (69.2%)	31 (15.7%)	8 (4%)	4 (2%)	10 (5.1%)
2 Self-Reported Ds/Fs/Unsatisfactory Grades Earned	115	1 (0.9%)	10 (8.7%)	54 (47%)	21 (18.3%)	12 (10.4%)	17 (14.8%)
3 Self-Reported Ds/Fs/Unsatisfactory Grades Earned	48	1 (2.1%)	0 (0%)	1 (2.1%)	11 (22.9%)	11 (22.9%)	24 (50%)
4 Self-Reported Ds/Fs/Unsatisfactory Grades Earned	25	1 (4%)	0 (0%)	1 (4%)	2 (8%)	8 (32%)	13 (52%)
5 or More Self-Reported Ds/Fs/Unsatisfactory Grades Earned	26	1 (3.8%)	1 (3.8%)	1 (3.8%)	0 (0%)	1 (3.8%)	22 (84.6%)
Item Non Respondents	255	124 (48.6%)	49 (19.2%)	24 (9.4%)	17 (6.7%)	12 (4.7%)	29 (11.4%)
Unit Non Respondents	4405	2444 (55.5%)	670 (15.2%)	385 (8.7%)	252 (5.7%)	175 (4%)	479 (10.9%)

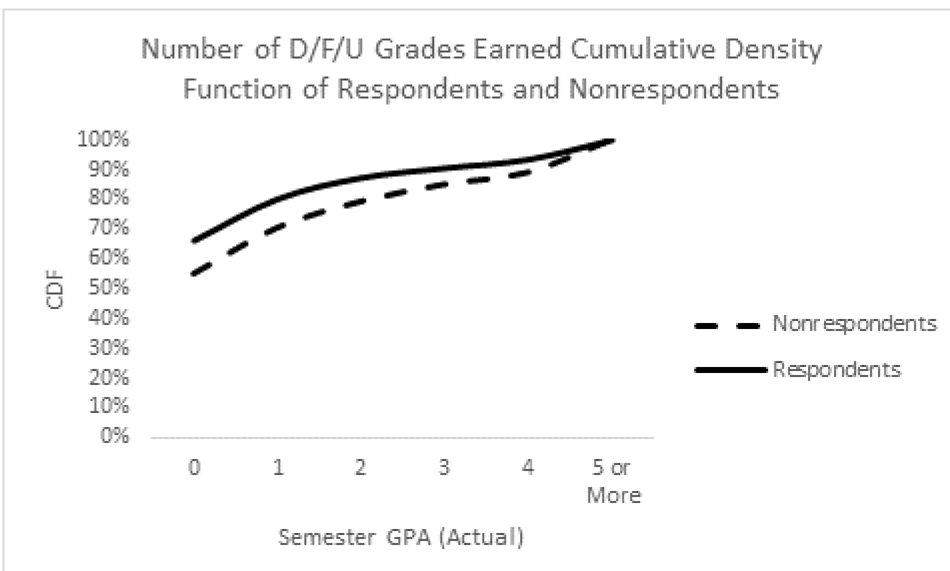

Figure D.1. CDF of question 4.

Appendix E

Question 5: A Grades Earned

Table E.1.

Summary of Question 5 Results.

		Underreporting	Correctly	Over Reporting	Actual A Grades Earned																				
Respondent Status	**N**	**As Earned**	**Reporting As**	**As Earned**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Respondents	1307				50(3.8%)	63(4.8%)	91(7%)	97(7.4%)	116(8.9%)	82(6.3%)	97(7.4%)	53(4.1%)	58(4.4%)	47(3.6%)	57(4.4%)	50(3.8%)	41(3.1%)	42(3.2%)	42(3.2%)	35(2.7%)	30(2.3%)	26(2%)	22(1.7%)	16(1.2%)	192(14.7%)
Non-Respondents	4651	N/A	N/A	N/A	333(7.2%)	289(6.2%)	348(7.5%)	381(8.2%)	396(8.5%)	302(6.5%)	264(5.7%)	225(4.8%)	220(4.7%)	206(4.4%)	161(3.5%)	162(3.5%)	151(3.2%)	124(2.7%)	129(2.8%)	112(2.4%)	119(2.6%)	98(2.1%)	94(2%)	68(1.5%)	469(10.1%)
Sample	5958	N/A	N/A	N/A	383(6.4%)	352(5.9%)	439(7.4%)	478(8%)	512(8.6%)	384(6.4%)	361(6.1%)	278(4.7%)	278(4.7%)	253(4.2%)	218(3.7%)	212(3.6%)	192(3.2%)	166(2.8%)	171(2.9%)	147(2.5%)	149(2.5%)	124(2.1%)	116(1.9%)	84(1.4%)	661(11.1%)
Population	20775	N/A	N/A	N/A	1244(6%)	1169(5.6%)	1583(7.6%)	1657(8%)	1689(8.1%)	1441(6.9%)	1261(6.1%)	980(4.7%)	928(4.5%)	876(4.2%)	800(3.9%)	776(3.7%)	719(3.5%)	596(2.9%)	552(2.7%)	504(2.4%)	491(2.4%)	443(2.1%)	402(1.9%)	312(1.5%)	2352(11.3%)
Self Reported Data	**N**				**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
00 Self-Reported A Grades Earned	32	N/A	62.5%	37.5%	20(62.5%)	3(9.4%)	3(9.4%)	1(3.1%)	0(0%)	1(3.1%)	0(0%)	0(0%)	0(0%)	1(3.1%)	1(3.1%)	0(0%)	1(3.1%)	1(3.1%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
01 Self-Reported A Grades Earned	68	19.1%	27.9%	52.9%	13(19.1%)	19(27.9%)	10(14.7%)	6(8.8%)	4(5.9%)	2(2.9%)	2(2.9%)	3(4.4%)	1(1.5%)	0(0%)	1(1.5%)	1(1.5%)	0(0%)	1(1.5%)	3(4.4%)	0(0%)	1(1.5%)	0(0%)	1(1.5%)	0(0%)	0(0%)
02 Self-Reported A Grades Earned	142	24.6%	37.3%	38.0%	9(6.3%)	26(18.3%)	53(37.3%)	21(14.8%)	8(5.6%)	2(1.4%)	3(2.1%)	1(0.7%)	1(0.7%)	3(2.1%)	4(2.8%)	4(2.8%)	1(0.7%)	1(0.7%)	0(0%)	0(0%)	0(0%)	0(0%)	1(0.7%)	0(0%)	4(2.8%)
03 Self-Reported A Grades Earned	154	18.2%	31.8%	50.0%	2(1.3%)	8(5.2%)	18(11.7%)	49(31.8%)	32(20.8%)	3(1.9%)	8(5.2%)	6(3.9%)	4(2.6%)	4(2.6%)	3(1.9%)	2(1.3%)	0(0%)	4(2.6%)	2(1.3%)	1(0.6%)	3(1.9%)	0(0%)	2(1.3%)	0(0%)	3(1.9%)
04 Self-Reported A Grades Earned	133	14.3%	37.6%	48.1%	1(0.8%)	2(1.5%)	3(2.3%)	13(9.8%)	50(37.6%)	17(12.8%)	9(6.8%)	8(6%)	6(4.5%)	3(2.3%)	6(4.5%)	3(2.3%)	4(3%)	0(0%)	1(0.8%)	0(0%)	1(0.8%)	0(0%)	1(0.8%)	0(0%)	5(3.8%)
05 Self-Reported A Grades Earned	160	16.3%	25.0%	58.8%	3(1.9%)	3(1.9%)	2(1.3%)	4(2.5%)	14(8.8%)	40(25%)	27(16.9%)	6(3.8%)	13(8.1%)	6(3.8%)	5(3.1%)	4(2.5%)	4(2.5%)	3(1.9%)	4(2.5%)	3(1.9%)	2(1.3%)	2(1.3%)	2(1.3%)	1(0.6%)	12(7.5%)
06 Self-Reported A Grades Earned	90	18.9%	36.7%	44.4%	0(0%)	2(2.2%)	0(0%)	0(0%)	4(4.4%)	11(12.2%)	33(36.7%)	8(8.9%)	5(5.6%)	3(3.3%)	6(6.7%)	3(3.3%)	2(2.2%)	1(1.1%)	3(3.3%)	1(1.1%)	0(0%)	2(2.2%)	1(1.1%)	0(0%)	5(5.6%)
07 Self-Reported A Grades Earned	68	17.6%	13.2%	69.1%	1(1.5%)	0(0%)	1(1.5%)	0(0%)	1(1.5%)	1(1.5%)	8(11.8%)	9(13.2%)	6(8.8%)	5(7.4%)	10(14.7%)	4(5.9%)	4(5.9%)	3(4.4%)	6(8.8%)	3(4.4%)	1(1.5%)	1(1.5%)	1(1.5%)	0(0%)	3(4.4%)
08 Self-Reported A Grades Earned	54	13.0%	25.9%	61.1%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	2(3.7%)	1(1.9%)	4(7.4%)	14(25.9%)	3(5.6%)	4(7.4%)	4(7.4%)	0(0%)	1(1.9%)	4(7.4%)	3(5.6%)	2(3.7%)	2(3.7%)	1(1.9%)	1(1.9%)	8(14.8%)
09 Self-Reported A Grades Earned	44	13.6%	29.5%	56.8%	0(0%)	0(0%)	0(0%)	1(2.3%)	0(0%)	2(4.5%)	0(0%)	2(4.5%)	1(2.3%)	13(29.5%)	1(2.3%)	2(4.5%)	2(4.5%)	3(6.8%)	3(6.8%)	2(4.5%)	3(6.8%)	1(2.3%)	0(0%)	1(2.3%)	7(15.9%)
10 Self-Reported A Grades Earned	62	25.8%	17.7%	56.5%	0(0%)	0(0%)	0(0%)	0(0%)	1(1.6%)	1(1.6%)	5(8.1%)	1(1.6%)	5(8.1%)	3(4.8%)	11(17.7%)	8(12.9%)	1(1.6%)	5(8.1%)	3(4.8%)	1(1.6%)	2(3.2%)	3(4.8%)	0(0%)	2(3.2%)	10(16.1%)
11 Self-Reported A Grades Earned	18	11.1%	38.9%	50.0%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(5.6%)	1(5.6%)	7(38.9%)	2(11.1%)	2(11.1%)	1(5.6%)	1(5.6%)	0(0%)	0(0%)	0(0%)	0(0%)	3(16.7%)
12 Self-Reported A Grades Earned	33	30.3%	33.3%	36.4%	0(0%)	0(0%)	0(0%)	2(6.1%)	1(3%)	0(0%)	0(0%)	2(6.1%)	0(0%)	0(0%)	1(3%)	4(12.1%)	11(33.3%)	2(6.1%)	1(3%)	1(3%)	0(0%)	3(9.1%)	1(3%)	0(0%)	4(12.1%)
13 Self-Reported A Grades Earned	16	25.0%	50.0%	25.0%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(6.3%)	1(6.3%)	1(6.3%)	1(6.3%)	8(50%)	1(6.3%)	0(0%)	3(18.8%)	0(0%)	0(0%)	0(0%)	0(0%)
14 Self-Reported A Grades Earned	14	21.4%	21.4%	57.1%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(7.1%)	1(7.1%)	1(7.1%)	0(0%)	3(21.4%)	4(28.6%)	1(7.1%)	1(7.1%)	0(0%)	0(0%)	2(14.3%)
15 Self-Reported A Grades Earned	32	31.3%	28.1%	40.6%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(3.1%)	0(0%)	0(0%)	0(0%)	1(3.1%)	5(15.6%)	0(0%)	3(9.4%)	9(28.1%)	2(6.3%)	1(3.1%)	0(0%)	1(3.1%)	9(28.1%)
16 Self-Reported A Grades Earned	16	25.0%	43.8%	31.3%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(6.3%)	2(12.5%)	1(6.3%)	7(43.8%)	2(12.5%)	1(6.3%)	2(12.5%)	0(0%)
17 Self-Reported A Grades Earned	13	53.8%	30.8%	15.4%	1(7.7%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(7.7%)	0(0%)	1(7.7%)	0(0%)	0(0%)	0(0%)	1(7.7%)	1(7.7%)	0(0%)	1(7.7%)	1(7.7%)	4(30.8%)	1(7.7%)	0(0%)	1(7.7%)
18 Self-Reported A Grades Earned	13	38.5%	46.2%	15.4%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(7.7%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	1(7.7%)	0(0%)	1(7.7%)	0(0%)	2(15.4%)	6(46.2%)	1(7.7%)	1(7.7%)
19 Self-Reported A Grades Earned	8	0.0%	50.0%	50.0%	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	4(50%)	4(50%)
20 or More Self-Reported A Grades Earned	137	19.0%	81.0%	N/A	0(0%)	0(0%)	1(0.7%)	0(0%)	1(0.7%)	0(0%)	0(0%)	1(0.7%)	1(0.7%)	1(0.7%)	1(0.7%)	1(0.7%)	1(0.7%)	4(2.9%)	2(1.5%)	3(2.2%)	1(0.7%)	2(1.5%)	3(2.2%)	3(2.2%)	111(81%)
Item Non Respondents	288	N/A	N/A	N/A	39(13.5%)	16(5.6%)	23(8%)	20(6.9%)	21(7.3%)	18(6.3%)	19(6.6%)	8(2.8%)	15(5.2%)	15(5.2%)	11(3.8%)	6(2.1%)	5(1.7%)	6(2.1%)	6(2.1%)	9(3.1%)	6(2.1%)	10(3.5%)	3(1%)	4(1.4%)	28(9.7%)
Unit Non Respondents	4363	N/A	N/A	N/A	294(6.7%)	273(6.3%)	325(7.4%)	361(8.3%)	375(8.6%)	284(6.5%)	245(5.6%)	217(5%)	205(4.7%)	191(4.4%)	150(3.4%)	156(3.6%)	146(3.3%)	118(2.7%)	123(2.8%)	103(2.4%)	113(2.6%)	88(2%)	91(2.1%)	64(1.5%)	441(10.1%)

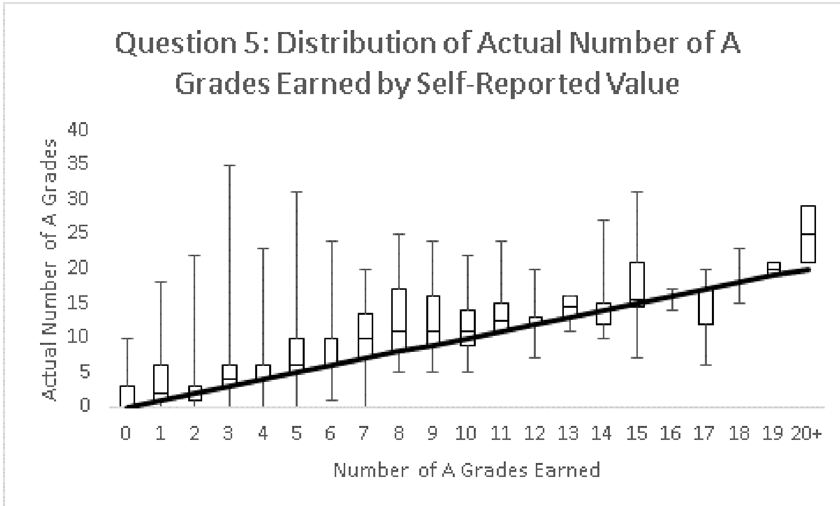

Figure E.1. Distribution of self-reported and institution-reported A grades earned.

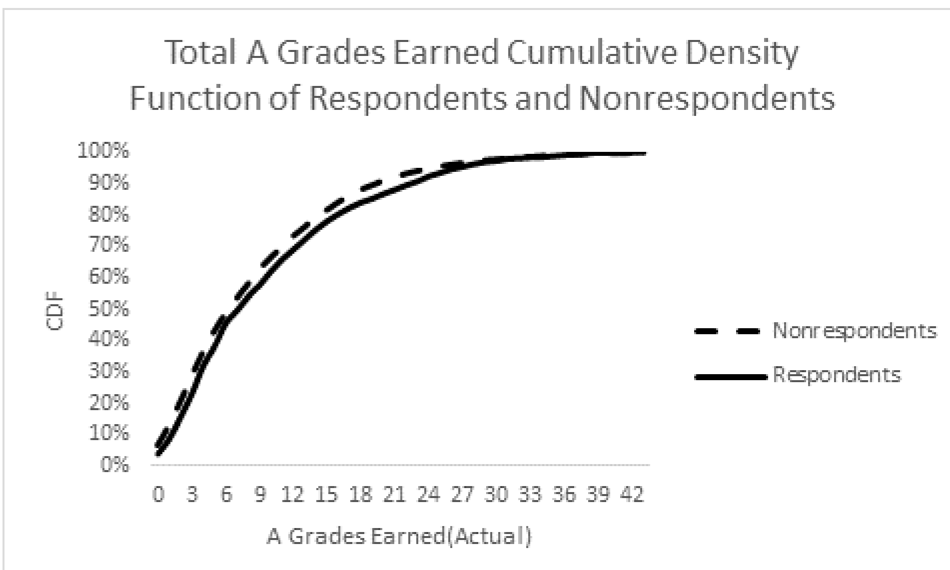

Figure E.2. CDF of question 5.

Appendix F

Question 6: PE/HES Courses Taken

Table F.1.

Summary of Question 6 Results.

		Actual HES Courses Taken				
Respondent Status	N	0	1	2	3	4 or More
Respondents	1316	302 (22.9%)	410 (31.2%)	416 (31.6%)	86 (6.5%)	102 (7.8%)
Non-Respondents	4642	1195 (25.7%)	1521 (32.8%)	1317 (28.4%)	367 (7.9%)	242 (5.2%)
Sample	5958	1497 (25.1%)	1931 (32.4%)	1733 (29.1%)	453 (7.6%)	344 (5.8%)
Population	20775	5128 (24.7%)	6807 (32.8%)	5929 (28.5%)	1662 (8%)	1249 (6%)
Self Reported Data	N	0	1	2	3	4 or More
0 Self-Reported HES Classes Taken	335	290 (86.6%)	29 (8.7%)	8 (2.4%)	5 (1.5%)	3 (0.9%)
1 Self-Reported HES Classes Taken	413	6 (1.5%)	363 (87.9%)	34 (8.2%)	5 (1.2%)	5 (1.2%)
2 Self-Reported HES Classes Taken	410	5 (1.2%)	13 (3.2%)	352 (85.9%)	23 (5.6%)	17 (4.1%)
3 Self-Reported HES Classes Taken	86	1 (1.2%)	4 (4.7%)	14 (16.3%)	49 (57%)	18 (20.9%)
4 or More Self-Reported HES Classes Taken	72	0 (0%)	1 (1.4%)	8 (11.1%)	4 (5.6%)	59 (81.9%)
Item Non Respondents	279	87 (31.2%)	86 (30.8%)	69 (24.7%)	18 (6.5%)	19 (6.8%)
Unit Non Respondents	4363	1108 (25.4%)	1435 (32.9%)	1248 (28.6%)	349 (8%)	223 (5.1%)

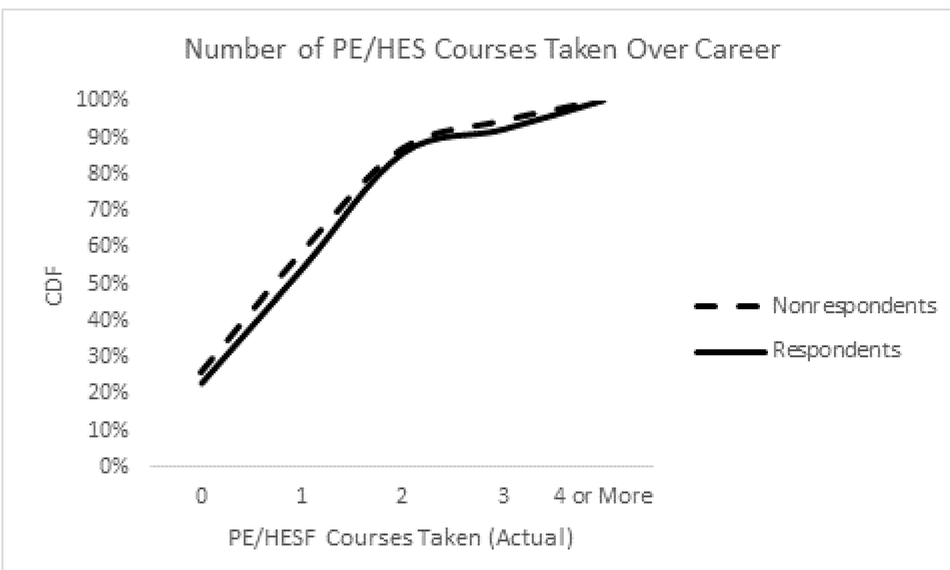

Figure F.1. CDF of question 6.

Appendix G

Question 7: Most Recent PE/HESF 100 Course Taken

Table G.1.

Summary of Question 7 Results.

Course	Description	Correctly Self Reported Percent	Respondents	Non Respondents	Sample	Population
Total		92.1%	1112(100%)	4783(100%)	6000(100%)	20778(100%)
None taken		98.2%	569(51.2%)	2778(58.1%)	3388(56.5%)	11605(55.9%)
Not Sure		0.0%	3(0.3%)			
PE/HESF 100	Cross Training	97.4%	38(3.4%)	122(2.6%)	163(2.7%)	606(2.9%)
PE/HESF 101	Fitness and Wellness	81.2%	133(12%)	370(7.7%)	484(8.1%)	1732(8.3%)
PE/HESF 102	Fitness Walking	93.2%	59(5.3%)	235(4.9%)	294(4.9%)	1074(5.2%)
PE/HESF 103	Water Aerobics	91.5%	47(4.2%)	173(3.6%)	217(3.6%)	728(3.5%)
PE/HESF 104	Swim Conditioning	96.9%	65(5.8%)	207(4.3%)	274(4.6%)	919(4.4%)
PE/HESF 105	Aerobics and Body Conditioning	91.7%	84(7.6%)	272(5.7%)	352(5.9%)	1203(5.8%)
PE/HESF 106	Triathlon	75.0%	8(0.7%)	21(0.4%)	28(0.5%)	79(0.4%)
PE/HESF 107	Run Conditioning	98.6%	74(6.7%)	262(5.5%)	337(5.6%)	1153(5.5%)
PE/HESF 108	Water Step Aerobics	100.0%	3(0.3%)	15(0.3%)	21(0.4%)	63(0.3%)
PE/HESF 109	Step Aerobics	0.0%	0(0%)	108(2.3%)	138(2.3%)	426(2.1%)
PE/HESF 110	Adapted Physical Education	0.0%	0(0%)	7(0.1%)	10(0.2%)	33(0.2%)
PE/HESF 111	Indoor Group Cycling	0.0%	29(2.6%)	213(4.5%)	294(4.9%)	1157(5.6%)

Appendix H

Question 8: Grade Earned in Most Recent PE/HESF 100 Level Course

Appendix H.1.

Summary of question 8 results.

		Actual PE/HESF 100 Grade Earned								
Respondent Status	N	Satisfactory	A-/A/A+	B-/B/B+	C-/C/C+	D-/D/D+	F/Unsatisfact.	Currently Enrolled	Incomplete or Withdrawn	Did Not Take PE/HESF 100
Respondents	724	369 (51%)	66 (9.1%)	42 (5.8%)	8 (1.1%)	2 (0.3%)	1 (0.1%)	113 (15.6%)	12 (1.7%)	111 (15.3%)
Non-Respondents	1986	1158 (58.3%)	209 (10.5%)	152 (7.7%)	38 (1.9%)	2 (0.1%)	22 (1.1%)	361 (18.2%)	44 (2.2%)	0 (0%)
Sample	2710	1527 (56.3%)	275 (10.1%)	194 (7.2%)	46 (1.7%)	4 (0.1%)	23 (0.8%)	474 (17.5%)	56 (2.1%)	111 (4.1%)
Population	9267	5273 (56.9%)	988 (10.7%)	591 (6.4%)	151 (1.6%)	13 (0.1%)	82 (0.9%)	1846 (19.9%)	212 (2.3%)	111 (1.2%)
Self Reported Data	N	Satisfactory	A-/A/A+	B-/B/B+	C-/C/C+	D-/D/D+	F/Unsatisfact.	Currently Enrolled	Incomplete or Withdrawn	Did Not Take PE/HESF 100
Satisfactory	398	342 (85.9%)	1 (0.3%)	5 (1.3%)	1 (0.3%)	0 (0%)	1 (0.3%)	16 (4%)	9 (2.3%)	23 (5.8%)
A-/A/A+	125	16 (12.8%)	62 (49.6%)	10 (8%)	0 (0%)	0 (0%)	0 (0%)	4 (3.2%)	1 (0.8%)	32 (25.6%)
B-/B/B+	43	7 (16.3%)	2 (4.7%)	26 (60.5%)	0 (0%)	0 (0%)	0 (0%)	2 (4.7%)	0 (0%)	6 (14%)
C-/C/C+	11	2 (18.2%)	1 (9.1%)	0 (0%)	7 (63.6%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	0 (0%)
D-/D/D+	2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
F/Unsatisfactory	4	1 (25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (75%)	0 (0%)	0 (0%)
Currently Enrolled (no grade)	109	1 (0.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	85 (78%)	2 (1.8%)	21 (19.3%)
Not Sure	32	0 (0%)	0 (0%)	1 (3.1%)	0 (0%)	0 (0%)	0 (0%)	2 (6.3%)	0 (0%)	29 (90.6%)

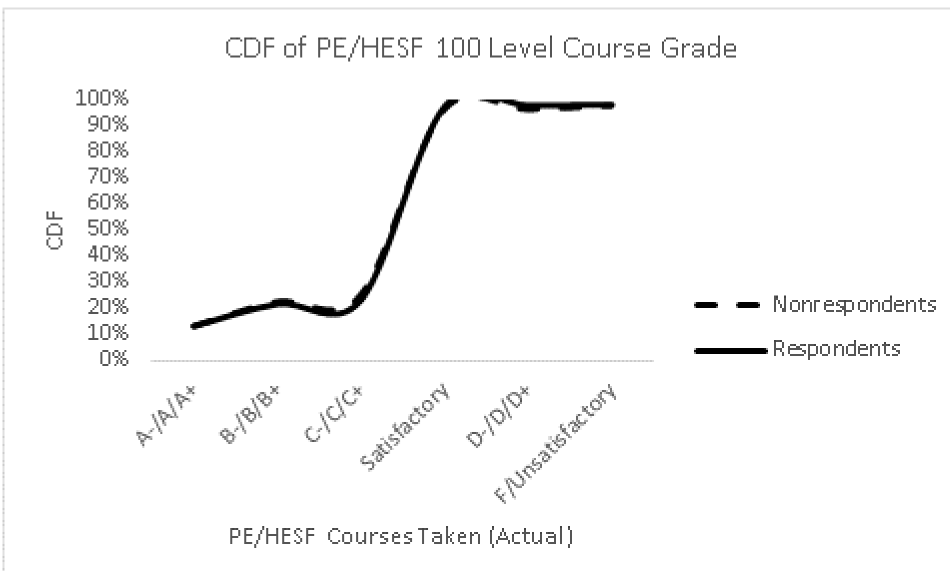

Figure H.1. CDF of question 8.

Appendix I

Question 9: PE/HESF 100 Level Course Class Skipping

Table I.1.

Summary of Question 9 Results.

		Actual PE/HESF 100 Classes Skipped								
Respondent Status	N	0 Never Skipped	Skipped 1	Skipped 2	Skipped 3	Skipped 4	Skipped 5	Skipped 6	Skipped 7 or More	Took A DE Section
Respondents	311	36 (11.6%)	37 (11.9%)	58 (18.6%)	58 (18.6%)	41 (13.2%)	28 (9%)	14 (4.5%)	17 (5.5%)	22 (7.1%)
Non-Respondents	985	95 (9.6%)	123 (12.5%)	143 (14.5%)	167 (17%)	143 (14.5%)	108 (11%)	38 (3.9%)	52 (5.3%)	116 (11.8%)
Sample	1296	131 (10.1%)	160 (12.3%)	201 (15.5%)	225 (17.4%)	184 (14.2%)	136 (10.5%)	52 (4%)	69 (5.3%)	138 (10.6%)
Population	4496	470 (10.5%)	563 (12.5%)	725 (16.1%)	773 (17.2%)	659 (14.7%)	441 (9.8%)	157 (3.5%)	252 (5.6%)	456 (10.1%)
Self Reported Data	N	0 Never Skipped	Skipped 1	Skipped 2	Skipped 3	Skipped 4	Skipped 5	Skipped 6	Skipped 7 or More	Took A DE Section
0	49	20 (40.8%)	11 (22.4%)	4 (8.2%)	5 (10.2%)	2 (4.1%)	2 (4.1%)	1 (2%)	2 (4.1%)	2 (4.1%)
1	72	11 (15.3%)	11 (15.3%)	22 (30.6%)	11 (15.3%)	8 (11.1%)	4 (5.6%)	2 (2.8%)	2 (2.8%)	1 (1.4%)
2	84	5 (6%)	11 (13.1%)	17 (20.2%)	27 (32.1%)	9 (10.7%)	10 (11.9%)	3 (3.6%)	2 (2.4%)	0 (0%)
3	46	0 (0%)	4 (8.7%)	12 (26.1%)	9 (19.6%)	9 (19.6%)	4 (8.7%)	5 (10.9%)	3 (6.5%)	0 (0%)
4	29	0 (0%)	0 (0%)	2 (6.9%)	5 (17.2%)	8 (27.6%)	5 (17.2%)	2 (6.9%)	7 (24.1%)	0 (0%)
5	11	0 (0%)	0 (0%)	1 (9.1%)	1 (9.1%)	4 (36.4%)	3 (27.3%)	1 (9.1%)	1 (9.1%)	0 (0%)
6	1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
7 or more	0	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
I took a distance education section	19	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	19 (100%)

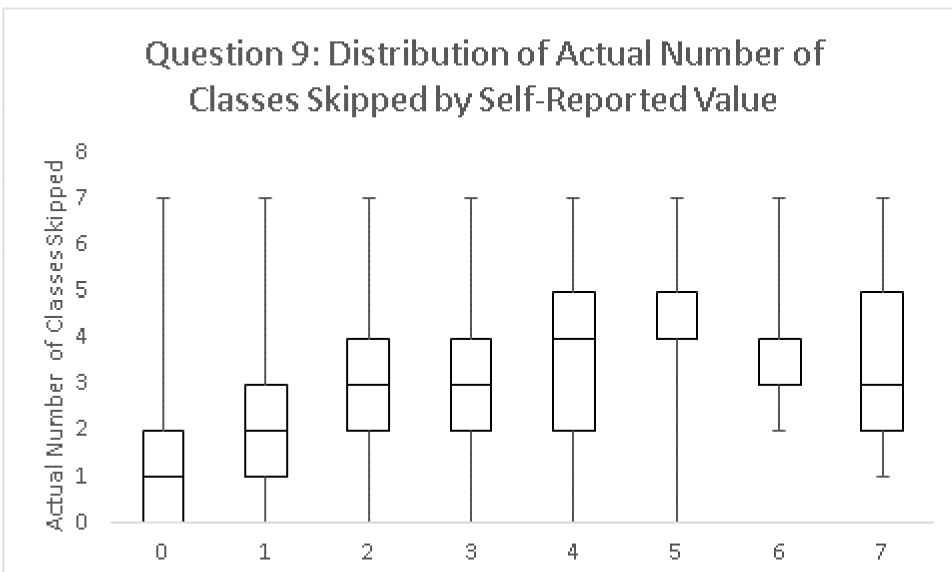

Figure I.1. Self-reported and institution-reported class skipping.

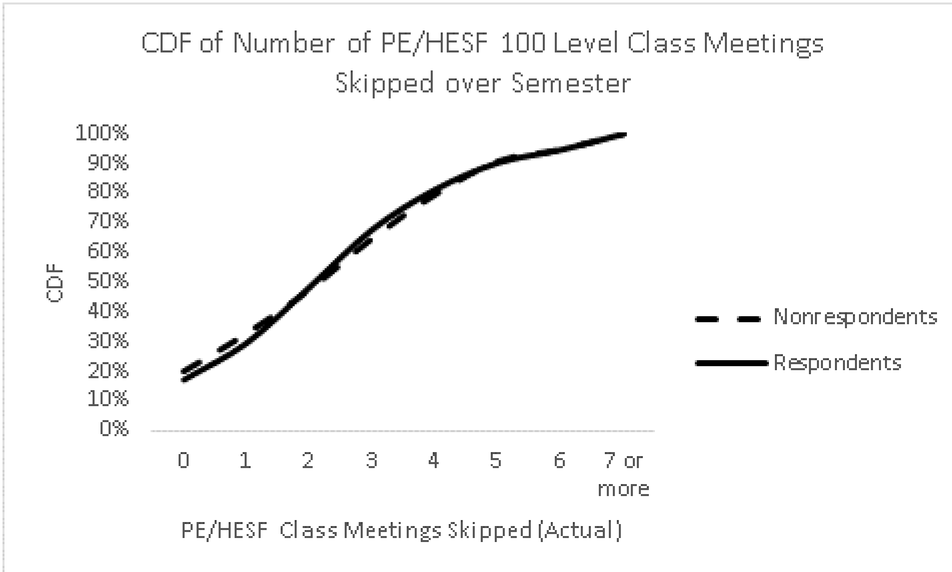

Figure I.2. CDF of question 9.

Appendix J

Question 10: Co-Curricular Participation

Question 10A: Intramural Athletics Participation

Table J.1.

Summary of Question 10A Results.

		Actual Rate of Participation			
Respondent Status	N	Never	One Semester	Multiple Semesters	All Semesters
Respondents	1253	780 (62.3%)	174 (13.9%)	217 (17.3%)	82 (6.5%)
Non-Respondents	4747	3046 (64.2%)	705 (14.9%)	811 (17.1%)	185 (3.9%)
Sample	6000	3826 (63.8%)	879 (14.7%)	1028 (17.1%)	267 (4.5%)
Population	20775	13295 (64%)	3003 (14.5%)	3510 (16.9%)	967 (4.7%)
Self Reported Data	N	Never	One Semester	Multiple Semesters	All Semesters
Never	751	715 (95.2%)	32 (4.3%)	3 (0.4%)	1 (0.1%)
One Semester	198	49 (24.7%)	113 (57.1%)	29 (14.6%)	7 (3.5%)
Multiple Semesters	200	9 (4.5%)	26 (13%)	141 (70.5%)	24 (12%)
All Semesters	104	7 (6.7%)	3 (2.9%)	44 (42.3%)	50 (48.1%)

Shading denotes self-reported and institution reported match

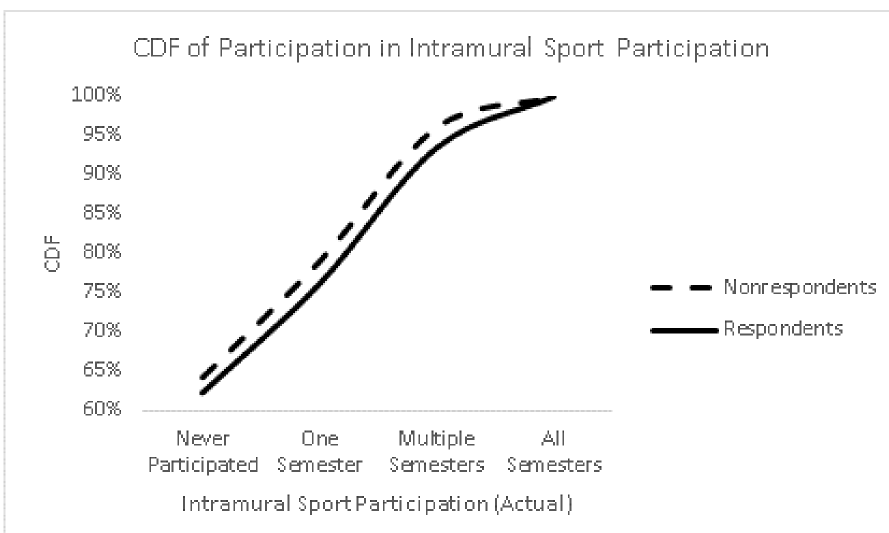

Figure J.1. CDF of question 10A.

Table J.2.

Summary of Question 10B Results.

		Actual Rate of Participation			
Respondent Status	N	Never	One Semester	Multiple Semesters	All Semesters
Respondents	1253	780 (62.3%)	174 (13.9%)	217 (17.3%)	82 (6.5%)
Non-Respondents	4747	3046 (64.2%)	705 (14.9%)	811 (17.1%)	185 (3.9%)
Sample	6000	3826 (63.8%)	879 (14.7%)	1028 (17.1%)	267 (4.5%)
Population	20775	13295 (64%)	3003 (14.5%)	3510 (16.9%)	967 (4.7%)
Self Reported Data	N	Never	One Semester	Multiple Semesters	All Semesters
Never	751	715 (95.2%)	32 (4.3%)	3 (0.4%)	1 (0.1%)
One Semester	198	49 (24.7%)	113 (57.1%)	29 (14.6%)	7 (3.5%)
Multiple Semesters	200	9 (4.5%)	26 (13%)	141 (70.5%)	24 (12%)
All Semesters	104	7 (6.7%)	3 (2.9%)	44 (42.3%)	50 (48.1%)

Shading denotes self-reported and institution reported match

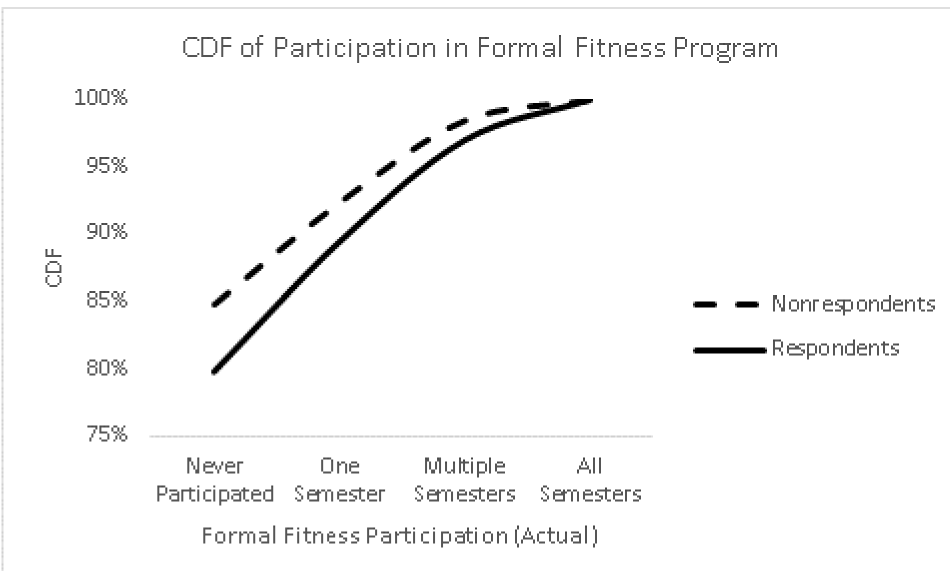

Figure J.2. CDF of question 10B.

Table J.3.

Summary of Question 10C Results.

		Actual Rate of Participation			
Respondent Status	N	Never	One Semester	Multiple Semesters	All Semesters
Respondents	1262	771 (61.1%)	218 (17.3%)	219 (17.4%)	54 (4.3%)
Non-Respondents	4738	3243 (68.4%)	633 (13.4%)	748 (15.8%)	114 (2.4%)
Sample	6000	4014 (66.9%)	851 (14.2%)	967 (16.1%)	168 (2.8%)
Population	20775	13933 (67.1%)	3034 (14.6%)	3172 (15.3%)	636 (3.1%)
Self Reported Data	N	Never	One Semester	Multiple Semesters	All Semesters
Never	679	611 (90%)	54 (8%)	11 (1.6%)	3 (0.4%)
One Semester	246	94 (38.2%)	103 (41.9%)	34 (13.8%)	15 (6.1%)
Multiple Semesters	250	61 (24.4%)	52 (20.8%)	121 (48.4%)	16 (6.4%)
All Semesters	87	5 (5.7%)	9 (10.3%)	53 (60.9%)	20 (23%)

Shading denotes self-reported and institution reported match

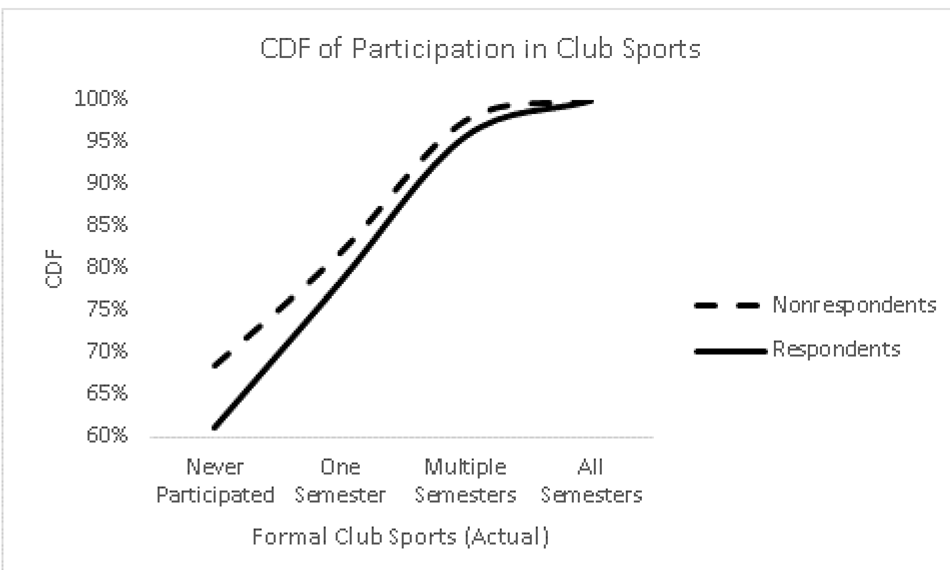

Figure J.3. CDF of question 10C.

Table J.4.

Summary of Question 10D Results.

		Actual Rate of Participation			
Respondent Status	N	Never	One Semester	Multiple Semesters	All Semesters
Respondents	1241	969 (78.1%)	184 (14.8%)	81 (6.5%)	7 (0.6%)
Non-Respondents	4759	4020 (84.5%)	503 (10.6%)	219 (4.6%)	17 (0.4%)
Sample	6000	4989 (83.2%)	687 (11.5%)	300 (5%)	24 (0.4%)
Population	20775	17309 (83.3%)	2371 (11.4%)	1001 (4.8%)	94 (0.5%)
Self Reported Data	N	Never	One Semester	Multiple Semesters	All Semesters
Never	1099	887 (80.7%)	158 (14.4%)	49 (4.5%)	5 (0.5%)
One Semester	82	44 (53.7%)	16 (19.5%)	20 (24.4%)	2 (2.4%)
Multiple Semesters	39	23 (59%)	7 (17.9%)	9 (23.1%)	0 (0%)
All Semesters	21	15 (71.4%)	3 (14.3%)	3 (14.3%)	0 (0%)

Shading denotes self-reported and institution reported match

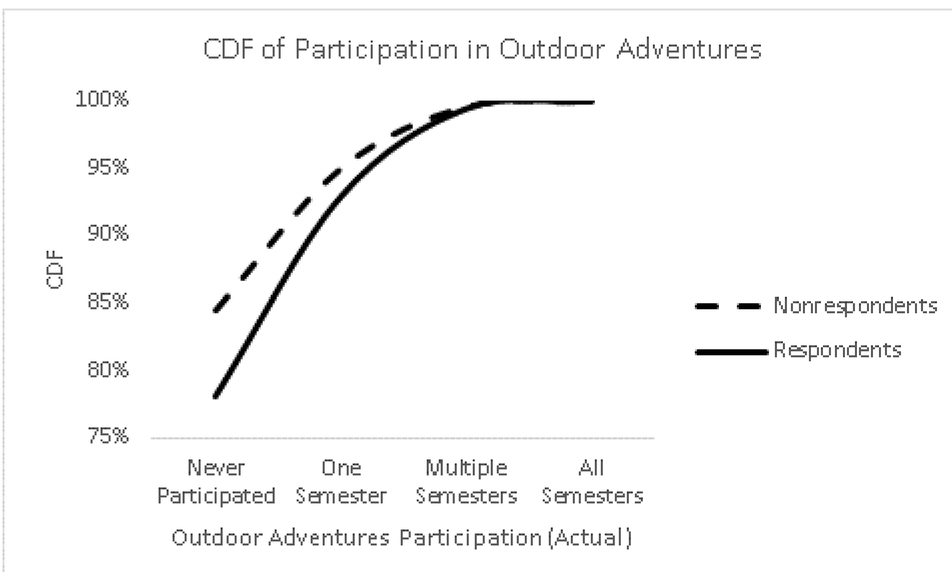

Figure J.4. CDF of question 10D.

Appendix K

Questions 11 to 23: Shortened Marlow-Crowne Social Desirability Scale

Table K.1.

MCSDS Scores by Demographics and Background.

MCSDS Score by Demographic Group				
	Respondent Group	N	Mean	Std Dev
Grand Total				
	Total	1246	6.57	2.775
Gender				
	Female	649	6.47	2.850
	Male	597	6.67	2.690
Ethnicity				
	African American	62	7.00	2.986
	Asian	50	6.70	2.873
	Hispanic	64	7.50	2.430
	International	25	7.36	2.914
	Two or More	57	6.72	2.596
	Unknown and Other Minority	28	6.11	3.143
	White	960	6.45	2.762
Age				
	19 and Younger	439	6.54	2.884
	20 to 22	674	6.50	2.709
	23 to 25	78	7.17	2.360
	26 or Older	55	6.67	3.174
Academic Year				
	Enrolled 1 Year	370	6.51	2.785
	Enrolled 2 Years	243	6.63	2.822
	Enrolled 3 Years	223	6.70	2.924
	Enrolled 4 Years	195	6.61	2.667
	Enrolled 5 Years	134	6.39	2.615
	Enrolled 6 or More Years	81	6.46	2.734
Entry Type				
	Freshman	1030	6.48	2.766
	Transfer	216	6.95	2.792
Cumulative GPA				

1st Quintile GPA (0.00 to 2.84)	250	6.86	2.888
2nd Quintile GPA (2.85 to 3.21)	249	6.65	2.702
3rd Quintile GPA (3.22 to 3.50)	249	6.56	2.511
4th Quintile GPA (3.51 to 3.82)	249	6.32	2.920
5th Quintile GPA (3.83 to 4.00)	249	6.43	2.825
SAT Score			
1st Quartile (720 to 1180)	275	6.92	2.597
2nd Quartile (1190 to 1250)	270	6.56	2.912
3rd Quartile (1260 to 1340)	295	6.34	2.733
4th Quartile (1350 to 1600)	248	6.18	2.790
No SAT Score	158	6.99	2.793
Pell Grant Status			
No FAFSA	569	6.53	2.805
No Pell Grant	466	6.39	2.757
Pell Grant Recipient	211	7.04	2.693

Table K.2.

Duncan Groupings by MCSDS Score.

	Under Reporters	Correct Reporters	Over Reporters	Other Misreporters
Q1A: CRF Use - Prior 7 Days	5.93	6.41	6.87	
Q1C: CRF Use - Weekly Rate	6.77	6.50	6.57	
Q2: Cumulative GPA	**6.72 (A,B)**	**6.34 (A)**	**6.75 (B)**	
Q3: Semester GPA	**6.57 (A,B)**	**6.31 (A)**	**6.76 (B)**	
Q4: D/F/U Grades Earned	6.69	6.55	6.16	
Q5: A Grades Earned	6.47	6.57	6.79	
Q6: HESF Course Taken	6.41	6.57	6.78	
Q7: Most Recent HESF Course		6.65		6.48
Q8: HESF Grade	5.67	6.57	6.97	6.78
Q9: HESF Class Skipping	6.66	6.68	6.37	6.47
Q10A: Intramural Sports Partic.	6.38	6.65	6.32	
Q10B: Fitness Program Partic.	6.61	6.55	6.50	
Q10C: Club Sports Partic.	6.51	6.55	6.57	
Q10D: Outdoor Adventures Partic.	**6.43(B)**	**6.53(A,B)**	**7.07(A)**	

Note. Bold denotes Duncan Test intra-question statistically significant mean differences (P<.05). Means with different letters are significantly different from means in other letter groups. "Other Misreports" describe incorrect self-reports that lack a numeric relationship and cannot be classified as under or over institution reports

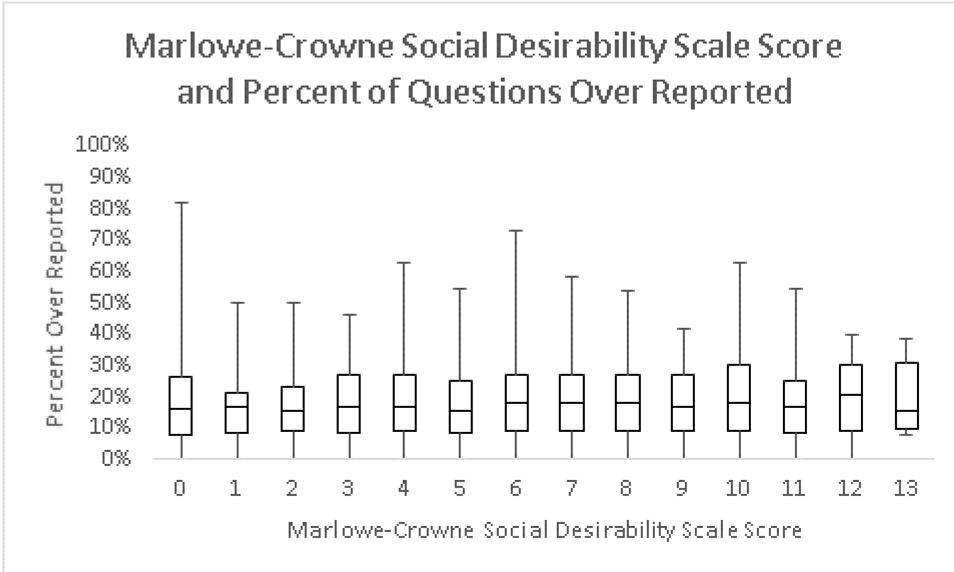

Figure K.1. MCSDS score and over reporting.

Table K.3.

Logistic Regression on Likelihood to Over Report 7 Days of CRF Use

Logistic Regression: Probability of Respondent Over Reporting on Prior 7 Days of CRF Usage

		All Respondents		First Wave Respondents Only		2nd & 3rd Wave Respondents Only	
		Model 1: No MSCDS Score	Model 2: MCSDS Score	Model 3: No MSCDS Score	Model 4: MCSDS Score	Model 3: No MSCDS Score	Model 4: MCSDS Score
N		398	398	227	227	171	171
χ^2 (Likelihood Ratio)		56.1	47.5	38.6	42.5	17.8	18.2
Intercept	B	-0.224	-0.812	3.240	2.410	-12.507	-12.764
	Std. Err.	1.3034	1.3406	1.7566	1.8002	232.8000	232.2000
Female	B	-.509*	-0.510*	-0.614*	-0.619*	-0.402	-0.399
	Std. Err.	0.2236	0.2252	0.3048	0.3085	0.3448	0.3454
African American	B	1.545	1.636	-0.076	-0.026	12.448	12.515
	Std. Err.	1.2080	1.2092	1.5726	1.5732	232.8000	232.2000
Asian	B	2.419	2.476*	0.797	0.848	13.282	13.323
	Std. Err.	1.2429	1.2436	1.6367	1.6392	232.8000	232.2000
Hispanic	B	1.025	0.901	-0.543	-0.825	11.473	11.473
	Std. Err.	1.2301	1.2328	1.6017	1.6078	232.8000	232.2000
International	B	0.691	0.703	-0.870	-0.891		
	Std. Err.	1.6226	1.6224	1.8658	1.8668		
Two Plus Races	B	1.329	1.401	0.203	0.274	11.141	11.187
	Std. Err.	1.2190	1.2209	1.5607	1.5698	232.8000	232.2000
White	B	1.528	1.574	0.262	0.274	11.740	11.7826
	Std. Err.	1.1235	1.1228	1.4421	1.4417	232.7000	232.2000
Semesters Enrolled	B	-.219**	-0.224**	-0.379**	-0.382**	-0.063	-0.067
	Std. Err.	0.0541	0.0544	0.0817	0.0826	0.0783	0.0787

Cumulative GPA	B	-0.372	-0.373	-0.732*	-0.694*	0.040	0.029
	Std. Err.	0.2065	0.2077	0.3012	0.3023	0.3306	0.3316
Number of Co-Curricular Activities	B	.570**	0.588**	0.644**	0.671**	0.476**	0.483**
	Std. Err.	0.1105	0.1125	0.1503	0.1541	0.1724	0.1742
Wave 1 Respondent	B	0.313	0.316				
	Std. Err.	0.2194	0.2204				
MCSDS Score	B		0.083*		0.104		0.040
	Std. Err.		0.0402		0.0532		0.0651

*denotes statistically significance at the .05 level

**denotes statistically significance at the .01 level

Table K.4.

Logistic Regression on Likelihood to Over Report Prior Year CRF Use

Logistic Regression: Probability of Respondent Over Reporting on Academic Year CRF Use			
		All Respondents	
		Model 1: No MSCDS Score	Model 2: MCSDS Score
N		408	408
χ^2 (Likelihood Ratio)		8.4	8.5
Intercept	B	1.029	1.119
	Std. Dev.	0.9189	0.9603
Female	B	0.205	0.203
	Std. Dev.	0.2062	0.2063
African American	B	0.288	0.306
	Std. Dev.	0.7423	0.7446
Asian	B	0.174	0.182
	Std. Dev.	0.7262	0.7270
Hispanic	B	0.012	0.020
	Std. Dev.	0.7457	0.7463
International	B	-0.387	-0.380
	Std. Dev.	0.8513	0.8527
Two Plus Races	B	0.147	0.171
	Std. Dev.	0.7520	0.7559
White	B	-0.226	-0.220
	Std. Dev.	0.5766	0.5772
Semesters Enrolled	B	-0.013	-0.014
	Std. Dev.	0.0458	0.0458
Cumulative GPA	B	-0.390	-0.396*
	Std. Dev.	0.1955	0.1962
Number of Co-Curricular Activities	B	0.046	0.047
	Std. Dev.	.0963	0.0963
Wave 1 Respondent	B	0.141	0.142
	Std. Dev.	0.2078	0.2078
MCSDS Score	B		-0.012
	Std. Dev.		0.0367

*denotes statistically significance at the .05 level

**denotes statistically significance at the .01 level

Appendix L

Question 24: Belief Exercise is Important to Academic Success

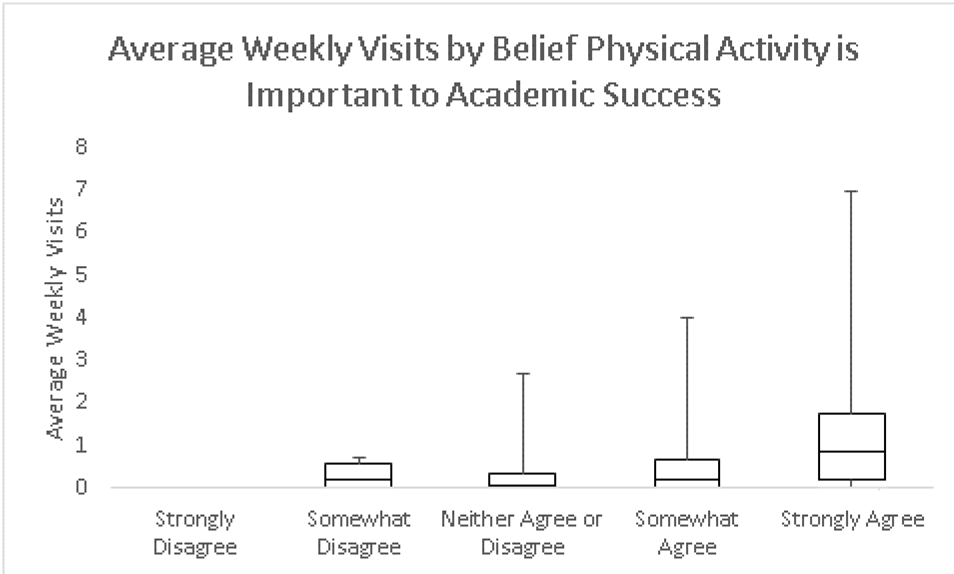

Figure L.1. CRF use by question 24.

Appendix M

Question 25: Share of Physical Activity Occurring Inside CRF

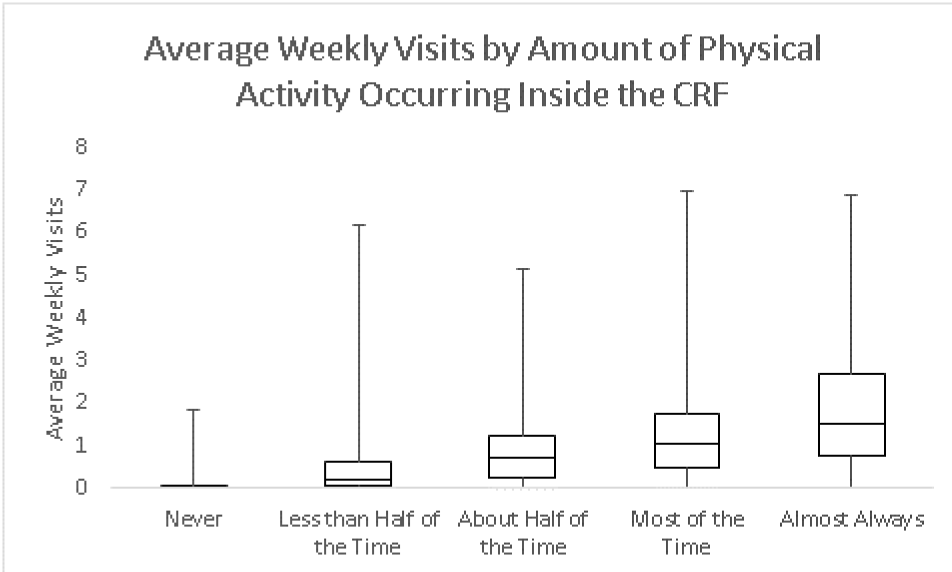

Figure M.1. Distribution of question 25.

Appendix N

Question 26: Open Ended Question 1

Table N.1.

Open-ended Question Length by Demographics and Background.

Question 26 Word Count by Demographic Groups				
	Respondent Group	N	Mean	Std Dev
Grand Total				
	Total	1595	7.53	7.735
Gender				
	Female	807	7.90	7.790
	Male	788	7.15	7.665
Ethnicity				
	African American	83	5.93	7.180
	Asian	68	6.99	7.349
	Hispanic	87	7.13	7.240
	International	48	4.06	6.366
	Two or More	72	8.74	8.294
	Unknown and Other Minority	31	8.68	7.812
	White	1206	7.73	7.803
Age				
	19 and Younger	558	7.78	7.754
	20 to 22	860	7.57	7.874
	23 to 25	109	5.91	6.677
	26 and Older	68	7.59	7.207
Academic Year				
	Enrolled 1 Year	484	7.36	7.694
	Enrolled 2 Years	313	7.42	7.644
	Enrolled 3 Years	277	7.63	7.719
	Enrolled 4 Years	241	8.21	7.927
	Enrolled 5 Years	174	7.61	8.224
	Enrolled 6 or More Years	106	6.65	6.978
Entry Type				
	Freshmen	1309	7.70	7.827
	Transfer	286	6.73	7.261
Cumulative GPA				

	1st Quintile GPA (0.00 to 2.84)	385	5.89	7.501
	2nd Quintile GPA (2.85 to 3.21)	316	7.34	7.586
	3rd Quintile GPA (3.22 to 3.50)	314	6.99	7.561
	4th Quintile GPA (3.51 to 3.82)	294	8.93	7.845
	5th Quintile GPA (3.83 to 4.00)	286	9.08	7.793
SAT Score				
	1st Quartile (720 to 1180)	349	7.34	7.540
	2nd Quartile (1190 to 1250)	337	8.06	7.801
	3rd Quartile (1260 to 1340)	389	7.50	7.934
	4th Quartile (1350 to 1600)	308	7.83	7.728
	No SAT Score	212	6.60	7.561
Pell Grant Status				
	No FAFSA	753	7.28	7.617
	No Pell Grant	586	7.63	7.884
	Pell Grant Recipient	256	8.03	7.735

Table N.2.

Duncan Groupings of Question 26 Word Count.

	Under Reporters	Correct Reporters	Over Reporters	Other Misreporters
Q1A: CRF Use - Prior 7 Days	7.9	8.1	8.3	
Q1C: CRF Use - Weekly Rate	**10.2(A)**	**8.2(A,B)**	**6.8(B)**	
Q2: Cumulative GPA	**8.3(B)**	**9.7(A)**	**8.5(B)**	
Q3: Semester GPA	**8.9(A,B)**	**9.8(A)**	**8.5(B)**	
Q4: D/F/U Grades Earned	9.2	8.9	8.4	
Q5: A Grades Earned	**8.4(B)**	**10.1(A)**	**8.4(B)**	
Q6: HESF Course Taken	9.6	9.1	8.7	
Q7: HESF 100 Course Taken		9.5		9.4
Q8: HESF Grade	9.0	9.7	10.1	7.7
Q9: HESF Class Skipping	**9.7(A)**	**9.7(A)**	**9.0(A,B)**	**7.7(B)**
Q10A: Intramural Sports Partic.	9.2	9.5	9.4	
Q10B: Fitness Program Partic.	9.6	9.5	9.7	
Q10C: Club Sports Partic.	9.9	9.3	9.8	
Q10D: Outdoor Adventures Partic.	12.4	9.4	10	

Note. Bold denotes Duncan Test intra-question statistically significant mean differences (P<.05). Means with different letters are significantly different from means in other letter groups. "Other Misreports" describe incorrect self-reports that lack a numeric relationship and cannot be classified as under or over institution reports

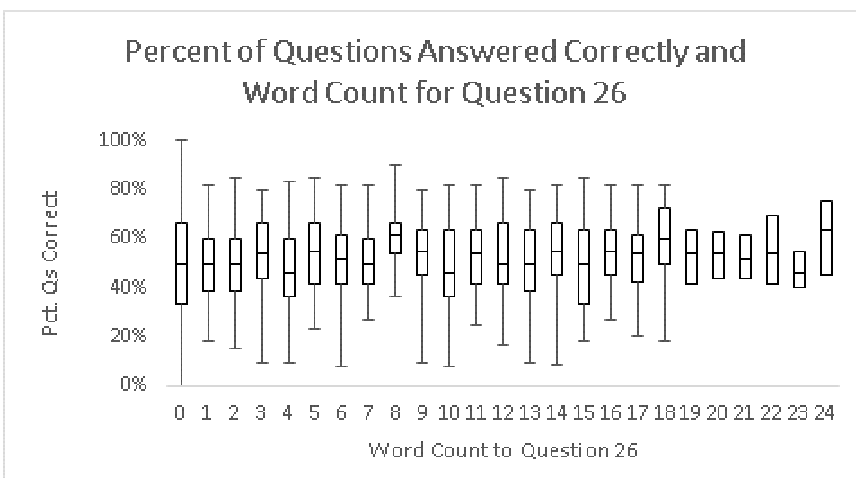

Figure N.1. Response accuracy and word count.