

**Participation Decisions and Measurement Error in Web Surveys**

**by**

**Andrey A. Peytchev**

**A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Survey Methodology)  
in The University of Michigan  
2007**

**Doctoral Committee:**

**Research Associate Professor Mick P. Couper, Chair  
Professor Robert M. Groves  
Professor Trivellore E. Raghunathan  
Professor Norbert W. Schwarz**

## Table of Contents

List of Figures .....	iii
List of Tables .....	iv
Chapter 1. Introduction .....	1
Chapter 2. Response Behaviors in Web Surveys: Breakoffs Are Different from Unit Nonresponse .....	14
1. Introduction .....	14
2. Background and Significance .....	16
3. Data and Methods .....	22
4. Results .....	26
5. Discussion and Conclusions .....	49
6. Appendix .....	53
Chapter 3. Causes of Survey Breakoffs .....	59
1. Introduction .....	59
2. Background and Significance .....	61
3. Data and Methods .....	74
4. Results .....	80
5. Discussion and Conclusions .....	87
Chapter 4. Web Survey Design and Correlated Measurement Error .....	93
1. Introduction .....	93
2. Background and Significance .....	94
3. Data and Methods .....	110
4. Results .....	117
5. Discussion and Conclusions .....	124
6. Appendix .....	128
Chapter 5. Causes of Correlated Measurement Error by Layout Design .....	129
1. Introduction .....	129
2. Background and Significance .....	130
3. Data and Methods .....	143
4. Results .....	147
5. Discussion and Conclusions .....	159
6. Appendix .....	163
Chapter 6. Conclusion .....	164
References .....	175

## List of Figures

Figure 1. Conceptual Model for Web Survey Participation .....	6
Figure 2. Student Life Surveys. ....	53
Figure 3. Living-Learning Program Surveys. ....	53
Figure 4: Breakoff by Page Number in Study 1 .....	81
Figure 5: Kaplan-Meier Survival Estimates by Study and Sample Source .....	82
Figure 6. Schematic Representation of the Full (Hybrid) Structural Equation Model Used and Its Components .....	106
Figure 7. Measurement Model in the Multi-group Analysis .....	116
Figure 8. (a) Related Instructions, (b) Different Pages Layout, Related Instructions, (c) Same Page Separate Question Layout, Independent Instructions, Ordered Topics, and (d) Same Page Grid Layout, No Instructions, Mixed Topics. ....	145
Figure 9. Variability of Responses Across the Six Questions by Layout Design, Location in Survey, and Order of Questions (Least Squares Adjusted Means). ....	151
Figure 10. Variability of Responses Across Questions in the Same Page Layout Designs, by Instructions on Relatedness of Questions (Least Squares Adjusted Means for Gender, Age, and Education). ....	153
Figure 11. Responses to How Closely Related the Questions Were Perceived By Layout Design and Instructions on Relatedness of Questions When Questions are (a) Ordered By Topic and (b) Mixed. (Scale 1-7, 1=Completely unrelated, 7=Completely related) .....	155
Figure 12. Variability of Responses by Layout and Instructions on Relatedness of Questions (Least Squares Adjusted Means). ....	156
Figure 13. Percent Changing Responses and Time in Seconds by Layout Design. ....	158

## List of Tables

Table 1. Response, Breakoff, Nonresponse, Completion Rates, Type of Incentive, and Informed and Median Time to Complete the Survey by Final Response Disposition.....	27
Table 2. Nonresponse, Breakoff by Location in Survey, Completes, and Completion Rates for 4,432 Respondents Invited to Both the Living Learning Program 2003 and the Student Life 2003 Surveys.....	31
Table 3. Current Survey Response Outcomes by Previous Survey Outcomes .	38
Table 4. Response Dispositions by Living Learning Program Participation across the Three Surveys.....	40
Table 5. Multinomial Regression of Breakoff and Nonrespondent Outcomes Relative to Complete Respondent in SL 2001, LLP 2003, SL 2003, LLP 2004, SL 2005 on Respondent Characteristics, Survey Characteristics, and Order of Survey (41,981 observations from 29,894 individuals selected for at least one survey).....	43
Table 6. Response Outcome for the SL 2003 Survey by Whether LLP Eligible Sample Persons were Invited to LLP 2003 a Month Earlier .....	47
Table 7. Multinomial Regression of Breakoff and Nonrespondent Outcomes Relative to Complete Respondent in LLP 2003, SL 2003, LLP 2004, SL 2005 on Respondent Characteristics, Survey Characteristics, and Response Outcome in Previous Survey (12,086 observations from 9,476 individuals selected for at least two surveys) .....	48
Table 8. Complete, Breakoff, and Nonrespondent Outcome in the Student Life and Living Learning Surveys by Gender and Year in School of Sampled Individuals.....	54
Table 9. Complete, Breakoff, and Nonrespondent Outcome in the Student Life and Living Learning Surveys by Race of Sampled Individuals. ....	55
Table 10. Response Behavior across Similar Surveys. ....	56
Table 11. Response Behavior across Different Surveys. ....	57
Table 12. Kappa Statistic and Standard Error for Each Cross Classification.....	58
Table 13: Response Dispositions for Both Studies by Sample Source.....	75
Table 14: Discrete Hazard Survival Models with Page-Varying Covariates Predicting Breakoff in the Combined Studies, SSI samples. ....	84
Table 15: Discrete Hazard Survival Models with Page-Varying Covariates Predicting Breakoff in the Combined Studies, SSI and AOL samples. ....	86
Table 16. Tests for Differences in Means between Groups for Each Question	118
Table 17. Covariances for All Questions under All Conditions .....	119
Table 18. Full Structural Model Path Coefficients and Model Fit Indices by Experimental Condition .....	120

Table 19. Multi-group Invariance Analysis Using $\chi^2$ Difference Tests .....	124
Table 20. Experimental Design (except Early vs. Late Location in the Survey): Layout, Instructions, and Topic.....	144
Table 21. Cronbach's Coefficient Alpha Estimates by Condition for Diet and Exercise Questions, and Percent of Respondents Leaving at Least One Question Unanswered.....	148
Table 22. Percent of Variability in BMI Explained by the Questions, by Layout and Instructions.....	154
Table 23. Percent of Variability in BMI Explained by the Questions, by Layout and Location in the Survey.....	157

## Chapter 1. Introduction

Inferences about a target population from sample surveys are hindered by the extent of differences between the intended measures from the selected sample and the measures actually obtained from those who cooperated with the survey request. There are two main types of deviations from the intended design: survey data are not collected from all sample elements (nonresponse), and the measures that are obtained from the respondents may deviate from those intended (measurement error). Attempts are generally made to reduce and adjust for different errors of nonobservation and for errors of observation, often informed by knowledge about some of the causal mechanisms. Both types of survey errors and their causes can vary across modes of data collection. While in-person surveys have received much attention over the past century, more recently developed modes, such as web surveys, have received far less methodological attention.

While a relatively small proportion of respondents in interviewer-administered surveys may terminate the interview prematurely, such behavior often occurs at substantial rates in web surveys (Musch and Reips, 2000; Lozar-Manfreda and Vehovar, 2002). This necessitates the addition of another response outcome to survey cooperation. Sampled persons who start the survey

but fail to reach the end (breakoffs) may be very different from those who complete the survey, but at the same time may also be different from the unit nonrespondents who do not even start the survey.

A first step toward understanding the multiple response outcomes in web surveys is to establish whether they are a different from unit nonresponse. Creating post-survey adjustments that do not differentiate between unit nonresponse and breakoff makes an implicit assumption that the two response behaviors have common causes. For example, using demographic characteristics to adjust survey estimates for all sample members who did not provide a complete survey assumes that both nonresponse and breakoff propensities vary across people, and furthermore, this variation is a function of the same set of covariates.

Instead, breakoffs may be driven by survey design characteristics, including features intrinsic to the questionnaire, such as the type of cognitive demands made by particular questions. Respondents are exposed to different demands on each page, reevaluating their cooperation decision throughout the survey and becoming breakoffs at various rates. That is, respondents make multiple cooperation decisions throughout the survey, and those are likely informed by stimuli specific to each question and page. Breakoffs are less likely to be driven by individual response propensities (conditional on the survey protocol) as these propensities are expected to have been exhibited in the initial response behavior. Unlike unit nonrespondents, the individuals who break off have begun the survey and have decided to stop at some point within the survey

instrument. A few survey features that are only seen by those who start the survey have been found to induce breakoff – these include the order of questions and the use of progress indicators (e.g. MacElroy, 2000; Crawford, Couper and Lamias, 2001; Frick, Bächtiger and Reips, 2001; Hogg and Miller, 2003). While many may break off at the very beginning of a survey, yielding the same lack of usable data as unit nonrespondents, the decision processes to start and continue the survey can be affected by very different factors and hence treating unit nonrespondents and breakoffs in the same way may be inappropriate.

It is likely that there is some overlap between the causal mechanisms producing nonresponse and breakoff. For example, prepaid incentives may motivate sampled persons to start the survey, while promised incentives may or may not have the same effect (Berk et al., 1987; Singer, Van Hoewyk and Maher, 2000). However, promised incentives should be more effective for reducing breakoff rates as they require completion of the survey, a hypothesis that remains untested (Bosnjak and Tuten (2003) try to address this question, but with under 15 breakoffs in each condition). Factors affecting response behavior include survey design, questionnaire and question characteristics, and respondent characteristics. Identifying the common and unique links between the factors affecting both nonresponse and breakoff would provide insights into each response behavior, anticipate their effects on survey inference, direct survey design that minimizes the effect of nonresponse and breakoff on survey statistics, and allow better informed post-survey adjustments.



As some of the causes of unit nonresponse are expected to be the same as those of breakoff, the systematic studying of such response behaviors can be facilitated by placing them in the same conceptual framework. Bosnjak and Tuten (2001) present a framework for web surveys that includes unit-nonresponse, breakoff, and item-nonresponse, but explain these behaviors as caused by three orthogonal factors – motivation (to start the survey), opportunity (time and absence of technological problems to finish the survey), and ability (defined as willingness to provide responses to sensitive questions). While some factors are likely unique to certain response behaviors, there must be commonalities in the mechanisms that warrant a unified framework and empirical work that test causal links among the different response behaviors. Furthermore, a framework that is more specific in the types of causes and classifies them in a meaningful way would be more beneficial as it would provide a structure for the systematic examination of well-defined causes.

Groves and Couper (1998) propose a framework for household survey cooperation that classifies different factors affecting the cooperation decision into Environment, Respondent, Survey, and Interviewer. Such a framework encourages studies of nonresponse that do not look at one cause in isolation. While more complete and specific than Bosnjak and Tuten's (2001) framework, Groves and Couper's (1998) framework applies for a single cooperation decision in household surveys. Breakoffs do occur in household surveys, but at rates that do not warrant special attention. However, in web surveys these rates can be substantial as noted earlier, and as breakoffs can occur throughout a survey,

they can be viewed as another cooperation decision that is made multiple times throughout a survey following the initial cooperation.

I propose a framework that builds on previous work by incorporating factors relevant to response behavior in web surveys, multiple cooperation decisions, and multiple response outcomes. My classification has three types of causal factors: Respondent, Survey design, and Question/Page characteristics (Figure 1). Each arrow represents an association between a set of factors and a decision, and again between a decision and a potential outcome. Recalling that nonresponse and breakoff can have common and unique factors, there are paths from both survey design and page/question characteristics to the breakoff decision, yet there is only a path from page/question characteristics to the initial cooperation decision. In addition, these paths vary in magnitude of influence, as for example the announced survey length in the invitation may affect initial cooperation more strongly than it affects subsequent breakoff rates, which in turn may be more affected by any discrepancy between the announced and actual survey length. This level of detail can not be included in the framework, nor is it my intention to do so.

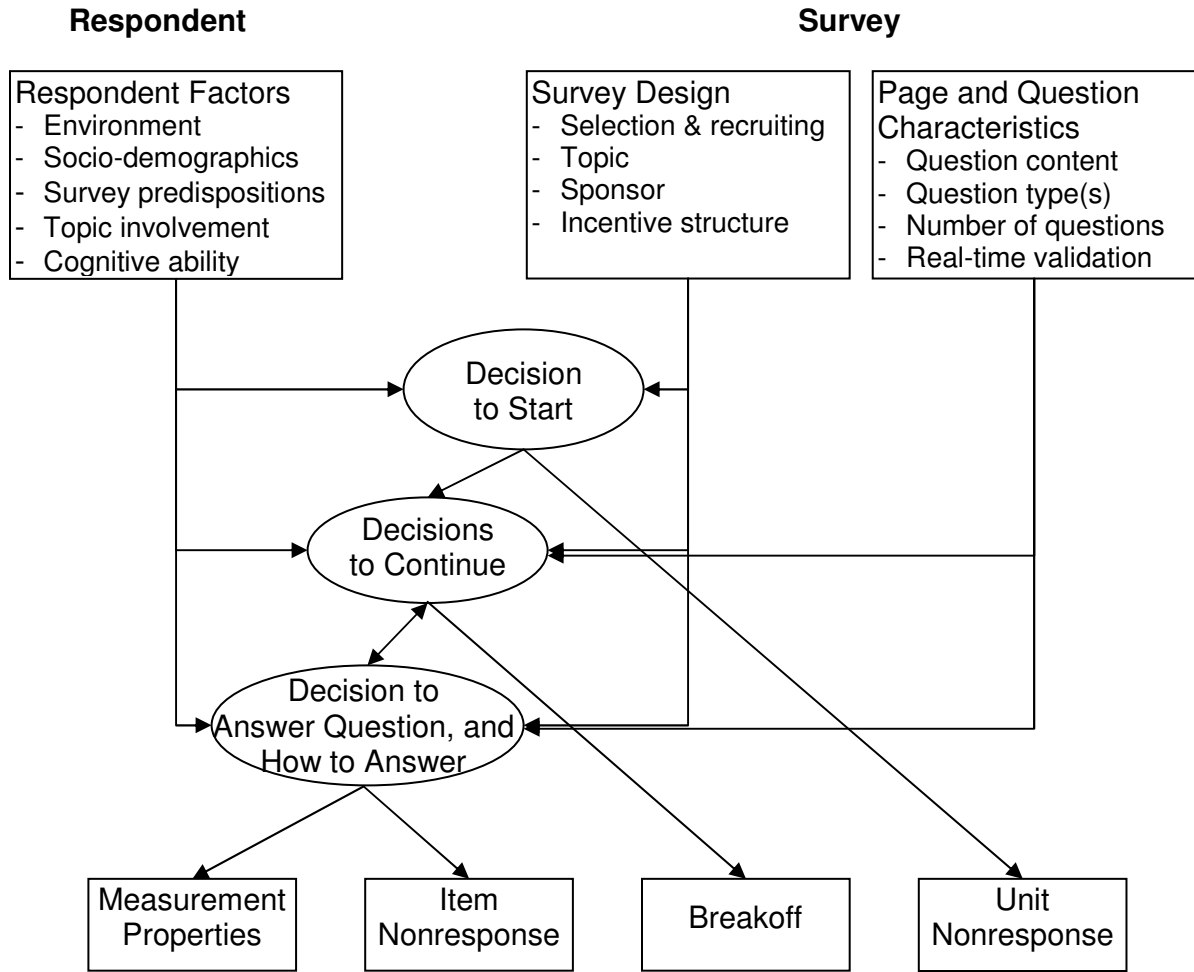


Figure 1. Conceptual Model for Web Survey Participation

This is truly a framework and it provides only a structure for the study of cooperation decisions in web surveys. It does not intend to inform the reader how a particular survey design feature affects a cooperation decision. The latter is often unique to a design feature or a respondent characteristic; attempts to identify such specific mechanisms are made in the next chapters.

This framework helps to classify the types of factors that otherwise may be omitted; it also underscores what is under the researcher's control (for example, factors on the left side of Figure 1 usually are not). By conceptually presenting

the entire process of web survey cooperation, the proposed framework may help identify causal mechanisms.

The first set of factors, respondent characteristics, can be predictive of initial cooperation, although in many cases without a causal explanation. Such factors include gender, age, and race, and can also be specific to a study, such as seniority of students. To the extent to which the association between some of these characteristics and the likelihood of cooperation are fundamentally unchanged by mode, findings from the past decades in other modes can be informative. However, some respondent characteristics can be used as proxies for causes of different response behaviors. Age has been used as a proxy indicator for cognitive ability in survey responding (Herzog and Dielman, 1985; Andrews and Herzog, 1986; Knauper et al., 1997) – people’s ability to perform certain cognitive tasks worsens as they age. Education has been used as a proxy indicator of cognitive sophistication in surveys (Sigelman, 1981; Krosnick and Alwin, 1987; Knauper, 1999). There are respondents who start the survey but have difficulty answering the questions - such respondent characteristics are likely to be associated with breakoff, as addressed in both Chapters 2 and 3. Other respondent factors are much more difficult to study - some are difficult to obtain; others, such as survey predispositions, are difficult to observe regardless of effort. To the extent that respondents are consistent in their response decisions across surveys and such consistency can not be explained sufficiently well with socio-demographic covariates, better attempts at measuring predispositions are needed.

Another set of factors are survey design features. Those have received the most attention as many are under the researcher's control and can be manipulated at relatively low cost in web surveys. Such features include the announced survey length (e.g. Crawford et al., 2001; Trouteaud, 2004), prenotification (e.g. Kaplowitz, Hadlock and Levine, 2004), reminders (e.g. Deutskens et al., 2004; Kaplowitz et al., 2004), content of invitation (e.g. Trouteaud, 2004), timing of reminders (e.g. Crawford et al., 2001), type and amount of incentives (Bosnjak and Tuten, 2003; Deutskens et al., 2004). Other survey design features are known to have an impact on cooperation decisions, often in a manner that introduces systematic bias in the survey estimates – these include the survey topic and sponsor.

There are also questionnaire characteristics that do not affect initial nonresponse decisions as they are not seen prior to starting the survey, but can affect decisions to continue the survey. These include the type of questions asked, such as open-ended (Crawford et al., 2001), the actual length of the survey (e.g. MacElroy, 2000; Hogg and Miller, 2003; Deutskens et al., 2004), the use of grid layout (O'Neil, Penrod and Bornstein, 2003), and even technical difficulties (e.g. Schwarz and Reips, 2001).

Chapter 2 examines the similarities and differences between nonresponse and breakoff as a function of the factors affecting cooperation. While demonstrating that sampled persons who exhibit either of the two response behaviors should be treated separately due to their differences, there is lower predictive ability for breakoff behavior. As noted in the framework, initial

nonresponse to a survey attempt is a single decision with a binary outcome at any given contact, while breakoff stems from decisions that can be reevaluated throughout the entire survey.

The within-survey multiple decision process implies that existing nonresponse theories may not be suitable for breakoff. Since it is the questions that mostly vary within surveys, theories for breakoff could be imported from question-answer process models that are usually used to explain measurement error. For example, as respondents break off on various pages in the survey, questions with particular cognitive demands may elicit higher rates of breakoff. A framework for the processes that respondents undergo in answering a question has been developed and proposed as the survey response process model (Cannell, Marquis and Laurent, 1977; Cannell, Miller and Oksenberg, 1981; Tourangeau, 1984; Strack and Martin, 1987; 1987). The survey response process model breaks down question answering into separate processes: comprehension of the question, retrieval of relevant information, judgment based on the retrieved information, mapping of the response onto the requested response format, and editing of the response for reasons such as social desirability. Questions can vary in the intensity to which each of these processes applies. For example, it is possible that a particular question requires a lot of effort to be comprehended (such as having a large amount of text to be read by the respondent) and may elicit a higher rate of breakoff compared to the rest of the survey. This model is not expected to perfectly predict breakoff rates on each page. There are technical disruptions and other factors that are not under

the researcher's control that can affect breakoff as well. The model, however, can be informative about the specific causes of breakoff that in the most part are under the researcher's control and have the potential of producing bias in the survey estimates. The causes of survey breakoff as a function of different demands made on each page are the primary focus of Chapter 3.

Employing the same causal mechanism used for explaining measurement error to breakoff implies that nonresponse behaviors and measurement error are not independent. Studying this cooperation decision in isolation from measurement error would be incomplete as tradeoffs are likely to exist – some survey design features that decrease breakoff may also increase measurement error. This is illustrated in the framework in Figure 1 as a double-headed arrow between the decision to break off and the decision of whether and how to answer a question. The potential outcomes from these decisions are termination of the survey by the respondent, not providing a response to that particular question, or providing a response with various measurement consequences.

The causal mechanisms affecting breakoff, data quality, and measurement error are not limited to the type of demands of the questions, but also include other page characteristics. A question may be seen as more burdensome if a lot of information needs to be retrieved by the respondent, but even the layout of the questions within the page may make the respondent task more burdensome. Specifically, pages with multiple questions may produce higher rates of breakoff as the size of the task is larger on that particular page.

This design may also affect measurement error, whether through decreased effort, altered context, or some other mechanism.

Choosing the number of questions per page is a design decision based on little empirical evidence. When the questions measure the same construct, inter-item correlations are higher, item-missing data rates are lower, and the set of questions take less time when the questions are presented on the same page (Couper, Traugott and Lamias, 2001). If evaluation criteria like reliability, as commonly estimated through Cronbach's Coefficient  $\alpha$ , are used, placing questions together on the same page would be preferred. As many researchers who use items in scales have criteria tied to Cronbach's  $\alpha$  and believe that values of .7 and above are indicative of a reliable measurement of a construct (e.g., Nunnally and Bernstein, 1994), multiple questions per page designs would be preferred.

However, in the presence of correlated measurement error, such guidance can be entirely misleading. If placing questions together increases correlated measurement error instead of decreasing uncorrelated measurement error, then the higher inter-item correlation and Cronbach's  $\alpha$  would be undesirable. In that case, the lower item-missing data rates and shorter time to answer the questions in that design may be interpreted as lower levels of effort in the question answering task. The reason this has not been addressed is the practical difficulty of separating the correlation between the intended measures, or true scores, from the correlated measurement error; especially in the absence of known true values.



A method that addresses this issue through the use of a criterion measure, by comparing the validity of the identified construct across different layout designs, is proposed and implemented in Chapter 4. If the higher inter-item correlations are indeed the result of higher correlated measurement error rather than higher reliability, lower validity of the measured construct will be found.

Applying a criterion that would allow conclusions about preferable layout designs would help researchers facing the same decision under very similar circumstances. Without understanding the causes of the observed differences, however, the generalizability of these results is limited. Identification of the causes also allows for reduction of measurement error through targeted design changes rather than altering sets of different features. For example, the difference between presenting a single or a few questions per page has multiple design components, including layout features such as grids.

There are different potential causal mechanisms, one or more of which could be producing differences across survey layout designs. Taking advantage of the different design components, the causal mechanism(s) could be addressed experimentally. One likely mechanism is that respondents may perceive the task as more burdensome when faced with numerous questions in a page and exert less effort in answering the questions. Another potential mechanism is respondents' interpretation of the questions as more related to each other when they are presented together, resulting in more similar responses. Yet a third mechanism is related to short-term memory as

respondents may not recall all their responses to preceding similar questions on previous pages. These causal mechanisms are tested in Chapter 5 through a factorial experimental design of four survey features.

The entire proposed framework for web survey participation cannot be tested within the scope of this dissertation. The purpose of this dissertation is to show that breakoff is different from nonresponse; to examine the mechanisms likely to produce breakoff; to demonstrate a method for evaluation of correlated measurement error in the absence of true values, and compare different survey layout designs; and to identify the mechanism through which layout design induces correlated measurement error.

## Chapter 2. Response Behaviors in Web Surveys: Breakoffs Are Different from Unit Nonresponse

### 1. Introduction

Once respondents are contacted, their subsequent decision to participate in the survey is commonly viewed as a binary outcome, their becoming either respondents or nonrespondents for the entire survey. That is, rarely does nonresponse research focus on changes of decisions among those who begin a survey. This is evident in both the theories on survey nonresponse (Groves, Cialdini and Couper, 1992; Groves, Singer and Corning, 2000; Groves, Presser and Dipko, 2004, and the framework by Groves and Couper, 1998) and in the methods for adjusting for nonresponse as they model a binary respondent-nonrespondent outcome (for review of methods see Little, 1982; Kalton and Kasprzyk, 1986).

However, many contacted respondents start the survey and fail to finish it, particularly in web surveys (Musch and Reips, 2000; Lozar-Manfreda and Vehovar, 2002). While many individuals who start the survey and later abandon it (i.e., breakoffs) yield unusable data, such sample persons may be different

from unit nonrespondents to the extent that the mechanisms producing breakoff and unit nonresponse are different.

Unlike nonrespondents, breakoffs started the survey and decided to stop participating only after being exposed to some part of the survey instrument. Potential differences in the causal mechanisms between response and breakoff would warrant the comparisons of the three participatory outcomes in web surveys. Furthermore, differences between breakoff and unit nonresponse carry implications for the measurement, reduction, and adjustment of nonresponse error (to the extent that breakoffs, particularly early in the questionnaire, are subjected to methods for adjustment for unit nonresponse). These implications may prompt the use of additional covariates of each response behavior in surveys, suggest changes to the design of surveys, and use of approaches to post-survey adjustments that treat unit nonresponse and breakoffs separately.

Theories explaining nonresponse in surveys, such as social isolation, social exchange, and altruism (Dillman, 1978; Goyder, 1987; Groves et al., 1992) imply that some individuals are more likely to be unit nonrespondents than others. While nonresponse is expected to have a degree of consistency for individuals across surveys, as well as being related to respondent characteristics, breakoff should not have that consistency as these individuals have made the decision to start the survey and it is likely that features of the particular survey instrument lead a respondent to break off.

This study contrasts unit nonresponse and breakoff in web surveys as a function of respondent characteristics, survey design features, and consistency of each response behavior across different surveys.

## 2. Background and Significance

Measures of survey participation use the ratio of completed surveys that have a sufficient amount of information collected to the total number of contacted eligible sampled elements (CASRO, 1982; AAPOR, 2006). That is, the response outcomes are collapsed into two categories: respondents and nonrespondents. This dichotomous evaluative criterion is also carried over to post-survey adjustments, typically modeling the variability in response outcome with available respondent characteristics (Little, 1982; Kalton and Kasprzyk, 1986).

Nonresponse theories (Groves et al., 1992; Groves et al., 2000; Groves et al., 2004) and conceptual models (Groves and Couper, 1998) also treat the decision to participate as a dichotomous cooperation decision. Groves and Couper (1998) propose a model in which different sets of factors that affect the decision to cooperate in household surveys are classified into environment, respondent, interviewer, and survey design categories. Theories used in survey nonresponse explain effects in this model by proposing how the different factors interact to lead to a respondent's cooperation decision. For example, respondents who can be identified as being less interested in the survey topic are less likely to cooperate (Groves et al., 2000; Groves et al., 2004). Another

line of research has focused on finding covariates of unit nonresponse that may be proxies of the real mechanisms causing survey participation, such as respondent demographic characteristics (Heberlein and Baumgartner, 1978; Goyder, 1982). Survey design features, such as incentives (Singer et al., 2000; Singer, 2002) and survey length (Heberlein and Baumgartner, 1978) have been yet another focus, with motivation and burden as possible explanations.

Meta-analyses have shown substantial rates of breakoff<sup>1</sup> in web surveys, with median rates of 16-34% (Musch and Reips, 2000; Lozar-Manfreda and Vehovar, 2002). Although breakoffs can be perceived as nonrespondents who may have provided some information, the causes producing breakoff may be different from those producing unit nonresponse. The few factors that have been found to induce breakoffs, such as order of the questions and use of progress indicators (e.g. MacElroy, 2000; Crawford et al., 2001; Frick et al., 2001; Hogg and Miller, 2003) are only experienced by those who start the survey. The theories and covariates of unit nonresponse may not apply to breakoffs.

Thus, causes of unit nonresponse may be very different from causes of breakoffs. However, there could also be an overlap between the two response behaviors. Surveys that are announced to take longer to complete have lower response rates (Heberlein and Baumgartner, 1978; Goyder, 1982; Dillman, Sinclair and Clark, 1993), while fewer respondents may reach the end of the survey as survey length is one type of respondent burden (Bradburn, 1978).

Furthermore, there is evidence for an inverse relationship between nonresponse

---

<sup>1</sup> Breakoffs should not be confused with partials; both describe respondents who began but failed to complete the survey, but partials are further defined by having obtained responses to certain (key) survey questions.

and breakoff based on length, as a survey with longer announced length can produce more unit nonresponse but achieve the same number of completed surveys through fewer breakoffs in an identical survey instrument (e.g. Crawford et al., 2001). The length of the survey can act as a common cause of nonresponse and breakoff, by motivating respondents to start the survey but burdening them to break off, for example, yet the degree to which the common causes are dominant in producing unit nonresponse and breakoff remains untested. The extent of commonality and difference between nonresponse and breakoff needs to be established to more effectively reduce each separately or together, and build effective adjustment models that reflect these differences, without inducing unnecessary variance. This latter point is that while separate treatment of nonresponse and breakoff can minimize bias to the extent that they are affected by different causes, treating them separately when this is not necessary can produce less certain and more variable adjustments through smaller sample sizes in multiple models.

There are different ways in which the relationship between the two undesirable response behaviors can be examined, including: (1) the consistency of each behavior across surveys, as well as transitions from one behavior to the other; (2) the association of each with survey design characteristics; and (3) the association of each with respondent characteristics.

There is a certain degree of consistency in people's response behavior. For example, in a small mail survey experiment Brennan and Hoek (1992) found that only 26% of those who refused to complete a survey later cooperated with a

request for a different survey, while 89% of the respondents in the first survey also cooperated in the second survey. Relatively stable attitudes and predispositions towards surveys may produce some of the consistency. However, these attitudes may have weaker effects on breakoffs during the survey, where features within the questionnaire may have a larger effect.

Difficulty in answering questions in the survey could be a potential cause, yet these can not be tested if the sample is restricted in terms of age and education that have been used to proxy cognitive ability and cognitive sophistication.

Another way unit nonresponse and breakoff can be distinguished is through how they are affected by survey features. Incentives reduce unit nonresponse. A mechanism that has been proposed for incentives is that they motivate the respondent to start the survey (e.g., Singer, Groves and Corning, 1999; Singer et al., 2000; Bosnjak and Tuten, 2003; Deutskens et al., 2004), and if they also motivate respondents to persevere to the end of the survey, then incentives should reduce breakoffs too. Different types of incentives may affect each response decision differently; prepaid incentives may have a larger effect on unit nonresponse, while conditional incentives should have a larger effect on breakoff. Respondent interest in the survey topic can affect initial cooperation rates as well (e.g. Goyder, 1987; Groves et al., 2004). If the questions are less interesting than anticipated by those who choose to participate in the survey, lack of topic interest may also increase breakoff rates. But questions may also have differential effect on breakoff through exerting different cognitive demands on the



respondent. Another cause can be attributed to respondent burden, one facet of which is survey length (Bradburn, 1978). When web surveys are introduced as being longer, fewer respondents start the survey (e.g. Crawford et al., 2001; Trouteaud, 2004). However, respondents who start the survey knowing its length should not break off at higher rates. The actual survey length, however, affects the breakoff rate (e.g., MacElroy, 2000; Hogg and Miller, 2003; Deutskens et al., 2004). Other questionnaire characteristics that do not affect initial nonresponse decisions, but can affect decisions on whether to continue the survey include technical difficulties such as the internet connection and software requirements and errors (e.g. Schwarz and Reips, 2001), the type of questions asked, such as open-ended (Crawford et al., 2001), and the use of grid layout (O'Neil et al., 2003).

Apart from the characteristics of the current survey, and the response behavior on previous surveys, nonresponse may be the result of fatigue or lack of interest due to prior exposure to survey requests; i.e., the number of similar surveys to which the respondent has been invited. Presently, interviewer-administered surveys in the U.S. take more effort (measured as call attempts) and achieve lower response rates than a decade ago, suggesting that people may become less willing to respond to new survey requests. This is further supported by ongoing national surveys, as Groves and Couper (1998) found that the declining response rates in the Consumer Expenditure Survey and the National health Interview Survey were driven by increasing refusal rates rather than accessibility. In web surveys, organizations are increasingly concerned with

over-surveying – when members of a survey panel are asked to complete multiple surveys in a relatively short period (Interactive Marketing Research Organization, 2006). There are psychological principles that would explain this phenomenon; for example, scarcity (Cialdini, 1993) when respondents feel that the survey is a unique opportunity, hence participating at higher rates to early survey requests. If an effect of previous survey requests is present, it should not affect breakoffs as such respondents have already started the survey regardless of their past survey taking history.

In summary, we expect that nonresponse and breakoff share common mechanisms to the extent that they are both affected by survey design characteristics influencing motivation, burden, and interest. However, the two response behaviors appear different in that unit nonresponse may be affected by response propensities that can have some consistency across survey requests. This should be evident both at the respondent level, and more generally, by demographic subgroups. In contrast, breakoffs have all started the survey and breakoff is influenced by the content and design of the survey questionnaire and protocol. I propose four hypotheses:

1. Survey design characteristics, such as incentives, affect breakoff in the same manner as they affect nonresponse.
2. Respondent characteristics, such as gender and race/ethnicity, are associated more with nonresponse than with breakoff.

3. Number of previous survey requests is more likely to be associated with lower response propensity, than with the propensity to break off.
4. Nonrespondents tend to be nonrespondents to subsequent surveys, but breakoff propensities vary more across surveys.

### 3. Data and Methods

In March 2001 a Student Life (SL 2001) Survey was conducted at the University of Michigan, inviting 7,000 registered undergraduate students. Half of the students were assigned to a mail survey and the other half to a paging web survey (the survey is presented in multiple pages, with little or no scrolling required). The invitations to the survey came from the principal investigators on the project, sent from the data collection company whose logo also appeared on the survey. Students were promised a coupon for \$20 to a book store upon completion. The main topics covered in the survey were alcohol consumption, tobacco and drug use. The study was repeated in late March 2003 (SL 2003) by the same group of researchers and data collection firm, with much of the same questions and design (Appendix, Figure 2). This time, the sampling frame included the entire university undergraduate population. The incentive was entry in a sweepstakes drawing, with a chance to win cash prizes totaling \$3,000. The survey was repeated yet again in 2005 (SL 2005), on a sample of 7,000 students. In this survey all selected respondents were mailed a

prenotification/invitation and prepaid \$2 incentive and upon completion they had the opportunity to be entered in a sweepstakes drawing for a variety of prizes.

In early February 2003 a Living-Learning Program (LLP 2003) survey of undergraduate students living in residence halls at the University of Michigan was launched as part of a program aiming to involve students in learning outside of the classroom. The survey used sweepstakes drawing for a Palm Pilot and a \$300 bookstore voucher as incentives. The invitation, as in the other surveys, was signed by the Principal Investigator and sent exclusively by e-mail. This survey had an almost identical design to that of the Student Life surveys apart from two distinguishing features (Appendix, Figure 3). The logo was not that of the data collection firm (although much of the design features were identical as the same firm did the SL 2001 and SL 2003), but the University seal was used. The other difference was that the survey used grids – most of the questions were presented in tables of about 5 questions each, where the items were in the rows and the response options were in the columns. The explicit topic of the survey was life on campus and residence halls, and there were a few questions on topics that were covered in the Student Life Surveys, such as alcohol use. The survey was repeated a year later (LLP 2004) using the same design – invitation, signature, incentive, and questionnaire layout.

The respondent identifiers in all five surveys were the unique names (used as computer identification) which the University provides students upon acceptance, and which students use for e-mail and university business. These unique identifiers were used to link the different surveys for this study.

Descriptive analyses first looked at nonresponse and at breakoff behavior by survey, explored bivariate associations of respondent characteristics with each response behavior, compared breakoffs throughout surveys, and presented the consistency of response behavior across the five surveys. In addition, the association between participation in the residential program that was being evaluated by the LLP surveys and each response behavior was examined as a mechanism that had implications for producing (nonresponse or breakoff) bias in survey estimates.

Due to the multivariate nature of modeling response behavior, three models were fit to test the key hypotheses. The data were reorganized so that each record represented the response outcome for a particular respondent to a particular survey. The response behavior outcome was coded as a trichotomy: nonrespondent (did not log on the web survey), breakoff (logged on to the survey but did not reach the end), and complete survey (submitted the page with the last question). There were 41,981 records, of which 20,417; 14,408; 5,808; and 1,348 were for respondents who were invited to one, two, three or four surveys respectively.

Some characteristics for all sampled individuals in all surveys were made available from the registrar's office at the university. These included gender, school standing (Freshman, Sophomore, Junior, or Senior), race/ethnicity (White, Black, Hispanic, Asian, or Other), and age (not used due to limited variability and collinearity with school standing).

Four survey characteristics were coded into indicator variables. As the LLP and SL surveys differed in more than one feature, such as survey topic, specific questions, and question layout in grids, an indicator was included for LLP. Another feature that was likely to affect response behaviors was the announced expected duration of the survey, coding the LLP 2003 survey that stated less than 20 minutes together with the SL 2001 survey that had no declared length. Two other indicators associated with unit nonresponse and also hypothesized to affect breakoff through motivation were incentives: whether the incentives were guaranteed, and also whether they were monetary. An indicator uniquely identifying each survey was not needed as all estimable effects at the survey level were exhausted – knowing the values of all four indicator variables uniquely identified a survey. Subsequently, model parameter estimates for other variables such as respondent characteristics were identical when using the four indicators above and when using an indicator for each survey.

To address the question whether being invited to previous surveys influences the likelihood of a sampled individual to become a nonrespondent, or a breakoff, a variable was created to represent the order of the survey request – whether it was the first, second, third, or fourth survey for which the respondent had been selected and invited.

Lastly, to examine the notion of consistency of a particular response behavior across surveys, as well as transitioning from one particular behavior to another, a variable was created for the response behavior in the previous survey. Since approximately half of the records in this person-survey data set were from

students who were invited to only one survey, this variable held previous response outcome for 12,086 cases (survey requests) from 9,476 students.

The resulting three models were fit as multinomial logistic regressions of breakoff and nonresponse relative to complete as dependent variable. The models accounted for clustering within sampled persons.

#### 4. Results

##### 4.1. Descriptive Analyses

The overall response rates were very different in the three SL studies, varying between 43.4% and 64.8% (treating all breakoffs as incompletes), shown in the first column in Table 1. This variation was consistent with the attributes of the designs, as respondents in the 2001 SL were guaranteed an incentive upon completion of the survey, in the 2003 SL they were offered to be entered for chance to win a prize upon completion, and in 2005 all respondents were sent a prepaid incentive in addition to being entered in a lottery. In addition, respondents in 2003 were informed that the survey would take about 30 minutes to complete, while in 2005 of the announced survey length was 20 minutes. Therefore, these findings warrant further multivariate analyses that take many of these factors into account.

Table 1. Response, Breakoff, Nonresponse, Completion Rates, Type of Incentive, and Informed and Median Time to Complete the Survey by Final Response Disposition

Survey (in Chronological Order)	Complete 1 1/(1+2+3)	Breakoff 2 2/(1+2+3)	Nonrespondent 3 3/(1+2+3)	Total 1+2+3 1+2+3	Completion Rate 1/(1+2)	Guaranteed Incentive	Prepaid Incentive	Informed Time in Minutes	Median Time in Minutes for Completes*
Student Life 2001**	61.6%	2.7%	35.7%	3,500 100%	95.8%	Yes	No	Undisclosed	Unavailable
Living Learning Program 2003	28.7%	10.1%	61.2%	5,040 100%	74.1%	No	No	15-20	32
Student Life 2003	43.4%	4.3%	52.4%	21,294 100%	91.0%	No	No	30	23
Living Learning Program 2004	19.0%	6.89%	74.1%	5,228 100%	73.4%	No	No	20	28
Student Life 2005	64.8%	1.8%	33.4%	6,919 100%	97.4%	Yes	Yes	20	15

\* The mean time was approximately 1.5 minutes longer than the median time for each of the five surveys.

\*\* Half of the cases were excluded (3,500), as they received a mail survey.

The LLP surveys achieved much lower overall response rates and although the two had very similar designs, LLP 2003 achieved a response rate of 28.7%, while the response rate for LLP 2004 was only 19.0%. Like the SL 2003, both surveys used lottery sweepstakes incentives only. However, the advertised topics of the surveys were also different than the SL surveys. To the extent that residential life elicits lower interest in the survey than topics like substance use, it may help explain why smaller proportions attempted the LLP surveys.

Another observation is that higher response rates corresponded to higher completion rates (smaller proportion of breakoffs given initial response to the invitation)<sup>2</sup>. This is consistent with a common mechanism, that if respondents are less motivated to start the survey by survey design features, they can also be less motivated to continue once they started. It is likely that initial response to a

<sup>2</sup> Note that the term “breakoff rates” is not used for the percent of the whole sample who abandoned the survey, but the percent of those who started, i.e. (100 - completion rate). This is based both on a theoretical argument that breakoffs are conditional on initial response (vs. a class of people who are breakoffs and do not start the survey with an intention to finish it), and on limitations of many web surveys where the nonrespondents could not be identified.



survey request is affected by the announced survey length, and that the completion rate is affected by the actual length of the survey. In addition, there is the interaction between the two. Crawford, Couper, and Lamias (Crawford et al., 2001) found that if respondents are informed of a shorter survey length, more begin the survey but higher breakoff rates occur. In the end the same response rate is observed as if longer survey length was announced and fewer respondents started the survey. Unfortunately, this comparison confounds the effect of the initial motivation with the effect of the difference between the announced and the empirical time to complete the survey. The SL surveys achieved lower initial nonresponse rates (36%, 52%, and 33%, Table 1) than the LLP surveys (61% and 74%). The SL surveys also had a shorter actual length (23 and 15 minutes) than announced (30 and 20 minutes), while the LLP surveys took longer to complete (32 and 28 minutes) than announced (15-20 and 20 minutes). Considering that the LLP surveys also had a different layout design by the frequent use of grids, the multivariate analysis will address a few of the likely causes.

Less than three out of four respondents who started the LLP 2003 and 2004 surveys actually completed them, while more than nine out of ten did so in the SL 2003. The different topics could not explain much in the different completion rates – in fact there was nontrivial overlap in question topics. However, the logo at the top of each page starting from the introduction and consent screens was the University seal in the LLP and the graphic of the data collection firm in the SL, which may explain differences in initial breakoff. The

other major difference was the layout of the LLP and SL surveys – the vast majority of the questions in the LLP surveys were placed in grids, starting with the first page of questions. The SL placed far fewer questions in grids and the first screens of the survey did not employ grids. Based on the cause of the differences in breakoff rates, breakoffs would be observed in different parts of the surveys. Lastly, only some of the selected respondents for each survey were also invited to any other of the five surveys. While it is plausible to expect that the sample composition is similar within the SL and within the LLP studies, this same expectation may not hold across SL and LLP if students living in residence halls are different from the rest of the students in their survey response behavior. For example, students in residence halls may be more likely to be in freshman and sophomore standing than the general undergraduate population, and we control for such differences in target populations in multivariate models later in this analysis.

#### *4.1.1. Survey Design Characteristics*

Table 1 shows that there is large variability in the nonresponse and breakoff rates across the different surveys.

In order to measure whether survey design can also affect where in the instrument breakoffs occur, a larger sample size is needed to have sufficient breakoff rates in different parts of the survey, and different survey designs are required. The two surveys that had the largest sample overlap, the LLP 2003

and the SL 2003 (administered only one month apart and the SL 2003 was a complete census of the undergraduate population), provided information on the location within the survey instrument where breakoffs occurred.

Table 2 compares the location of breakoff among respondents who were invited to both surveys. The response and completion rates of these students were very similar to those of the full samples. Yet the completion rates for the same respondents in the two surveys were significantly — higher in the 2003 LLP (74.3%) than in the SL 2003 (90.5%),  $\chi^2_{(1 \text{ d.f.})}=186.01, p<.001$ . The breakoff rates on the introduction and consent screens (where the different logos, academic vs. corporate, were displayed) were relatively similar. However, the breakoff rate in the first three screens was more than twice as high for the LLP as for the SL (2.2% vs. 0.5%, Table 2). This relative difference persisted through the rest of the survey (21.5% vs. 7.1%, Table 2). The section of the survey (Introduction & Consent, first three screens with questions, and the rest of the survey) where respondents broke off in each of the two surveys was significantly different ( $\chi^2_{(2 \text{ d.f.})}=21.43, p<.001$ ). This is consistent with the finding by O'Neil, Penrod, and Bornstein (2003) that placing the questions that share response options in grids increases breakoff rates. However, the results are not consistent with an embedded experiment on grid layout in a single page of a larger web survey (Chapter 3), where differences in breakoff rates associated with a grid format were not detected. While the use of grids is a notable distinction between the two surveys, the instruments also used different questions and survey

protocols. Therefore, this analysis cannot identify specific factors causing the differences in rate and location of breakoff.

Table 2. Nonresponse, Breakoff by Location in Survey, Completes, and Completion Rates for 4,432 Respondents Invited to Both the Living Learning Program 2003 and the Student Life 2003 Surveys

		Non-respondent	Breakoff				Complete
			Introduction	Consent	First Three Screens	Remainder of the Survey	
LLP 2003	Invited	60.7%	0.8%				1,295
	% of total						29.2%
SL 2003	Invited	49.1%	0.7%				2,040
	% of total						46.0%
	% of started	--	2.0%				74.3%
	% of started	--	1.4%	0.3%	0.3%	3.6%	90.5%
	% of started	--	1.4%	0.5%	0.5%	7.1%	90.5%

Note: This table includes only students who were invited to both surveys and were not assigned to the experimental group receiving a scrolling version of SL (n=4,432), as this breakdown of breakoff location was not possible for the scrolling version. Breakoffs on the introduction versus consent screens could not be distinguished in the LLP data.

In order to examine the survey factors affecting breakoff, there should be variability in the question layout and characteristics, and variation in the placement of the different page designs within the questionnaire in order to be able to separate the effect of location in the questionnaire from the effect of a particular design feature. Neither of these occurred in the current studies, but there was some limited evidence for question characteristics affecting breakoff. In the SL 2003 there was a single screen towards the middle of the questionnaire that elicited far more breakoffs than even the first screen. One question elicited 35 breakoffs and the following three screens were grids of questions that cumulated 66 survey abandonments – overall, 10% of all the breakoffs during the questionnaire. This rate of breakoff was much more than what was observed on other screens with grids that came earlier or later in the survey. The questions

were among the more sensitive in the survey, such as: “Been in trouble with the police or University authorities [as a result of being drunk]?” and “Were you sexually assaulted by someone drunk?”

Another observation is that only the first sensitive question elicited a notably high number of breakoffs, with the fourth question in the sequence producing only 3 breakoffs. It could be speculated that different respondents break off for different reasons. Once the characteristic causing a particular set of respondents to break off has been presented and breakoffs occur, subsequent presentation of the stimulus may no longer produce breakoffs, among those remaining as respondents.

Breakoffs throughout the survey did not have a uniform distribution over question sequences or even a smooth temporal function, and seemed to be affected by questionnaire characteristics that warrant further study. As noted above, these surveys were not suitable for examining the reasons for breakoff, yet there is some evidence that higher breakoffs occur on more complex pages with more sensitive questions.

In summary, survey design comprised of features such as topic and layout of the questions, has an effect on the initial cooperation decision and on breakoff during the questionnaire pages with questions. However, survey design does not have an effect on breakoff on the initial pages that do not present survey questions and may show redundant information with the survey invitation, such as topic and sponsor.

#### *4.1.2. Respondent Characteristics*

Respondent demographic characteristics were expected to be more associated with unit nonresponse than with breakoff. Contrary to prior findings in mail surveys (Goyder, 1982)<sup>3</sup> but consistent with findings from interviewer-administered surveys (see Gray et al., 1996), men completed the surveys at significantly lower rates in all five studies (Appendix, Table 8). In terms of breakoff (conditional on starting the survey), only the SL 2003 study showed a significant difference in completion rates between men and women, 89.4% and 92.3%, respectively (Appendix, Table 8). In the other surveys, the gender differences did not have a consistent direction. This descriptive analysis warranted further testing through multivariate models that control for relevant factors affecting response behavior and also account for conducting multiple tests of statistical significance.

Another respondent attribute of possible interest is year in school, with attachment to the institution possibly varying across those in different years. Students in different years also vary in the total number of survey requests they have received from the University, as they receive five to ten survey requests every year. Students in their later years at the University seemed to be more likely to be nonrespondents (Appendix, Table 8). The students in later years in their programs, who did respond, were more likely than students in their earlier years to complete the survey to the end, although the difference in completion

---

<sup>3</sup> The result found by Goyder is based mostly on comparison of response rates to surveys of special populations that happen to be either male or female.

rates was not significantly different in SL 2001 and SL 2005 (Appendix, Table 8). These differences in completion rates were not very large, as the juniors and seniors in SL 2003 for example, had a completion rate of 92% compared to 90% for the freshmen and sophomores. It could not be determined whether higher nonresponse among those in later years of their degree was due to specific factors such as the number of surveys they had been invited to or whether such students simply had less time. Regardless, the key finding from this bivariate examination was that year in school seemed to be positively associated with both initial nonresponse and breakoff.

Similarly, race/ethnicity groups with higher nonresponse rates also had higher breakoff rates among those who started the survey (Appendix, Table 9). Among the largest race groups represented at the University, Blacks had the highest nonresponse rates, followed by Hispanics, Asians, and Whites.

From these bivariate tabulations some respondent characteristics, year in school and race/ethnicity, were associated with both unit nonresponse and breakoff, while gender was associated only with unit nonresponse. These findings confirm the expectation that respondent demographics predict unit nonresponse, but require further multivariate analysis to determine whether demographic characteristics are associated with breakoff.

#### *4.1.3. Response Behavior Across Surveys*

In addition to looking at response behavior between subgroups of sampled persons in each survey, response behavior for individual sampled persons was examined across repeated survey requests. If individual response propensities are largely influenced by stable person-level characteristics, people should exhibit some degree of consistency in response behavior across surveys. Yet once sampled persons exhibit this consistency through either starting the survey or not, their propensity to break off is expected to be mostly driven by survey design features.

Response behavior in pairs of consecutive surveys showed that unit nonrespondents in one survey were much more likely to be unit nonrespondents in another survey than to be complete respondents or breakoffs (Appendix, Table 10). However, as hypothesized, breakoffs did not exhibit this consistency across surveys. For instance, from the 40 respondents who broke off in SL 2001, only 3 broke off also in 2003 with the remainder becoming either complete respondents or nonrespondents, and 4 of the 144 breakoffs in SL 2003 also broke off in SL 2005. Similarly, only 15 of the 144 breakoffs in the LLP 2003 also broke off in the LLP 2004. This breakoff rate was similar to that of the entire sample. Furthermore, the same pattern for people who break off in one survey but are either respondents or nonrespondents in another survey held for initial respondents who broke off at the introduction or the informed consent screens



and for those who abandoned the survey later on (Appendix, Table 11a), ( $p=.343$ ,  $\chi^2$  test).

As expected, nonrespondents in one survey were nonrespondents to other surveys at higher rates. While the degree of consistency is affected by the marginal distribution (i.e., the nonresponse rate to each of the surveys), 1,848 of the 3,005 (61.5%) nonrespondents in the LLP 2003 were also nonrespondents in the SL 2003 (Appendix, Table 11), the two surveys that provided the largest overlap in sample and also had the most similar response rates for the sampled persons common to the two surveys. The response behavior across all seven pairs of surveys that had sufficiently large overlap (1,000 or more students invited to both surveys) were significantly associated ( $p<.001$ ,  $\chi^2$  test).

There was no evidence for transitioning from one behavior to another, which is consistent with the hypothesis that nonrespondents and breakoffs tend not to be a single set of respondents who share a higher propensity to not respond/breakoff. Those who did not break off in 2003 became complete respondents and nonrespondents at rates similar to the overall rates (e.g. 58% became nonrespondents, while the overall nonresponse rate was 53% in Table 10a). These findings held for the other two similar-survey cross classifications, in Table 10b and c, including the LLP surveys.

In order to quantify the consistency of respondents' behavior across surveys, kappa statistics were computed for all cross-classifications. Despite the consistency found for nonresponse, there was a high degree of switching behaviors between the surveys (Appendix, Table 12). Interestingly, pairs of

different surveys have about the same consistency in respondent behavior as pairs of similar surveys – the kappa statistics for SL 2003 with LLP 2003 and with SL2005 were 0.274 and 0.276, and for LLP 2004 with LLP 2003 and SL 2003 were 0.201 and 0.213. There was evidence for two effects: the pairs of surveys in the same year had higher consistency in respondent behaviors, and this applied for pairs where both surveys had relatively high response rates. That is, two surveys that are temporally close and have low nonresponse rates, may be likely to have respondents who are completers in both or nonrespondents who are nonrespondents in both.

Another way to look at variability in nonresponse and breakoff is to examine survey response outcome by response outcome in the previous surveys to which respondents were invited, pooling this information across all five surveys. Table 3 shows that about half of those who were nonrespondents in a previous survey, were also nonrespondents in the next survey to which they were invited (51.7%). This was not the case for breakoffs where that percentage was only 4.4% ( $p < .001$ ,  $\chi^2$  test). However, this descriptive analysis did not take into account the very different response rates across the five surveys, which will be revisited in the multivariate models.

Table 3. Current Survey Response Outcomes by Previous Survey Outcomes

Previous Survey Disposition	Current Survey Disposition			Total
	Complete	Breakoff	Nonrespondent	
Complete				5708
% within Complete	49.8%	4.3%	45.9%	100%
% of Total	23.5%	2.0%	21.7%	47.2%
Breakoff				473
% within Breakoff	46.7%	4.4%	48.8%	100%
% of Total	1.8%	0.2%	1.9%	3.9%
Nonrespondent				5905
% within Nonrespondent	44.8%	3.5%	51.7%	100%
% of Total	21.9%	1.7%	25.3%	48.9%
Total				12086
% of Total	47.2%	3.9%	48.9%	100%

Significant at  $\alpha=.001$ ,  $\chi^2$  test with 4 d.f.

#### 4.1.4. Effect of Program Participation on Response Behavior

The effect of topic involvement and program participation on response behavior to surveys evaluating a program is of particular interest due to the potential biasing effect on survey statistics that are related to the topic or program. Response dispositions for those who participated in the Living Learning Program (with faculty members living in the student resident halls) were compared to those who did not participate in the program. Approximately one third of the respondents in the Living-Learning Program surveys participated in residence hall programs within their dormitories. Program participation should affect nonresponse in the LLP surveys relative to the control condition, as the purpose and topic of the LLP surveys was evaluation of the program.

Breakoff could be similarly affected as respondents who are more/less motivated through having participated in the program would breakoff at lower/higher rates. A counter-hypothesis is that those who participated in the

program would exhibit any lower motivation resulting from the topic of the survey by not even starting it, as they have been made aware of the topic in the invitation. Under this hypothesis, the breakoff rates for program participants and nonparticipants, conditional on having started the survey, would not be affected by whether the survey is on the topic of the program. Since topic is the underlying mechanism affecting nonresponse and breakoff rates in this comparison, this difference between program participants and nonparticipants is not expected to exist in a survey of a different topic, like the SL 2003.

Program participation had a significant effect on nonresponse rates in both surveys evaluating the program, LLP 2003 and LLP 2004, and no effect in the survey on a different topic, SL 2003 (Table 4). In the LLP surveys the program participants had unit nonresponse rates that were 4 percentage points higher than the nonresponse rates for the program nonparticipants (54% vs. 50%,  $p < .05$  and 76% vs. 72%,  $p < .05$ ), while there was no difference in the SL 2003 survey (45% vs. 44%,  $p = .981$ ). In this case, program participation created a negative affect toward a survey on the same topic and increased nonresponse and breakoff rates for those in the program. While this seems contrary to the forwarded hypothesis for program participants to be nonrespondents at higher rates, it is very possible that the participants do not like the program or at least do not like discussing it.<sup>4</sup>

As expected, the completion rate among program participants and nonparticipants was not different in the SL 2003 (90% vs. 91%). Turning to the LLP surveys, the results were inconclusive due to the relatively small number of

---

<sup>4</sup> The substantive variables are not available in order to test this explanation.

breakoffs. While completion rates were higher for nonparticipants in the program in LLP 2003 and LLP 2004 (71% vs. 81% and 64% vs. 71%), this difference was statistically significant only in the LLP 2003.

Table 4. Response Dispositions by Living Learning Program Participation across the Three Surveys

	LLP 2003		SL 2003		LLP 2004	
	Residential Program	Non-Participants	Residential Program	Non-Participants	Residential Program	Non-Participants
1.Completes	33%	40% *	50%	50%	15%	20% *
2.Breakoffs	14	10	5	5	9	8
3.Nonrespondents	54	50	45	44	76	72
Total	573 100%	605 100%	573 100%	605 100%	573 100%	605 100%
Completion Rate 1/(1+2)	71%	81% *	91%	90%	64%	71%

\* Significant at  $\alpha=.05$ ,  $\chi^2$  test.

Program participation was associated with higher unit nonresponse in the surveys evaluating the program, and there was limited evidence that breakoff rates (conditional on starting the survey) were also higher. This finding is based on a relatively small number of breakoffs and requires replication due to the implications for bias in statistics related to the survey topic.

#### 4.2. Multivariate Analyses

The hypotheses I set out to test were on the association of survey design features, respondent characteristics, number of previous survey invitations, and the consistency of response behavior for nonresponse and for breakoff.

The descriptive analyses indicated that unit nonresponse and breakoff may share causes, as surveys with higher nonresponse rates also obtained lower completion rates. This was true even when constraining analysis to the same set of sampled persons. However, breakoff was associated with only some of the sample person characteristics that were related to nonresponse. Response behavior across surveys was relatively consistent for nonrespondents, but not for breakoffs. The surveys varied in their response rates, survey design characteristics, respondent characteristics, and had a fixed sequential order. In order to measure the marginal impact for the respondent and survey characteristics three multivariate models were fit.

The two data sets needed for these analyses and described earlier, contained separate records for each sampled person, for each survey. The first contained 41,981 records from all five surveys from 29,894 individuals. The second data set was needed for the consistency and transitioning of response behavior across surveys, and therefore excluded respondents with only one survey and excluded data from the first survey for the remainder of the respondents, yielding 12,086 observations from 9,476 individuals.

#### *4.2.1. Survey and Respondent Characteristics*

The first multinomial logistic regression predicts the probability of being a breakoff and the probability of being a nonrespondent relative to being a complete respondent as a function of the respondent characteristics and survey

design features. The respondent characteristics such as gender, race, and year in school were discussed in the bivariate analyses. The four predictors measuring survey design features served a dual purpose: providing parameter estimates for causes of breakoff and nonresponse, and accounting for marginal differences in response outcomes for each of the five surveys. The survey design features were guaranteed incentive, prepaid incentive, differences in topic and questionnaire design (Living Learning Program surveys indicator), and whether the survey was announced as less than 20 minutes to complete.<sup>5</sup>

If breakoffs are driven mostly by survey design features (as individuals are willing to start the survey but later break off), while nonresponse is driven by the survey design and characteristics of the sample persons that are associated with an initial survey cooperation decision, this should become apparent from the odds ratios in the multinomial regression model.

The results were consistent with the above hypothesis. The four survey design features had a statistically significant (Type III analysis Wald  $\chi^2$ ,  $p < .001$ ) and similar impact on the likelihood of breakoff and on the likelihood of nonresponse for three of the factors. Providing a guaranteed incentive lowered the odds of breakoff by .446 and the odds of nonresponse by .482, relative to being a complete respondent (Table 5). Similarly, prepaid incentives significantly decreased the odds of both breakoff and nonresponse. Sampled persons in the Living Learning Program surveys (that differed from the Survey Life survey on multiple characteristics, including design such as the frequent placement of

---

<sup>5</sup> The actual length of the survey could not be used as it highly collinear with the announced length and the LLP indicator, apart from the lack of degrees of freedom.

questions in grids, the topic and purpose of the survey, and unmeasured differences in sample characteristics) were much more likely to break off and not respond (odds ratios of 3.471 and 3.503, respectively). Informing sampled individuals that the survey would take less than 20 minutes to complete reduced the odds of nonresponse by .544. Not surprisingly, those who started the survey having been exposed to the expected time to complete the survey were no more or less likely to break off.

Table 5. Multinomial Regression of Breakoff and Nonrespondent Outcomes Relative to Complete Respondent in SL 2001, LLP 2003, SL 2003, LLP 2004, SL 2005 on Respondent Characteristics, Survey Characteristics, and Order of Survey (41,981 observations from 29,894 individuals selected for at least one survey)

Predictor	Breakoff		Nonresponse		Type III Analysis p-value
	Odds Ratios	Parameter p-value	Odds Ratios	Parameter p-value	
Lottery Incentive, SL Surveys, 20 Minutes or more	1.000	--	1.000	--	<.001
Guaranteed Incentive (1=yes)	0.436	<.001	0.481	<.001	
Prepaid Incentive (1=yes)	0.554	<.001	0.804	<.001	
Living Learning Program Surveys (1=yes)	3.145	<.001	3.395	<.001	
Less than 20 Minutes (1=yes)	1.023	0.800	0.558	<.001	
First Survey	1.000	--	1.000	--	0.111
Second Survey	0.933	0.307	1.000	0.991	
Third Survey	1.292	0.040	1.112	0.069	
Fourth Survey	1.576	0.229	1.230	0.091	
Female	1.000	--	1.000	--	<.001
Male	1.223	<.001	1.583	<.001	
White	1.000	--	1.000	--	<.001
Black	1.506	<.001	1.900	<.001	
Hispanic	1.173	0.154	1.340	<.001	
Asian	1.236	0.004	1.295	<.001	
Other	1.195	0.025	1.127	0.003	
Freshman	1.000	--	1.000	--	
Sophomore	1.057	0.405	0.969	0.328	
Junior	0.861	0.060	0.982	0.596	
Senior	0.794	0.005	1.074	0.032	

Likelihood Ratio  $\chi^2$  (30 d.f.)= 4611.20, p<.001, Pseudo R<sup>2</sup>=.104



Although the sample persons' characteristics were associated with both breakoff and nonresponse, the association was larger for nonresponse. Gender was significantly associated with response outcome, and when contrasted, the odds ratio of males to break off (1.218) was significantly lower than for males to be nonrespondents (1.579),  $p < .001$ . Race and ethnicity were significant, although the parameter estimates again tended to be larger for nonrespondents and the odds ratio for Hispanics to break off was not significantly different from 1. Year in school was also a significant predictor of response behavior. Unlike in the bivariate analysis, when controlling for the other variables in the model only students in their Senior year had significantly different propensities to break off and not to respond. Senior students had significantly lower odds of breaking off (.832) and higher odds of being nonrespondents (1.090).

Findings were not definitive but largely consistent with expectations. Survey design features had a significant effect on the odds to break off and the odds to be nonrespondent and that effect was similar for both response behaviors with the exception for informed survey length. Respondent characteristics on the other hand were more strongly associated with nonresponse than with breakoff.

#### *4.2.2. Survey Order: Changing Respondent Predispositions toward Surveys*

Researchers have been concerned with over-surveying target populations, potentially causing higher nonresponse rates and especially so in university student populations where e-mail addresses can often be obtained for all students (e.g., <http://www.ccs.uottawa.ca/webmaster/survey/best-practices.html>). This fear has not been supported by empirical evidence. If such an effect is found, it has implications for sampling (optimizing samples for multiple surveys), data collection procedures (e.g., number of reminders), survey design (including measures for predispositions and prior survey experience), and post-survey adjustments that may incorporate measures for the number of surveys respondents have been exposed to. There is a mixture of possible factors that can be manipulated – for example, undermining the scarcity principle that is commonly used in invitation and persuasion letters, and considering the notion of survey fatigue.

The data from the five surveys can provide a limited test of the above hypothesis, utilizing the large overlap in sampled students. Particularly, as much of this overlap is between surveys that are temporally close, the effect of being invited to other surveys on response propensity can be observed in the response behavior in a subsequent survey.

The model in Table 5 presents the order of the survey, i.e., whether a particular response outcome is from the sampled student's first, second, third, or

fourth survey from the five surveys in this study.<sup>6</sup> Survey order was not statistically significant ( $p=.111$ ) with an indication of later survey requests being associated with higher odds of breakoff and nonresponse. As students who received requests for three or four of the five surveys were mostly in their Junior and Senior years in school, the analysis did control for year in school.

Some coefficients could be affected by the structure of the combined data, as for example the five surveys have different sample sizes – since the SL 2003 is a census and is the largest, it would also be the survey with the largest share of respondents' first survey. No person could be sampled for LLP 2004 before SL 2003. Such confounds could not be fully addressed by the multivariate model, but can be partially addressed by using only a subset of the data.

Therefore another test of whether inviting respondents to participate in one survey decreases the probability of participation in another survey shortly after, is to use a smaller but identical set of respondents. The two LLP surveys were used to identify a larger pool of students who were eligible to be selected but only part of them were invited to the LLP 2003. About a month after LLP 2003, students were invited to the SL 2003 survey. Among those in SL 2003 who were eligible for the LLP surveys, those who were invited (and sent reminders) to participate in the LLP 2003 survey the previous month failed to respond at significantly higher rates than those who were not invited to the LLP 2003 (44.6% vs. 38.7%, respectively, Table 6). However, the completion rates were not significantly different ( $p=.161$ ). Despite the small number of breakoffs, this is limited evidence that earlier survey requests only affect nonresponse.

---

<sup>6</sup> No students were invited to all 5 surveys.

Table 6. Response Outcome for the SL 2003 Survey by Whether LLP Eligible Sample Persons were Invited to LLP 2003 a Month Earlier

		Student Life Survey 2003 (April)			Total
		Completes	Breakoffs	Nonrespondents	
Living Learning Program Survey 2003 (March)	Not Invited	267 57.1%	20 4.3%	181 38.7%	468 100%
	Invited	589 50.0%	64 5.4%	526 44.6%	1179 100%

\* Significant at  $\alpha=.05$ ,  $\chi^2$  test.

The combination of the multivariate analysis and the more targeted comparison of response behavior by whether the student was invited to another survey a month earlier indicated that there might be an adverse effect on nonresponse from multiple survey requests, but it was not very clear whether this also affected breakoff.

#### 4.2.3. Previous Response Behavior

Breakoff and nonresponse are expected to be different through the consistency of each behavior across surveys. Under the proposed hypothesis, nonrespondents will tend to be nonrespondents to other survey requests, while breakoffs will be just as likely to be completes and nonrespondents as they are likely to be breakoffs. Testing this in a multivariate model allows controlling for differences between surveys and differences between respondents.

Table 7 presents the unadjusted odds ratios (not controlling for the other variables that are in the model) and the odds ratios under the model. The two sets of odds ratios are quite similar, and also similar to those in the model in

Table 5. Nonrespondents in the previous survey had odds of 1.255 being nonrespondents in the following survey than of being completes. However, breakoffs in the previous survey request were no more likely to be breakoffs again in the current survey.

Table 7. Multinomial Regression of Breakoff and Nonrespondent Outcomes Relative to Complete Respondent in LLP 2003, SL 2003, LLP 2004, SL 2005 on Respondent Characteristics, Survey Characteristics, and Response Outcome in Previous Survey (12,086 observations from 9,476 individuals selected for at least two surveys)

Predictor	Bivariate		Model				Type III Analysis p-value
	Breakoff Odds Ratios	Non-response Odds Ratios	Breakoff Odds Ratios	Parameter p-value	Nonresponse Odds Ratios	Parameter p-value	
Lottery Incentive, SL Surveys, 20 Minutes or more	1.000	1.000	1.000	--	1.000	--	<.001
Guaranteed Prepaid Incentive (1=yes)	0.202	0.430	0.332	<.001	0.520	<.001	
Living Learning Program Surveys (1=yes)	6.913	4.566	4.852	<.001	4.046	<.001	
Less than 20 Minutes (1=yes)	2.596	1.938	0.595	0.433	0.418	0.011	
Complete in Previous Survey	1.000	1.000	1.000	--	1.000	--	<.001
Breakoff in Previous Survey	1.097	1.132	1.018	0.942	1.045	0.664	
Nonrespondent in Previous Survey	0.898	1.249	0.929	0.458	1.255	<.001	
Female	1.000	1.000	1.000	--	1.000	--	<.001
Male	1.090	1.467	1.348	0.003	1.671	<.001	
White	1.000	1.000	1.000	--	1.000	--	<.001
Black	1.269	1.734	1.488	0.012	2.016	<.001	
Hispanic	0.681	1.091	1.001	0.997	1.380	<.001	
Asian	1.126	1.331	1.228	0.167	1.380	<.001	
Other	1.423	1.338	1.198	0.267	1.173	0.036	
Freshman	1.000	1.000	1.000	--	1.000	--	
Sophomore	0.897	0.900	0.748	0.035	0.782	<.001	
Junior	0.501	0.682	0.648	0.006	0.781	<.001	
Senior	0.265	0.611	0.425	<.001	0.801	0.001	

Model 3: Likelihood Ratio  $\chi^2$  (26 d.f.)= 1387.16, p<.001, Pseudo R<sup>2</sup>=.108

No evidence for transitioning was present, as nonrespondents were not more likely to be breakoffs in the next survey, and breakoffs were not more likely to become nonrespondents.

The degree of consistency in nonresponse behavior across surveys means that some sample persons are more likely to be nonrespondents, while the lack of consistency in breakoff behavior means that different people are equally likely to be breakoffs, and it must be factors like survey design and circumstances that lead to such response behavior. The lack of transitioning between nonresponse and breakoff further supports the idea that treating breakoffs as part of nonrespondents erroneously assumes that they come from the same pool of respondents with low propensities to respond.

## 5. Discussion and Conclusions

Breakoff and nonresponse did show to be similar through some survey design features that could affect individuals' motivation, less similar in terms of sample person characteristics, and completely different in terms of behavior consistency across surveys.

We found support for the first hypothesis that we tested, that survey design features would affect both nonresponse and breakoff in a similar manner as they affect respondent decisions to cooperate both prior to the survey and during the survey. Offering guaranteed incentives (instead of only prize drawings) and prepaid incentives decreased both the likelihood to be a nonrespondent and the likelihood to be a breakoff. The set of sample, survey design, and questionnaire differences between the Student Life and the Living Learning Program surveys that could not be disentangled with these data also

showed significant differences. The LLP surveys were on a residential program evaluation topic, placed most of the questions in grids, and were constrained to students who live on campus. These surveys achieved higher nonresponse and higher breakoff rates even after controlling for some of the survey design features and student characteristics that were available. The announced expected time to complete the survey did have a significant effect only on initial nonresponse and no effect on breakoff. This is the one survey design feature that can be expected to affect mostly the decision to start the survey. That is, surveys that announced 20 minutes or more needed to complete the instrument yielded higher unit nonresponse rates but no different breakoff rates compared to the surveys that announced a shorter survey length or did not announce length at all.

Complete support for the second hypothesis that only nonresponse would be associated with person characteristics was not found. While both unit nonresponse and breakoff were associated with person characteristics, this association was stronger and more prevalent for unit nonresponse. Males were more likely to break off than females, yet males were also significantly more likely to be nonrespondents than breakoffs. While all non-white race/ethnicity groups were more likely to be nonrespondents than completes compared to non-hispanic whites, this was not the same for breakoffs as Hispanics were not more likely to be breakoffs than completes. Breakoffs were most different from nonrespondents by year in school, which among other things can be viewed as a

proxy for survey experience - students in their Senior year were less likely to be breakoffs but more likely to be nonrespondents.

Concerns about the effect of “over-surveying” or at least the effect of previous survey requests on survey cooperation were not supported in the multivariate models. The test that looked at response behavior in the common sample between two consecutive surveys showed that those who were invited to a survey in the previous month were more likely to be nonrespondents than those who were not invited, but they were not more likely to be breakoffs.

Strong support was found for the key fourth hypothesis, that unit nonresponse behavior is more consistent across surveys than breakoffs. Nonrespondents were 1.255 times more likely to be nonrespondents to the next survey request. This consistency was not found for breakoffs where this odds ratio was not significantly different from 1, implying that this response behavior was predominantly driven by survey design and circumstantial factors, such as incentives, question difficulty, events during the completion of the survey, etc.

Looking at breakoff rates for different parts of the surveys in this study we found that abandonments did not occur exclusively in the very beginning. Furthermore, the introduction and consent screens resulted in similar proportion of respondents breaking off in the different surveys. Rates did differ throughout the different surveys. The survey that had many screens with multiple questions in grid layout evoked more breakoffs, albeit confounded with other design differences.




While the common methods for decreasing both nonresponse and breakoff have an inherent efficiency, it is the differences between the two response behaviors that may have the greatest implications for survey practice. The lack of consistency in breakoff behavior and the lower predictive ability of respondent characteristics of this response behavior may be erroneously interpreted as a lack of breakoff bias, that breakoffs are like complete respondents. This could not be tested here and needs future research. The exact mechanism producing breakoffs needs to be identified, so that bias in survey statistics as a result of breakoffs can be anticipated, reduced, and adjusted.

There is certainly reason to be concerned about breakoff bias. Topic involvement, indicated by participation in a program, had a significant effect on initial nonresponse and breakoff. However, program participants had higher initial nonresponse rates and lower completion rates. This could be associated with the effectiveness or other features of the program, as the survey on a different topic found no differences in completion or nonresponse by involvement group, as expected.

This study is a necessary first step in understanding breakoffs as a survey response behavior, that despite anticipated commonalities is very different from nonresponse. However, it needs to be explored in other populations and in other modes of data collection. The same level of attention that has been devoted to studying the causal mechanisms of nonresponse needs to be given to breakoffs separately.

## 6. Appendix



Questions about this survey?  
 Email us at [umsl@msiresearch.com](mailto:umsl@msiresearch.com)  
 or call toll free 1.866.674.3375

Consent

About You

Tobacco & Alcohol

Other Drugs

Perceptions

Sex & Driving


**Where do you live during the school year?**

- Residence hall
- Fraternity/sorority
- Family housing
- Co-op (ICC)
- House/apartment in Ann Arbor
- Outside Ann Arbor
- Other
- Refused

---

Next Screen

Figure 2. Student Life Surveys.



Questions about this survey?  
 Email us at [housingwi@msiresearch.com](mailto:housingwi@msiresearch.com)  
 or call toll free 1.866.674.3375

Before College

College

Campus Life

About You

Living and Learning

**Your Perceptions Before Enrolling In College**

**Thinking back to before you started college, what activities did you think were going to be very important to you during college?**

*(Click one response in each row)*

	Not at all important	Somewhat important	Important	Very important
Gaining a broad education and appreciation of ideas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting to know your instructors outside of class	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Discussing ideas and intellectual topics with friends and other students	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Being active in extra-curricular activities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Being a top student academically	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

---

Next Screen

Previous Screen

Figure 3. Living-Learning Program Surveys.

Table 8. Complete, Breakoff, and Nonrespondent Outcome in the Student Life and Living Learning Surveys by Gender and Year in School of Sampled Individuals.

Survey	Disposition	Gender		Year in School			
		Male	Female	Freshman	Sophomore	Junior	Senior
SL2001	1.Complete	943 55.3%	1,204 * 67.5%	372 57.3%	511 64.0%	590 63.0%	674 61.0%
	2.Breakoff	49 2.9%	46 2.6%	19 2.9%	26 3.3%	22 2.4%	28 2.5%
	3.Nonrespondent	713 41.8%	534 29.9%	258 39.8%	261 32.7%	325 34.7%	403 36.5%
	Completion Rate 1/(1+2)	95.0%	96.3%	95.1%	95.2%	96.4%	96.0%
SL2003	1.Complete	3,965 37.8%	5,237 * 48.8%	1,464 43.6%	2,239 44.2%	2,427 44.1%	3,072 * 42.0%
	2.Breakoff	469 4.5%	435 4.1%	158 4.7%	255 5.0%	220 4.0%	271 3.7%
	3.Nonrespondent	6,056 57.7%	5,067 47.2%	1,732 51.6%	2,567 50.7%	2,851 51.9%	3,973 54.3%
	Completion Rate 1/(1+2)	89.4%	92.3% *	90.3%	89.8%	91.7%	91.9% *
SL2005	1.Complete	2,033 60.1%	2,451 * 69.3%	1,249 71.3%	1,082 62.7%	1,028 62.2%	1,125 * 62.9%
	2.Breakoff	50 1.5%	71 2.0%	34 1.9%	33 1.9%	34 2.1%	20 1.1%
	3.Nonrespondent	1,298 38.4%	1,016 28.7%	469 26.8%	610 35.4%	592 35.8%	643 36.0%
	Completion Rate 1/(1+2)	97.6%	97.2%	97.4%	97.0%	96.8%	98.3%
LLP2003	1.Complete	529 26.3%	918 * 30.4%	519 25.4%	660 30.4%	227 32.9%	41 * 33.3%
	2.Breakoff	164 8.2%	342 11.3%	213 10.4%	220 10.1%	60 8.7%	13 10.6%
	3.Nonrespondent	1,316 65.5%	1,759 58.3%	1,311 64.2%	1,293 59.5%	402 58.3%	69 56.1%
	Completion Rate 1/(1+2)	76.3%	72.9%	70.9%	75.0%	79.1%	75.9% *
LLP2004	1.Complete	313 14.9%	681 * 21.8%	642 19.0%	259 18.3%	66 18.6%	27 * 33.3%
	2.Breakoff	130 6.2%	230 7.4%	208 6.2%	121 8.5%	29 8.2%	2 2.5%
	3.Nonrespondent	1,657 78.9%	2,217 70.9%	2,526 74.8%	1,037 73.2%	259 73.2%	52 64.2%
	Completion Rate 1/(1+2)	70.7%	74.8%	75.5%	68.2%	69.5%	93.1% *

\* Significant at  $\alpha=.05$ ,  $\chi^2$  test.

Table 9. Complete, Breakoff, and Nonrespondent Outcome in the Student Life and Living Learning Surveys by Race of Sampled Individuals.

Survey	Disposition	None	White	Black	Hisp.	Asian	Native Am.	Not Incl.	Total
SLS2001	1.Complete	213 60.0%	1,455 63.3%	106 42.1%	75 58.1%	273 65.8%	12 63.2%	13 65.0%	2,147 * 61.5%
	2.Breakoff	11 3.1%	67 2.9%	7 2.8%	3 2.3%	6 1.4%	0 0.0%	1 5.0%	95 2.7%
	3.Nonrespondent	131 36.9%	777 33.8%	139 55.2%	51 39.5%	136 32.8%	7 36.8%	6 30.0%	1,247 35.7%
	Completion Rate 1/(1+2)	95.1%	95.6%	93.8%	96.2%	97.8%	100.0%	92.9%	95.8% *
SLS2003	1.Complete	635 40.2%	6,337 46.6%	562 32.2%	386 39.6%	1,133 38.3%	72 41.9%	77 41.2%	9,202 * 43.3%
	2.Breakoff	78 4.9%	550 4.0%	71 4.1%	45 4.6%	134 4.5%	7 4.1%	19 10.2%	904 4.3%
	3.Nonrespondent	865 54.8%	6,726 49.4%	1,114 63.8%	544 55.8%	1,690 57.2%	93 54.1%	91 48.7%	11,123 52.4%
	Completion Rate 1/(1+2)	89.1%	92.0%	88.8%	89.6%	89.4%	91.1%	80.2%	91.1% *
SLS2005	1.Complete	294 64.2%	2,411 68.5%	568 56.2%	550 61.3%	593 63.0%	26 74.3%	42 72.4%	4,484 * 64.8%
	2.Breakoff	13 2.8%	52 1.5%	21 2.1%	16 1.8%	19 2.0%	0 0.0%	0 0.0%	121 1.7%
	3.Nonrespondent	151 33.0%	1,056 30.0%	421 41.7%	331 36.9%	330 35.0%	9 25.7%	16 27.6%	2,314 33.4%
	Completion Rate 1/(1+2)	95.8%	97.9%	96.4%	97.2%	96.9%	100.0%	100.0%	97.4% *
LLP2003	1.Complete	128 29.6%	941 30.6%	110 23.8%	65 25.4%	176 25.7%	9 15.8%	18 29.0%	1,447 * 28.8%
	2.Breakoff	40 9.3%	316 10.3%	52 11.2%	21 8.2%	63 9.2%	8 14.0%	6 9.7%	506 10.1%
	3.Nonrespondent	264 61.1%	1,817 59.1%	301 65.0%	170 66.4%	445 65.1%	40 70.2%	38 61.3%	3,075 61.2%
	Completion Rate 1/(1+2)	76.2%	74.9%	67.9%	75.6%	73.6%	52.9%	75.0%	74.1% *
LLP2004	1.Complete	123 22.9%	635 20.0%	60 12.3%	31 14.4%	119 17.1%	10 20.8%	16 23.2%	994 * 19.0%
	2.Breakoff	28 5.2%	225 7.1%	31 6.3%	14 6.5%	53 7.6%	1 2.1%	8 11.6%	360 6.9%
	3.Nonrespondent	385 71.8%	2,316 72.9%	398 81.4%	171 79.2%	522 75.2%	37 77.1%	45 65.2%	3,874 74.1%
	Completion Rate 1/(1+2)	81.5%	73.8%	65.9%	68.9%	69.2%	90.9%	66.7%	73.4% *

\* Significant at  $\alpha=.05$ ,  $\chi^2$  test.

Table 10. Response Behavior across Similar Surveys.

Table 10.a. Response Behavior in SL 2003 by Previous Behavior in SL 2001.

		SL 2003			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
SL 2001	Completes	536 56.9%	26 2.8%	380 40.3%	942 100.0%	63.2%
	Breakoffs	14 35.0%	3 7.5%	23 57.5%	40 100.0%	2.7%
	Nonrespondents	105 20.7%	13 2.6%	390 76.8%	508 100.0%	34.1%
	Total	655 44.0%	42 2.8%	793 53.2%	1490 100.0%	100.0%

p<.001,  $\chi^2$  test

Table 10.b. Response Behavior in SL 2005 by Previous Behavior in SL 2003.

		SL 2005			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
SL 2003	Completes	1,004 77.6%	14 1.1%	275 21.3%	1293 100.0%	44.7%
	Breakoffs	89 61.8%	4 2.8%	51 35.4%	144 100.0%	5.0%
	Nonrespondents	671 46.1%	27 1.9%	759 52.1%	1457 100.0%	50.3%
	Total	1764 61.0%	45 1.6%	1085 37.5%	2894 100.0%	100.0%

p<.001,  $\chi^2$  test

Table 10.c. Response Behavior in LLP 2004 by Previous Behavior in LLP 2003.

		LLP 2004			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
LLP 2003	Completes	137 31.2%	48 10.9%	254 57.9%	439 100.0%	36.3%
	Breakoffs	21 14.6%	15 10.4%	108 75.0%	144 100.0%	11.9%
	Nonrespondents	56 8.9%	37 5.9%	534 85.2%	627 100.0%	51.8%
	Total	214 17.7%	100 8.3%	896 74.0%	1210 100.0%	100.0%

p<.001,  $\chi^2$  test

Table 11. Response Behavior across Different Surveys.

Table 11.a. Response Behavior in SL 2003 by Previous Behavior in LLP 2003.

		SL 2003				Total	Percent of Total
		Completes	Intro & Consent	Breakoffs During	Nonrespondents		
LLP 2003	Completes	1,026 71.7%	17 1.2%	42 2.9%	346 24.2%	1431 100.0%	29.0%
	Intro & Consent	24 46.2%	2 3.8%	1 1.9%	25 48.1%	52 100.0%	1.1%
	Breakoffs During Survey	207 47.2%	7 1.6%	31 7.1%	194 44.2%	439 100.0%	8.9%
	Nonrespondents	1,022 34.0%	20 0.7%	115 3.8%	1,848 61.5%	3005 100.0%	61.0%
	Total	2279 46.3%	46 0.9%	189 3.8%	2413 49.0%	4927 100.0%	100.0%

$p < .001, \chi^2$  test

Table 11.b. Response Behavior in LLP 2004 by Previous Behavior in SL 2003.

		LLP 2004			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
SL 2003	Completes	246 28.7%	101 11.8%	509 59.5%	856 100.0%	52.0%
	Breakoffs	5 6.0%	9 10.8%	69 83.1%	83 100.0%	5.0%
	Nonrespondents	50 7.1%	23 3.3%	634 89.7%	707 100.0%	43.0%
	Total	301 18.3%	133 8.1%	1212 73.6%	1646 100.0%	100.0%

$p < .001, \chi^2$  test

Table 11.c. Response Behavior in SL 2005 by Previous Behavior in LLP 2003.

		SL 2005			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
LLP 2003	Completes	312 81.5%	3 0.8%	68 17.8%	383 100.0%	30.5%
	Breakoffs	78 64.5%	4 3.3%	39 32.2%	121 100.0%	9.6%
	Nonrespondents	405 53.9%	12 1.6%	335 44.5%	752 100.0%	59.9%
	Total	795 63.3%	19 1.5%	442 35.2%	1256 100.0%	100.0%

$p < .001, \chi^2$  test

Table 11.d. Response Behavior in SL 2005 by Previous Behavior in LLP 2004.

		SL 2005			Total	Percent of Total
		Completes	Breakoffs	Nonrespondents		
LLP 2004	Completes	249 87.7%	3 1.1%	32 11.3%	284 100.0%	18.6%
	Breakoffs	92 77.3%	1 0.8%	26 21.8%	119 100.0%	7.8%
	Nonrespondents	625 55.6%	23 2.0%	477 42.4%	1125 100.0%	73.6%
	Total	966 63.2%	27 1.8%	535 35.0%	1528 100.0%	100.0%

$p < .001$ ,  $\chi^2$  test

Table 12. Kappa Statistic and Standard Error for Each Cross Classification.

	LLP 2003	SL 2003	LLP 2004	SL 2005
SL 2001	--	0.303 (0.021)	--	--
LLP 2003		0.274 (0.011)	0.201 (0.021)	0.190 (0.020)
SL 2003			0.213 (0.016)	0.276 (0.016)
LLP 2004				0.159 (0.014)

## Chapter 3. Causes of Survey Breakoffs

### 1. Introduction

A survey breakoff occurs when a respondent starts the survey but stops providing answers at some point during the survey, whether by choice or technology. Like unit nonresponse and other behaviors causing errors of nonobservation, breakoffs may present a threat to survey inference. This threat is large in web surveys. Meta-analyses find median reported breakoff rates of 16% and 34% (Musch and Reips, 2000; Lozar-Manfreda and Vehovar, 2002). However, despite the magnitude of the problem, breakoffs have received very little attention in the research literature.

There is very little that is known about breakoffs as a form of nonresponse, yet they occur at relatively high rates in web surveys. Without understanding the mechanisms generating breakoffs, they tend to be combined in administration reports with unit-nonresponse in most studies. Understanding the causes of breakoffs might allow for reduction of breakoffs through survey design, as well as for better post-survey adjustments for breakoff.

This paper has the goal of delineating breakoffs from initial nonresponse in web surveys, and of studying the causes of breakoff. The fundamental



premise is that since breakoffs occur after the survey is started by a respondent, their causes might be different than those of unit nonresponse. Certainly, common causes such as motivation and interest could link the two response behaviors to some degree. However, breakoff could also be related to the different cognitive demands of various types of questions and the cognitive abilities of respondents, among other factors like survey design features and browser problems.

In addition to the need for theories relevant to the causal mechanisms of breakoff, there are also empirical difficulties of studying breakoffs. Turning to the framework proposed in Chapter 1, in addition to survey design factors, respondent characteristics have to be available for those who break off. Furthermore, in order to study influences on breakoff there has to be sufficient variation in question characteristics. As breakoffs occur during the questionnaire, and could do so at varying rates, there must be question and page characteristics that induce breakoff; for example, some pages may make the respondent's task seem more burdensome. Finally, to create contrasts to analyze the effect of question characteristics on breakoffs, respondents should receive different forms of the questionnaire to be able to separate the effect of location in the questionnaire from that of a particular characteristic. This is both because the sequence of the characteristics can otherwise be completely confounded with order of the questionnaire, and because the perceived burden of certain features could interact with what the respondent has already experienced in the survey.

## 2. Background and Significance

Breakoffs are not unique to web surveys (Groves and Kahn, 1979; Catania et al., 1996; Stussman, Taylor and Riddick, 2003; Olson et al., 2004; Blumberg et al., 2005), and research in other modes has identified some factors affecting breakoffs. For example, allowing respondents in a telephone survey on sex topics to choose the gender of the interviewer significantly reduced breakoffs, while using longer questions increased breakoffs (Catania et al., 1996). However, the findings from telephone and face-to-face methods of data collection are not directly relevant to web surveys. The nature of the social interaction in interviewer-administered surveys may attenuate the effect of some causes of breakoffs (that could otherwise play a large role in web surveys), such as lack of interest in the topic. Therefore most of the breakoffs in interviewer-administered surveys, such as in the Behavior Risk Factor Surveillance System Survey (Osborn, Blumberg and Olson, 2000), the National Health Interview Survey (Stussman et al., 2003), the National Survey of Children's Health Interview Survey, and the National Immunization Survey (Olson et al., 2004), occur during the screening and respondent selection stages. This seems analogous to nonresponse to a request for a larger main survey or proxy nonresponse on behalf of a household member, rather than question demands causing selected persons to break off during the survey. One exception to these findings comes from Groves and Kahn (1979), where breakoffs were also

observed on section introductions within the survey. New sections in the survey provide a natural breaking point of the conversation and foreshadow additional material to come.

Mail surveys lack the social interaction of interviewer-administered surveys and therefore questionnaire characteristics can more heavily affect respondent cooperation. Unfortunately, in mail surveys only a binary outcome is generally observable – the respondent either returned the survey or failed to do so-- without a clear distinction between breakoff and unit nonresponse. Breakoffs in a web survey can be seen as cooperation decisions on every page of the survey. As pages can vary in a multitude of design features, numerous factors affecting this cooperation decision could be explored. Instead of merely measuring what proportion of sampled persons do not respond to a mail survey, in a web survey we can also measure the proportion of sampled persons who break off.

However, even treating survey participation as having three outcomes – respondent, nonrespondent, and breakoff – instead of two outcomes is insufficient depending on the purpose of the research. For a researcher who utilizes weighting for a single group of nonrespondents and breakoffs, recognizing that there are three outcomes should yield an improvement in post-survey adjustment, to the extent that breakoffs tend to be different from nonrespondents on key variables of interest. However, in order to be able to better reduce breakoff rates and anticipate biases in statistics resulting from them, identification of the causes of this response behavior is desirable. There

are differences between unit nonresponse and breakoff that should be expected, as unit nonrespondents generally do not see the questions in the survey. One likely set of causes for breakoff is the multitude of features within the questionnaire. The features that induce higher rates of breakoff could be informative about the processes causing this behavior.

Surveys can vary in multiple aspects throughout the questionnaire – topic, sensitivity of questions, burden of answering the question, format of the questions, number of questions on a page, layout, etc. To the extent that some of these features are associated with higher rates of breakoff, their identification can lead to understanding of the reasons for why people decide to stop cooperating. The decision to participate can be reevaluated on every page of a web survey, compared to an interviewer-administered survey where far more respondents seem to make a permanent cooperation decision at the time of the survey request. This view of multiple decisions to participate in a survey is valuable, both to be able to influence them and to be better able to adjust for them. Unlike unit-nonresponse, breakoffs are analogous to obtaining non-independent replicate measures of response propensities for each respondent. This view allows the estimation of reasons for nonresponse as a function of particular characteristics on pages with certain cognitive tasks. Furthermore, it permits the discovery of factors that have not been considered and tested before, as discussed in Chapter 1.

To help study survey breakoff, we need a theoretical framework, such as the one for unit-nonresponse developed by Groves and Couper (1998). While

this framework was constructed for a single cooperation decision and specifically for household surveys, it classifies the factors affecting survey participation into environment, respondent, survey, and interviewer, encouraging studies of nonresponse that do not look at one factor in isolation. The additional decision to break off is conditional on the respondent having made the decision to start, and is likely to be influenced by individual features within the questionnaire. Since breakoffs occur at the question level, models for data quality and measurement error may be more relevant. Beatty and Herrmann's framework (2002) for item nonresponse may be relevant here as one of the reasons they identify as a cause of item-nonresponse is the respondent's inability to retrieve the information needed to answer the question. As respondents continuously reevaluate their participation in the survey, an alternative to item-nonresponse is breakoff. For example, one way to minimize item-nonresponse in web surveys is to use prompts. Mooney, Rogers, and Trunzo (2003) found that while prompts – messages reminding or requiring responses based on omitted questions by the respondent – decreased item-nonresponse, their use may increase breakoff rates. It may not be just item-nonresponse – respondents' difficulty, discomfort, threat, or some other effect associated with how the respondent processes the question, may be associated with breakoff. A theoretical framework for this potential mechanism is provided by the survey response process model (Cannell et al., 1977; Cannell et al., 1981; Tourangeau, 1984; Strack and Martin, 1987; 1987). How the respondent comprehends the question, retrieves the information, uses it to make a judgment, maps it onto the response options, and edits it based

on sensitivity or threat, affects measurement properties and is likely to be also affecting item nonresponse and breakoffs as possible alternatives for the respondent.

Bosnjak and Tuten (2001) present a framework for web surveys that includes unit-nonresponse, breakoff, and item-nonresponse, but explain these behaviors as continuous dimensions resulting from three orthogonal factors – nonresponse is affected by motivation (to start the survey), breakoff is affected by opportunity (time and absence of technological problems to finish the survey), and item-nonresponse is affected by ability (defined as willingness to provide responses to sensitive questions). While some factors are likely unique to a response behavior, this framework implies no overlap in the mechanisms producing each of the three outcomes. Furthermore, the factors are somewhat difficult to operationalize.

None of these existing frameworks is appropriate as a general framework for web survey participation, as web surveys do have unique response outcomes such as breakoffs, while breakoffs in turn may have common causes with measurement error and item-nonresponse that requires widening of the scope of the framework.

For these reasons there is a need for a framework for studying participation decisions in web surveys: there are *multiple participation decisions* leading to *multiple outcomes* in web surveys that are made *throughout* the survey administration, and *multiple factors* that can be *common* to more than one decision. While some factors can be shared with those for household interview

surveys, such as topic of the survey and incentives, there are also unique factors such as questionnaire layout, respondent equipment, etc. Although it can occur through multiple turns, the household interview cooperation model has a single decision, whether to cooperate or not. A model for web survey participation has to accommodate a decision (although it may be superficial) to start the survey that can result in unit-nonresponse, multiple decisions to continue observable as breakoffs on different pages, decision to answer a particular question within a page that can result in item-nonresponse, and a decision on how to answer the question affecting the measurement properties of the response. The focus of this research is to gain understanding of what factors affect nonresponse decisions throughout the survey, resulting in breakoff.

There may be three sets of factors worthy of study: *page and question characteristics, survey design*<sup>7</sup>, and *respondent characteristics*.

### 2.1. *Page and Question Characteristics*

Page and question characteristics do not affect the initial participation decision but can affect decisions on whether to continue the survey. Such characteristics are seen by the respondent only after starting the survey. These characteristics are page-specific and are likely to affect the likelihood of breakoff on particular pages rather than the overall likelihood of breakoff during the survey. It is also possible that these factors have a cumulative effect over the

---

<sup>7</sup> Survey design refers to study characteristics such as topic, sponsor, and data collection protocol.

entire questionnaire. Such survey attributes include the type of questions asked, such as open-ended (Crawford et al., 2001), the actual length of the survey (e.g. MacElroy, 2000; Hogg and Miller, 2003; Deutskens et al., 2004), use of grid layout (O'Neil et al., 2003), and technical difficulties on particular pages such as software requirements (e.g. Schwarz and Reips, 2001; Bayer and Thomas, 2004; Couper et al., 2006). Apart from technical difficulties, page characteristics should have cognitive mechanisms through which they affect respondent decisions to break off.

A convenient framework for studying these mechanisms is the survey response process model (Cannell et al., 1977; Cannell et al., 1981; Tourangeau, 1984; Strack and Martin, 1987; 1987) that stipulates the types of processes respondents have to undergo when presented with a survey question. Respondents have to comprehend the question, retrieve relevant information, make a judgment, map the response in the required format, and edit the response, although not necessarily in this order. Questions vary in the degree to which each of these processes is involved, and while all have different potential for dissuading respondents from continuing by exerting different types of cognitive demands, it is still unknown which processes induce breakoffs.

It is likely that long questions place more burden on respondents in terms of comprehension as they require more processing of text and as a result of this increased demand on the respondent, will be associated with higher breakoffs, controlling for other features of that question and page. Similarly, questions that ask respondents to recall behaviors will be retrieval-intensive; questions that ask



for attitudes to be constructed will be judgment-intensive; questions that require the respondent to type will require more effort in providing the answer; and sensitive or threatening questions will affect the respondent's editing of the response (e.g., Sudman, Bradburn and Schwarz, 1996; Tourangeau, Rips and Rasinski, 2000).

In addition to the type of cognitive demand, breakoffs could be affected by the amount of the demand – the perceived burden. This is particularly relevant when considering breakoffs in terms of pages in the survey. Each page could have a varying number of questions. At one extreme, all questions could be placed on the same page and at the other extreme there could be only one question per page. Comparing a 5-page scrolling design to a paging design with more than 100 pages, Peytchev, Couper, McCabe, and Crawford (2006) found no significant difference in breakoff rates. However, this comparison is close to extremes on the scrolling vs. paging survey design continuum, with the scrolling design presenting large sections of the survey similar to paper questionnaires. There are many differences between two such designs that could affect respondent cooperation decisions. A paging survey with variability in the number of questions per page could allow the examination of the effect of the number of questions placed per screen on breaking off, as an indicator of the amount of burden exerted on the respondent. These effects do not need to be assumed to be immediately activated: they could be delayed, such as the number of questions on the previous page increasing the likelihood of breakoff on the current page, as well as the number of questions answered so far into the survey

– a measure of the cumulative burden experienced by the respondent to that point in the survey. How a respondent weighs continuing vs. breaking off can change over the course of the survey, and there is evidence that prior binary decisions can affect subsequent decisions (Katz, 1964); deciding not to break off on a page can make a person less likely to break off on a subsequent one. Alternatively, the risk of breakoff could increase over the course of the survey. Busemeyer and Townsend (1993) stipulate that people have a tolerance threshold in decision tasks; in the context of breakoffs, this implies a cumulative effect of burden leading to survey termination. In a web survey experiment, Galesic (2006) found that breakoffs are associated with self-reported interest and burden on the previous page. Rather than using respondent self-reports, the idea of a threshold for burden can be tested by parameterizing burden as the cumulative number of questions asked and in addition to the immediate effect of the current page, a delayed effect of burden by the number of questions on the previous page. This is the parameterization used in this study; respondent burden in survey methodology has long been acknowledged to have multiple dimensions (Bradburn, 1978).

There is another kind of demand that occurs, that of repeated survey requests. Analogous to the initial survey invitation, introductory pages announce the beginning of a new set of questions, implicitly asking for further survey participation. They carry similarities with the initial request in that more survey responses are requested and often a new topic is introduced. Increased breakoff

at such points in the survey has been found in telephone surveys (e.g., Groves and Kahn, 1979) and is expected to occur in web surveys for the same reasons.

## 2.2. *Survey Design*

A second set of factors that can affect breakoff propensities are survey design features. Those are also under researcher control like the question and page characteristics, but respondents can be exposed to them without starting the survey; they are not manipulations on features within the survey. These features include the announced survey length (e.g. Crawford et al., 2001; Trouteaud, 2004), implied survey length by presenting progress indicators throughout the survey (Conrad et al., In press), sending prenotifications of the survey request (e.g. Kaplowitz et al., 2004), sending reminders to respond to the survey (e.g. Deutskens et al., 2004; Kaplowitz et al., 2004), content of invitation (e.g. Trouteaud, 2004), timing of reminders following a survey request (e.g. Crawford et al., 2001), type and amount of incentives (Bosnjak and Tuten, 2003; Deutskens et al., 2004), etc. For example, if some respondents are presented very slow feedback regarding their progress they can be discouraged from participation after starting the survey by implying a long survey.

### 2.3. *Respondent Factors*

Respondent factors include socio-demographic characteristics, predispositions, previous survey experience and outcome, as well as the respondent's survey taking environment. Common measurable characteristics are age, gender, race and ethnicity, and various study specific categorizations that are related to participation decisions. Most of these do not have a causal relationship with response outcomes, but are proxies for many of the causes that cannot be measured directly. Education has been used as a proxy measure of cognitive sophistication in a survey context (Sigelman, 1981; Krosnick and Alwin, 1987; Knauper, 1999), finding that higher levels of education are associated with smaller response order effects.

Age has also been used as an indicator of cognitive ability in survey responding (Herzog and Dielman, 1985; Andrews and Herzog, 1986; Knauper et al., 1997), with findings consistent with cognitive aging – older respondents associated with larger response order effects, more item-missing data, and lower proportion of true score variance explained. There is evidence from the Seattle Longitudinal Study that as respondents age, their mental abilities for verbal meaning, inductive reasoning, numeric ability, and word fluency decrease (Schaie, 1994). These are likely to adversely affect the comprehension, retrieval, and judgment stages of the survey response process model, while evidence that older respondents exhibit higher social desirability (Campbell, Converse and Rogers, 1976) would also suggest more editing of responses.

To the extent that some respondents have lower cognitive sophistication and ability, they can find questions more difficult than other respondents, which can dissuade them from continuing the survey. Under this hypothesis, older respondents and less educated respondents are more likely to break off.

If respondents who break off are those who tend to have cognitive difficulties in answering the questions, this should be evident not just through their demographic characteristics, but also through their behavior. Difficulty in comprehending the question, retrieving information, making a judgment, and responding would increase the time to move to the next question. It can be hypothesized that those who break off are not as serious about the survey task as those who continue – the group that Bosnjak and Tuten label “lurkers” (2001) – but if they actually spend more time on questions, this would mean that these respondents do put effort into the survey task and practitioners should minimize breakoff rates.

There are different patterns for breakoffs that could be postulated. At the beginning of the survey respondents could be trying to make a decision whether to complete the survey, which would result in early breakoffs. A second alternative is that respondents could be fatigued by the length of the survey (MacElroy, 2000; Hogg and Miller, 2003; Deutskens et al., 2004), become less interested, and more burdened (Galesic, 2006), that would result in increasing breakoff rate late in the survey. Thirdly, breakoffs may be a random phenomenon with a somewhat uniform rate of breakoffs throughout the survey.

We propose that breakoffs are instead a fourth alternative: in addition to the simplistic mechanisms described above, breakoffs are also a function of the different cognitive tasks in each page in the questionnaire, burden accumulated over the course of the survey, overall survey design features, respondent characteristics, respondent environment, and respondent behavior. In other self-administered modes, such as in mail surveys, respondents can see the entire questionnaire prior to completing it, providing the respondent the ability to make a cooperation decision that is informed by various characteristics of the questionnaire – topic, length, level of effort, requests for sensitive or threatening information, etc. In paging web surveys respondents are exposed to these questionnaire characteristics one page at a time, rather than at the very onset of the survey. If pages have random assignment of location in the survey across respondents, common characteristics of pages that elicit higher rates of breakoff would help understand the causes of breakoff.

Understanding the mechanisms for survey breakoff would help to minimize breakoff. For example, a set of questions that relies on the respondents' ability to retrieve information may induce higher rates of breakoff. Placing such questions at the beginning of the survey will result in collecting less data on these respondents. To the extent that those who break off are different from those who complete the survey on key variables, biases in survey estimates will be higher than if the retrieval-intensive questions had been placed at the end of the survey. Furthermore, to the extent that different types of questions lead to different respondents breaking off, order of the questions or sets of questions

within the questionnaire will influence the mix of respondents and affect survey estimates.

In sum, there are multiple response outcomes in web surveys. Breakoff can be affected by factors that could not affect unit nonresponse; therefore theories on nonresponse are not suitable for this type of response behavior. Mechanisms for web survey breakoff should include within-questionnaire causes, in addition to the respondent characteristics and survey design features that are observed prior to the start of the survey. Many of these design features could be linked to survey response processes. These cognitive causes could cumulate over the course of the survey, burdening respondents and increasing the risk of breakoff. Furthermore, respondents who have lower cognitive ability and cognitive sophistication may have more difficulty with the respondent role and will likely break off at higher rates. Finally, respondents who break off are not necessarily poor respondents that do not take the survey task seriously, but they are rather respondents who find the task difficult.

### 3. Data and Methods

Two web survey studies were conducted in December 2003 – January 2004 (Study 1) and December 2004 – January 2005 (Study 2). These studies had the objective of testing various visual design features in web surveys, which also provides a needed variety of questions for the current analysis of breakoffs. The surveys have identical design as part of the same program of research,

conducted by Market Strategies Inc. and designed by the same primary investigators with considerable overlap (replication in the second survey) of experiments. The samples were drawn in the same way from the same pool of respondents, just over half from the Survey Sampling Inc. (SSI) national web panel and the rest recruited through pop-up messages in America On-Line (AOL) web pages. Among the 3,195 respondents who started the first study, 478 broke off (15%), and among the 2,831 who started the second study, 244 broke off (9%); these rates are further broken down by sample in Table 13. Study 1 was longer, as the median time for completing it was 21.3 minutes versus 18.3 minutes for Study 2. Apart from a different mix of questions, there was an experiment with a slider bar response scale in Study 1 that produced high breakoff rates on a few questions, which is incorporated in the analyses.

**Table 13: Response Dispositions for Both Studies by Sample Source**

		Breakoff Rate	Total
Study 1	SSI	15.8%	1695
	AOL	14.0%	1500
Study 2	SSI	9.1%	1498
	AOL	8.0%	1333

The questionnaires involved no skip patterns, so that exposure to particular questions is not dependent on substantive responses and respondent characteristics, needed for unbiased estimation of causes for breakoff. Both studies incorporated experiments that randomly assigned conditions to respondents, such as the number of questions in particular pages and the display of progress indicators. In Study 2 the location of sets of questions within the



questionnaire was randomly assigned to respondents. Surveys commonly confound the page and question characteristics with location in the questionnaire. The random assignment of order allows the separation of effects of page characteristics from the location (i.e., early vs. late) and the context (i.e., characteristics of preceding pages). To illustrate, the fifth page for one respondent may follow an introductory page and have a single sensitive closed-ended question that requires retrieval, while for another respondent the fifth page may follow another page with a question and display four nonsensitive closed-ended questions that require more judgment processes.

Variables were selected and constructed to represent the different sets of factors expected to affect breakoffs.

### 3.1. *Questionnaire Characteristics*

Respondents received pages in different order. Characteristics of each page were coded, accounting for all random assignments that manipulated which questions were displayed, their format, and the order in the questionnaire<sup>8</sup>. The variables that were coded for each page included: number of questions in the page, number of questions in the previous page, cumulative number of questions asked prior to the current page, whether the question was long (multiple sentences), required an open-ended response, required a numeric response,

---

<sup>8</sup> While for some research questions it is preferable to discuss question characteristics as a lower level unit of analysis rather than page characteristics, when a respondent breaks off only the page is captured by the survey system. However, the two surveys typically placed one question per page, and when multiple questions were displayed, they tended to be very similar and the characteristics of all the questions were captured.

whether the page was an introduction to a new section of the questionnaire, whether any of the questions were on sensitive topics, were comprehension intensive (longer than one sentence), were retrieval intensive (typically behavioral questions), were judgment intensive (typically attitudinal questions), whether definitions were provided for key terms in the questions, and whether a Java-applet slider bar was used as a response scale.<sup>9</sup>

### 3.2. *Survey Design*

Another manipulation was coded at the questionnaire level that altered respondents' perception of the length of the questionnaire by displaying fast progress at the beginning, slow progress at the beginning, actual progress, or no progress at all (Conrad et al., 2005). To account for between study differences in breakoff rate, a dummy indicator for Study 2 was created. An indicator for sample source, AOL versus SSI was also included.

### 3.3. *Respondent Factors*

#### 3.3.1. *Respondent Characteristics*

Auxiliary demographic data for the breakoffs was available only for the SSI respondents as they had provided key demographic information when they became members of the panel. These variables included age in years, gender,

---

<sup>9</sup> High breakoff rates were observed for respondents who were presented these questions. It is unknown whether it displayed properly for all respondents causing the breakoffs, but analyses need to account for it to avoid bias in other parameter estimates.

race (collapsed into an indicator for White), marital status, education, occupation, and income (with nine categories including one for missing).

### 3.3.2. Respondent Behavior

In order to use time spent on the first question as a proxy of respondent effort in both surveys, time was standardized within each survey and experimental condition. Although the first questions in the two studies were almost identical, respondents in Study 2 were randomly assigned to answer a different part of the survey first, changing the first question seen by the respondent.

While the variation of question types and varying question order for respondents allows the page characteristics to be separated from location in the survey, it requires a more complex model. A discrete hazard survival model with page-varying covariates is employed, where the page characteristics are allowed to vary for across pages and respondents, and questionnaire-level experiments, respondent characteristics, and respondent behavior are added sequentially to the model as page-invariant covariates. In the model below,

$$\ln\left(\frac{P_{iq}}{1 - P_{iq}}\right) = \alpha_q + \beta_1 X_{i1} + \beta_2 X_{i2}(q)$$

$P_{iq}$  is the probability for respondent “i” to breakoff on page “q”,

$\alpha_q$  is the baseline hazard of breaking off on page “q”,

$X_{i1}$  is a vector of page-invariant covariates for respondent “i”, and

$X_{i2}(q)$  is a vector of page-varying covariates for respondent “i” for page “q”.

First a model is fit, testing whether the different cognitive tasks, and burden defined as immediate amount of request (number of questions per page), inoculation effect (number of questions on previous page), and cumulative burden (cumulative number of questions asked to that point) increase the risk of breakoff. This is tested using the variability in questionnaire characteristics and survey design variables. A second model is fit to test whether these effects remain after controlling for respondent characteristics, as well as testing the hypothesis of respondent cognitive ability and sophistication being associated with higher rates of breakoff. A third model builds upon question/page and respondent characteristics by testing the hypothesis that respondents who break off are not haphazard respondents, but rather respondents who try but may find the task difficult. This is tested by the addition of respondent behavior as time spent on the first question.

These three models use only the SSI samples as only they contain respondent socio-demographic characteristics on all respondents. The same models are replicated using both SSI and AOL samples, except for the model adding respondent characteristics, as Models 4 and 5. This allows a limited test of the robustness of the findings across samples, as well as providing greater degree of certainty in the results from the higher statistical power.

#### 4. Results

The pattern of breakoffs can be presented graphically for the first study because it did not vary the order of the pages (Figure 4). Breakoffs seem to be not just a function of location in the questionnaire as they do not monotonically increase or decrease, nor are they about the same rate with small variability. For example, some of the higher breakoff rates are on pages with grids of multiple questions, section introduction pages, a question on amount spent on alcohol, a page with multiple open-ended questions, and pages with Java slider bar scales. At the question/page level, the respondents who were assigned to no progress indicator or presented slow progress at the beginning seem to breakoff at higher rates throughout the entire survey. However, this figure confounds many factors by displaying only one of many page characteristics at a time. Therefore, conclusions about the effect of specific characteristics could not be made, requiring a multivariate analysis on the combined data that varies the order of the page characteristics within the questionnaire.

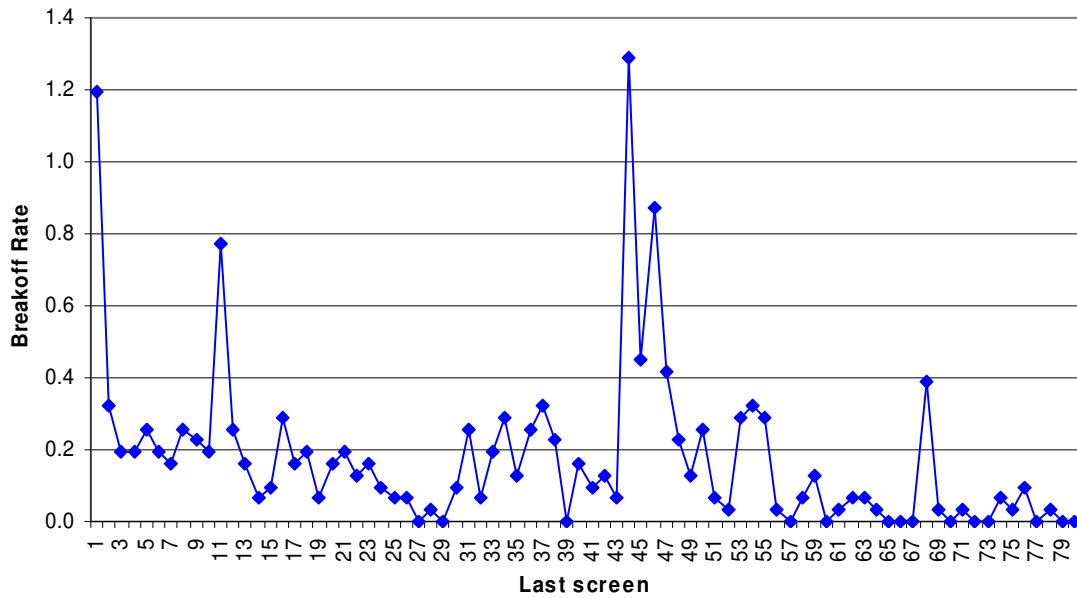


Figure 4: Breakoff by Page Number in Study 1

Combining the two studies, additional variation in question/page characteristics is achieved, with a total sample size of 6,026 respondents who started a survey, and a maximum of 87 pages. The Kaplan-Meier survival estimates are presented in Figure 5, showing somewhat smooth survival curves, censoring slightly earlier in Study 1 as it had fewer pages.

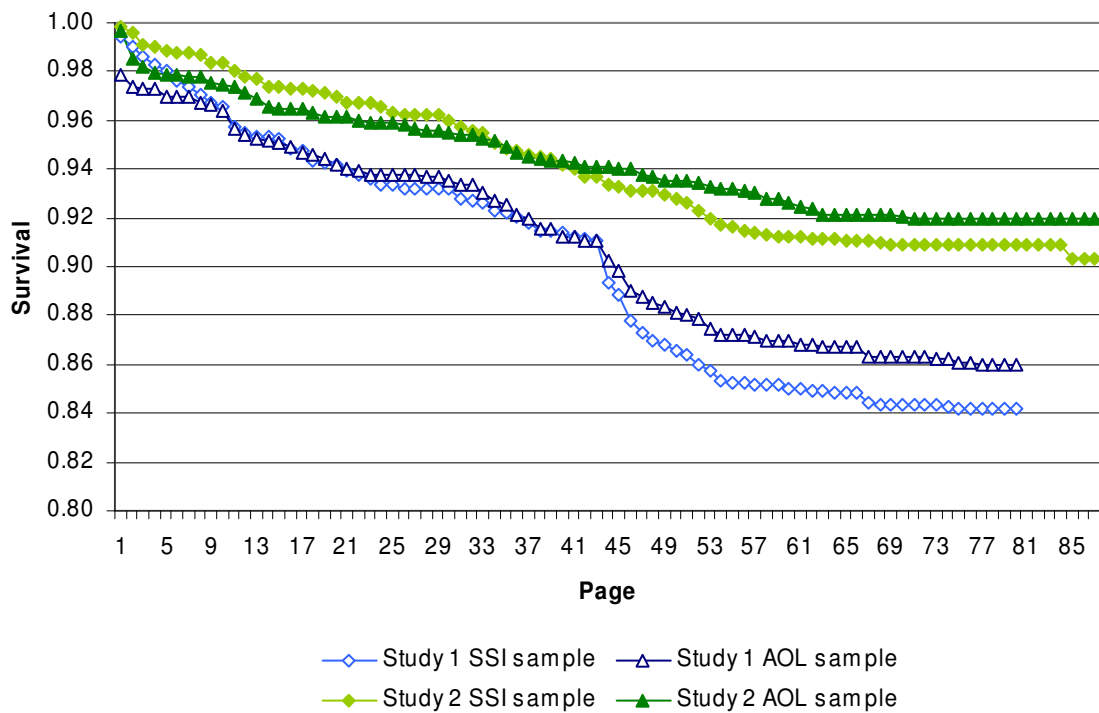


Figure 5: Kaplan-Meier Survival Estimates by Study and Sample Source

A model predicting breakoff was first estimated with page and questionnaire characteristics, presented in Table 14. The relative risk of breakoff increases by 23% for each additional question in a page, but neither the number of questions in the previous page, nor the cumulative number of questions asked to that point have an effect on breakoff. Open questions increase the risk by almost two and a half times, while long questions (those comprised of multiple sentences) triple the risk of breakoff. There is some indication that pages with questions that require retrieval tasks induce breakoffs, but the coefficient is not significant at the .05 level. Consistent with the finding from telephone surveys, pages that introduced a new section in the survey (and no questions were displayed) had more than twice the relative risk of breaking off, an influence of

breakoff that resembles the initial decision to cooperate. Respondents who were shown slow progress early in the questionnaire had more than twice the risk of breakoff relative to those who were not shown a progress indicator at all.



Table 14: Discrete Hazard Survival Models with Page-Varying Covariates Predicting Breakoff in the Combined Studies, SSI samples.

Factor	Predictor	Model 1		Model 2		Model 3	
		Relative Risk	p- value	Relative Risk	p- value	Relative Risk	p- value
<b>Questionnaire</b>	Number of questions	1.227	<.001	1.226	<.001	1.235	0.001
<b>Page and Question Characteristics</b>	Number of qns on previous pg	0.985	0.817	0.990	0.876	1.014	0.833
	Cumulative number of qns	1.008	0.419	1.009	0.393	1.010	0.346
	Numeric question (1=yes)	1.325	0.202	1.367	0.160	1.257	0.322
	Open question (1=yes)	2.394	0.029	2.413	0.029	2.408	0.031
	Long question (1=yes)	2.898	0.034	2.859	0.037	2.976	0.033
	Section Introduction (1=yes)	2.613	0.008	2.650	0.007	2.153	0.044
	Sensitive question (1=yes)	0.644	0.239	0.625	0.213	0.568	0.143
	Retrieval question (1=yes)	1.522	0.086	1.526	0.087	1.463	0.131
	Judgment question (1=yes)	1.461	0.189	1.487	0.172	1.330	0.337
	Definitions (1=yes)	1.099	0.773	1.094	0.786	1.055	0.873
	Slider-bar scale (1=yes)	4.844	<.001	4.871	<.001	4.876	0.000
<b>Survey Design</b>	<b>Progress Indicator:</b>		<.001		<.001		<.001
	No progress indicator	1.000	--	1.000	--	1.000	--
	Slow to fast progress	1.582	0.036	1.536	0.051	1.526	0.056
	Fast to slow progress	0.570	0.022	0.557	0.018	0.545	0.014
	Constant progress	0.950	0.822	0.934	0.764	0.900	0.646
	Study 2 (1=yes)	0.606	0.037	0.603	0.038	0.572	0.024
<b>Respondent Characteristics</b>	<b>Male (1=yes)</b>			1.124	0.314	1.123	0.327
	<b>Marital Status:</b>				0.439		0.515
	Married			1.000	--	1.000	--
	Single			1.034	0.815	1.069	0.642
	Separated/Div/Widowed			0.836	0.244	0.869	0.369
	<b>White (1=yes)</b>			0.767	0.059	0.822	0.172
	<b>Employment:</b>				0.279		0.217
	Employed full-time			1.000	--	1.000	--
	Employed part-time			1.083	0.597	1.018	0.910
	Not employed			0.843	0.236	0.795	0.117
	<b>Education:</b>				0.002		0.004
	High school or less			1.000	--	1.000	--
	Some college			1.163	0.815	1.243	0.642
	College			0.894	0.244	0.970	0.369
	Grad school or graduate			0.565	0.008	0.646	0.045
	<b>Age</b>			0.994	0.174	0.990	0.025
	<b>Occupation:</b>				0.341		0.229
	Mgt/IT/Doctor/Educ/Small Business			1.000	--	1.000	--
	Homemaker			1.315	0.251	1.432	0.140
	Student			1.474	0.119	1.539	0.090
	None of the above			1.301	0.083	1.344	0.059
	Missing occupation			1.719	0.126	1.891	0.074
	<b>Income:</b>				0.079		0.166
	Under \$20,000 income			1.000	--	1.000	--
	\$20,000-29,999			0.719	0.086	0.692	0.059
	\$30,000-39,999			1.016	0.928	1.030	0.872
	\$40,000-49,999			0.533	0.007	0.592	0.026
	\$50,000-59,999			0.835	0.395	0.842	0.429
	\$60,000-74,999			0.787	0.271	0.837	0.424
	\$75,000-99,999			0.621	0.051	0.720	0.189
	\$100,000+			0.748	0.258	0.851	0.540
	Missing income			0.965	0.857	1.044	0.830
<b>Respondent Behavior</b>	Standardized time on first question					1.345	<.001

Number at risk/number of breakoffs in each model: 3193/392; 3192/390; 3190/388; 3164/377.

Among the respondent characteristics available for the SSI panel samples, respondents with at least some graduate education were less likely to break off (Table 14, Model 2). There was limited support for white ( $p=.059$ ) and higher income ( $p=.079$ ) respondents also being less likely to break off.

Respondent behavior as time spent on the first question was significantly associated with risk of breakoff, with each standard deviation slower responding resulting in 34% higher risk of breakoff, shown in the last row for Model 3. In addition, age became statistically significant in the opposite direction to the cognitive ability hypothesis – older respondents were associated with a lower likelihood of breakoff. This could be because younger adults can be exposed to more surveys as they are more active online (older adults have access at lower rates, and use the Internet less intensively, Fox, 2005), and also because older adults may be more motivated to finish the task that they have started.

These models were replicated using both samples, excluding the model with Respondent Characteristics as those were not available on the AOL sample. Table 15 shows Models 4 and 5 based on the full set of data from both SSI and AOL samples. While the SSI sample is drawn from a volunteer web survey panel, the AOL sample is recruited through pop-up messages in AOL web pages, leaving the potential for unobserved differences between the samples. To verify whether the two samples can be pooled for this analysis, Model 5 was fit separately for SSI and AOL respondents. Among the variables in the model, the two samples differed in likelihood of breakoff associated with progress indicator condition as AOL respondents were not affected by being presented very slow

progress on the early pages. To account for these sample differences, interaction effects between sample and progress indicator were estimated in Models 4 and 5.

Table 15: Discrete Hazard Survival Models with Page-Varying Covariates Predicting Breakoff in the Combined Studies, SSI and AOL samples.

Factor	Predictor	Model 4		Model 5	
		Relative Risk	p-value	Relative Risk	p-value
<b>Questionnaire</b>	Number of questions	1.230	<.001	1.227	<.0001
	Number of qns on previous page	0.973	0.600	1.007	0.895
<b>Page and Question Characteristics</b>	Cumulative number of qns	1.011	0.142	1.012	0.107
	Numeric question (1=yes)	1.608	0.004	1.479	0.022
	Open question (1=yes)	2.515	0.003	2.448	0.005
	Long question (1=yes)	2.044	0.083	2.213	0.055
	Section Introduction (1=yes)	4.091	<.001	2.683	0.001
	Sensitive question (1=yes)	1.018	0.944	0.823	0.471
	Retrieval question (1=yes)	1.346	0.119	1.327	0.146
	Judgment question (1=yes)	1.882	0.004	1.566	0.046
	Definitions (1=yes)	1.259	0.368	1.187	0.512
	Slider-bar scale (1=yes)	4.393	<.001	4.603	<.0001
<b>Survey Design</b>	<b>Progress Indicator:</b>		<.001		<.001
	No progress indicator	1.000	--	1.000	--
	Slow to fast progress	1.735	0.004	1.583	0.015
	Fast to slow progress	0.625	0.032	0.570	0.011
	Constant progress	1.042	0.836	0.939	0.751
	AOL*Slow to fast	0.709	0.089	0.835	0.384
	AOL*Fast to slow	1.550	0.064	1.864	0.010
	AOL*Constant	0.915	0.685	1.066	0.775
	AOL sample (1=yes)	0.830	0.123	0.735	0.017
Study 2 (1=yes)	0.636	0.015	0.542	0.001	
<b>Respondent Behavior</b>	Standardized time on first question			1.377	<.0001

Number at risk/number of breakoffs in each model: 6026/673; 5944/641.

Compared to Model 1, Model 4 shows a significant effect of numeric questions, estimated as 61% more likely to break off relative to closed-ended questions, under the model. In addition to the significant predictors in Model 1, Judgment questions were also associated with higher risk of breakoff in Model 4. AOL respondents were less affected by Survey Design as interactions for progress indicator conditions with AOL were in the opposite direction of the main

effects of progress indicator speed. The risk of breakoff associated with each study remained approximately the same.

## 5. Discussion and Conclusions

There are four key conclusions that can be drawn from this research:

1. Breakoff is a function of questionnaire characteristics that are seen after the survey is started.

Since page and question characteristics are not seen by initial unit nonrespondents, the finding that they are significantly associated with breakoff delineates the two types of response behavior. Since there are different causes affecting breakoff, those who abandon the survey should not be simply combined with those who fail to start the survey. However, in order to treat breakoffs differently, the unique causal mechanisms need to be captured.

2. Comprehension, judgment, and response mapping induce higher rates of breakoff.

To some degree breakoffs are indeed a function of the survey response process model and amount of demand on the respondent. Pages that had questions requiring substantial comprehension (long questions), judgment, and

response mapping (numeric and open-ended questions) were associated with higher risk of break off (Table 15). Long questions require more effort by the respondent to read the question and were associated with an increased risk of breakoff. While the availability of definitions for some questions did not affect breakoff, it is likely that respondents make the distinction that unlike the question itself, the definitions are not essential to the task of answering a question.

The first set of models that were fit to the data from the SSI panel respondents showed some indication for retrieval questions to increase breakoff risk, compared to other mostly factual questions. This was not even marginally significant after controlling for respondent behavior and also in the analyses of the combined SSI and AOL data.

However, judgment questions were associated with almost twice the risk of breakoff in the combined data. A key difference between the two samples is that the AOL respondents were not members of a volunteer panel and had less prior survey experience. Furthermore, after controlling for respondent behavior, i.e., respondents spending more time on the first question were also more likely to break off, the effect of the judgment processes was reduced (Model 5).

Unfortunately the interaction between respondent and page characteristics could not be tested with sufficient statistical power (page characteristics have indicator variables for each page and respondent), and when it was attempted, the interactions were not significant. This interaction is key for future research, as it means that the respondents who may be putting in effort in the survey and/or

finding it more difficult than other respondents, are also more likely to break off on questions relying on judgment.

Open-ended questions require both formulation of an answer and typing it, in the final response process, and also led to higher risk of breakoff. This effect was smaller when the response was numeric.

While none of the models found the sensitive topic of questions to be associated with breakoff, it is also possible that more sensitive questions could produce such effects, or that this effect would be stronger for non-volunteering respondents. However, this is consistent with an earlier finding in a web survey experiment (Frick et al., 2001) where placing personal information questions early in the questionnaire decreased the breakoff rate.

The amount of the demand in the page had a similar effect, as each additional question in a page increased the likelihood of breakoff by about 23%. This finding needs to be taken with caution, given the limited number of pages that had multiple questions in these studies. The effect is immediate, as the number of questions in the previous page did not have an effect on breakoffs. The overall accumulated burden measured by the cumulative number of questions asked, also had no effect on breakoff, when controlling for page characteristics. In addition, the switching of topic by means of section introduction pages also increased the likelihood of breakoff, although this mechanism is likely similar to that of unit nonresponse as the refusal to start a set of questions.

Making respondents believe that the survey is longer by presenting a progress indicator that changes very little at the beginning more than doubles the risk of breakoff in all models, relative to not showing any progress indicator, confirming this finding by Conrad and colleagues (Conrad et al., 2005) when survey and respondent covariates are not used. This is also an example of how overall questionnaire design features could affect breakoffs in a survey, and the need for understanding the consequences of such design features.

3. Respondent cognitive sophistication is associated with lower risk of breakoff.

There is mixed support for this hypothesis that warrants further research. Higher cognitive sophistication as indicated by level of education was associated with lower risk of breakoff. However, cognitive ability indicated by age was not significant, and after controlling for response time on the first page, its coefficient became significant but in direction opposite to expectations – older age associated with lower risk of breaking off. This does not necessarily reject the cognitive aspect of survey breakoffs, but may in fact require better measures of respondent cognition. Just as age involves far more constructs than just cognitive ability, education can subsume many other respondent differences in addition to the cognitive sophistication used in the literature.

4. Those who break off are not haphazard respondents, but may be finding the survey task difficult.

While some may believe that respondents who break off are simply haphazard respondents that do not put effort in answering a question, it seems to be quite the contrary – it is those who spend more time on answering questions that may be finding the task difficult and therefore breaking off. In both the SSI sample and the combined samples those who spent a standard deviation longer on the first page with a question were 37% more likely to break off. While time only on the first question was used in order to keep most of the breakoff cases in the analysis, there is some degree of variability of time within respondents. For example, the correlation between the natural log of time between the first and the second question was .53 in Study 1 and .48 in Study 2. This finding needs further research, both empirical and theoretical. One likely hypothesis is that respondents who break off try to answer questions carefully, hence we should try to keep them in the survey whether by assisting them by avoiding questions that they find difficult or varying the location of these questions in the survey.

An additional finding involves multiple modes and links unit nonresponse to breakoff: section introductions within the survey, found to cause breakoffs in telephone surveys (Groves and Kahn, 1979), induced significantly higher rates of breakoff in the web surveys. These pages do not have any questions and the decision resembles that of initial cooperation – starting a set of questions on a topic.



The first two conclusions need to be taken with some caution, as although there was some randomization of page order, it was by section and there were many pages that always followed the same other pages. All the conclusions are subject to replication on other types of samples, specifically on non-self selected respondents that may be very different in terms of motivation and hence breakoff behavior.

In summary, breakoffs seem to have a cognitive component in that different cognitive demands of the questions and different cognitive sophistication of the respondents lead to breakoff, while at the same time those who break off seem to try to put in effort.

While further empirical research on breakoffs is warranted, some practical solutions can be proposed. If some questions induce higher rates of breakoff by particular respondents, while other question characteristics lead to breakoff by different respondents, then placement of modules of questions within the questionnaire can be varied. This way respondents breaking off for a particular reason will not break off at the same point in the survey. Another benefit from using this framework is using it to identify other informative variables that are associated with both the dependent variables and breakoff processes. One such example are questions that rely on judgment processes – respondents who break off on such questions can be different in terms of responses regardless of the topic involved, and collecting and using additional auxiliary variables such as time spent on responding that will improve adjustment models through imputation and/or weighting.

## Chapter 4. Web Survey Design and Correlated Measurement Error

### 1. Introduction

The ability to make inferences about some attributes of a target population from a survey depends on the degree to which the sampled persons respond with minimal error. Measurement error, defined as deviations from expected values, can be affected by survey design. When validation data are available, measurement error can be estimated for respondents under different survey design conditions, permitting judgments about preferable designs. Validation data are not commonly available for factual and behavioral questions, and records are not obtainable for attitudinal variables. This leads to an inability to determine preferable designs, even when experiments are embedded in surveys and differences are found between measurement conditions.

One such design decision in web surveys is whether to place related questions on the same page or on separate pages. When questions are in the same page, inter-item correlations increase (Couper et al., 2001). When the questions are used to form scales, a researcher could be misled to select this design, thinking it provides higher reliability. While higher *estimated* reliability could be from lower measurement error of the items identifying a latent construct,

it could also be from correlated measurement error across the items.

Mechanisms supporting either cause can be identified, yet in the absence of a method to evaluate the measurement properties of the questions under each design, no guidance could be provided on selection of a design that improves survey inference.

This study presents a method that could be used to evaluate the combined effect of uncorrelated and correlated measurement error in such web surveys in the absence of validation data.

The second concurrent purpose of this study is to employ this method to evaluate the effect of placing multiple attitudinal questions on the same page in terms of measurement error, a common yet uninformed design decision in web surveys that is known to produce measurement differences.

## 2. Background and Significance

In survey methodology, measurement error is an error of observation, defined as the difference between the truth (or, sometimes, the expected value) and the observed response. Often it is estimated by the difference between what is obtained from records and the observed survey responses. This has been done for a variety of behaviors, from medical visits to voting in elections. Such investigations have helped improve survey design to reduce measurement error, such as improving questions and instrument design.

There is considerable evidence that responses to attitudinal questions can be affected by survey design, such as whether the questionnaire is on paper or computerized (e.g., Ramos, Sedivi and Sweet, 1998). Yet for attitudinal questions, records with true values are rarely if ever available. This poses a problem for comparison of different survey designs, in terms of reduction of errors of observation.

## 2.1. Background on Estimation

Evaluation of measurement error in responses to questions measuring the same construct, in the absence of validation data, has been of interest to psychometricians for over a century (e.g., Spearman, 1904). Classical Test Theory (CTT), introduced by Spearman (1904; 1904) and further developed by Novick (1966), postulates that every observed measure (item, question) is an imperfect indicator of an underlying true score. Under model-based assumptions, the use of multiple indicators ( $x_i$ ) allows the identification of the true score ( $X$ ) and measurement error components ( $e_i$ ):

$$x_i = X + e_i \quad [1]$$

Under this model, where “i” denotes a specific item, assumptions are made that  $\text{cov}(e_i, e_j) = 0$  and  $\text{cov}(e_i, X) = 0$ , where “i” and “j” are different items ( $i \neq j$ ).

Multiple estimators of the ability of a set of indicators to identify the true score have been developed under this true score framework, the most common of which is Cronbach's Coefficient  $\alpha$  (1951), which is the proportion of the observed variability that is common to all items, adjusted for the number of items, shown in equation 2:

$$\alpha = \frac{n}{n-1} \left[ 1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_T^2} \right] \quad [2]$$

Where "n" is the number of items,  $\sigma_i^2$  is the variance of item "i", and  $\sigma_T^2$  is the variance of the total scale. In this classical view, sampling is the selection of items and measurement error is the property of the items. This estimate of reliability (the consistency with which a construct can be identified) has dominated the field since its introduction and although a large proportion of publications using scales fail to provide reliability estimates, Coefficient  $\alpha$  is the one most commonly used (Vacha-Haase, 1998) despite observations about its limitations (Cronbach, 2004).

Under the assumptions of CTT, less measurement error in the responses to the items yield higher reliability estimates. In order for Cronbach's Coefficient  $\alpha$  to be useful in evaluating measurement error, the critical assumption is that the measurement errors of the items are not correlated. Without an explicitly known structure of the errors, the correlation between the errors is confounded with

correlation between the true scores. Furthermore, correlation between the true score and the measurement errors is also confounded. A more recent theory, Generalizability Theory (GT) (Cronbach, Rajaratnam and Gleser, 1963), extends CTT in that it permits items to measure non-identical true scores, yet it still requires that the uncorrelated errors assumption is still made in order to identify the true score.

As the true score and the correlated measurement error are often confounded, the only choice left to a researcher is to redefine what the true score is. If the method of data collection induces correlated measurement error, under this often imposed approach the true score will be defined as “under the study’s measurement conditions,” as explained by Brennan (2001, p.7):

In Classical Theory and Generalizability Theory, error (as defined in or by the models) does not mean mistake, and it does not mean 'model misfit' in the usual sense of that term. Rather, important aspects of error are defined directly or indirectly by the investigator. This is eminently obvious in Generalizability Theory, which requires that the investigator define, among other things, the type of error under consideration. In Classical Theory, the investigator effectively defines error through specifying a data collection design and adopting the Classical Test Theory model. That is why traditional coefficients of internal consistency, stability and equivalence typically lead to different estimates of error variance. That is, the

error variances associated with these coefficients are *not* different estimates of the same quantity; rather, they are estimates of different quantities. In my opinion, it is difficult to overstate the importance of understanding that the investigator, through various overt or hidden choices, is actively involved in deciding what shall be considered as error.

Inter-item correlations between the same set of questions can be manipulated through survey design. Simply putting the questions together one after the other has been demonstrated to increase inter-item correlations (McGuire, 1960; Smith, 1983). The CTT-based methods would lead a researcher to place questions on the same page, possibly erroneously in the presence of correlated measurement error. The widely used Cronbach's  $\alpha$  has been found to vary not just across different selections of questions (for which it is intended), but also across subgroups of respondents and study sample size (Vacha-Haase, 1998). This requires other methods of evaluating measurement error that allow correlated measurement errors through the survey design.

Work has been done on violations of assumptions under CTT through simulation studies (e.g., Zimmerman, Zumbo and Lalonde, 1993; Komaroff, 1997; Raykov, 2001) and work on estimators of reliability that relax assumptions (e.g., Heise, 1970; Joreskog, 1971; Raykov, 1997; Tarkkonen and Vehkalahti, 2005), but not on the absence of the correlated errors assumption.

The two key issues that arise are: (1) to view measurement error not just as a property of the questions and of people, but also as a function of data collection methodology, and (2) to be able to evaluate different data collection designs without the limitations from relying on the CTT-based reliability estimates (which assume no correlated measurement error).

## 2.2. Background on Number of Questions per Page

A common design choice in web surveys is whether to use single questions per page, multiple questions per page, or both formats within the same instrument. There is experimental evidence that by placing multiple items in a page, respondents take significantly less time and yield less item-missing data<sup>10</sup> (Couper et al., 2001). Another web survey experiment also found significantly less time when placing more questions per page, yet also found a higher item-missing data rate (Toepoel, Das and Soest, 2005). In terms of measurement error, literature is largely based on theoretical suppositions (e.g. Gräf, 2002; Dillman and Bowker, 2001), such as placing questions together provides "necessary context." The absence of empirical evidence arises because few studies have experimented with this feature, and there are limited estimates of differences between the layout designs.

Literature on paper SAQ's has shown that when questions are spatially proximate, context effects are stronger (e.g. Schwarz, Strack and Mai, 1991). In the mail survey experiments, it has been assimilation and contrast effects that

---

<sup>10</sup> DK/NA responses



have been predominantly examined by using general-specific questions in varying order (e.g. Schwarz and Hippler, 1995). In the web experiments this has been replicated (Reips, 2002) and also carried over to related specific-specific questions (Couper et al., 2001), where context effects may include cognitive dissonance<sup>11</sup>, i.e. providing consistent answers to related questions. Placing questions together has been found to produce more consistent responses to them (McGuire, 1960) even if the relation between them is only topical (Smith, 1983).

Visual separation of questions has been shown to produce and/or alter context effects in paper questionnaires. Schwarz and Hippler (Sudman et al., 1996, p. 123) experimented with a general and a specific question by placing a question on marital satisfaction and another on life satisfaction either together in a single black frame or in two separate frames (boxes). They found that placing them visually together resulted in contrast effects, indicated by lower inter-item correlation, while separating the questions led to assimilation effects, as it yielded higher inter-item correlation. In a study focusing on context effects in web surveys, Ulf-Dietrich Reips (2002) replicated a paper SAQ experiment by Schwarz and Hippler (1995), where a question on willingness to donate was either preceded or followed by two questions on taxes, but also varied the number of questions per page. However, Reips only looked at means, concluding that multiple questions per page evoke different answers. As was the case with Schwarz and Hippler (Sudman et al., 1996, p. 123), the questions were

---

<sup>11</sup> Defined as "Inconsistency among related beliefs ... produces motivation to do whatever is easiest in order to regain cognitive consistency or consonance among beliefs." In Jones, E.E. and Gerard, H.B. (1967). Foundations of Social Psychology. New York, Wiley, p. 42.

measures of different constructs. In a web survey experiment, Couper, Traugott, and Lamias (2001) applied the method used by Schwarz and Hippler to analysis of context effects for items in scales (multiple items per construct). A total of 11 items in an attitude scale and 5 items in a knowledge scale were placed either as one question per page or multiple questions per page in grids (4, 4, 3, and 5), and the conditions were randomly assigned to 1,602 student respondents. They found higher inter-item correlation (internal consistency) among the questions, when they appeared with other questions on the page. However, they recognized that this method does not reveal which design yields preferable results, but that it merely establishes whether there are response differences. In another web study varying the number of question per page even more, placing 1, 4, 10, or 40 question per page, Toepoel, Das, and van Soest (2005) found not less, but more item-missing data when placing more questions per page. They report slightly higher inter-item correlations that may not be statistically significant, and find the highest Cronbach's  $\alpha$  coefficient when there are multiple questions per page but can be seen at the same time. Together with the finding that there are no differences in means and variances, they conclude that none of the designs elicits more measurement error. They also conduct a second administration of the survey and find no effect on test-retest reliability. The limitations of this study include the small sample size per cell, as in the test-retest analysis there were about 120 respondents in each unique condition, and that the questions used could be suspect in both unidimensionality and discriminating ability (questions such as: "I like to run through fallen leaves," and "Designs and

patterns should be bold and exciting”) as seemingly different yet yielding Cronbach’s  $\alpha$  of .879 and above.

Both Couper, Traugott, and Lamias (2001) and Toepoel, Das, and van Soest (2005) used a grid design when placing multiple questions per page. This can potentially be inducing higher correlations among similar question as in Schwarz and Hippler (Sudman et al., 1996, p. 123) for general and specific questions. However, Bell, Mangione, and Kahn (2001) compared grid versus listed (repeating the response options for each question) designs and found no differences in Cronbach’s  $\alpha$  coefficient for the SF-36 scale, just that grid took 13% less time to complete.

There are different mechanisms that support either layout design. An untested assertion is that placing one question per page results in lack of necessary context, i.e. the respondents may lose their concentration and get frustrated when trying to back up to the previous question (Dillman and Bowker, 2001). If this assertion is true, then the expectation is that measurement error will be higher when placing questions in separate pages. This seems to be an argument that requires a set of very ambiguous questions for which comprehension will be aided if the link between them is emphasized through the layout design. The notion of satisficing<sup>12</sup> was brought to this topic by Dillman, Tortora, and Bowker (1998), claiming that the need to scroll or click for the next screen will increase the likelihood of satisficing. When questions are on separate

---

<sup>12</sup> The satisficing concept was originally presented by Simon, H. A. (1957). Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York, Wiley. According to this notion, people try to minimize psychological costs, searching for solutions that are acceptable, even if not perfect.

pages, respondents have to click through individual pages, while pages with multiple questions would not even require scrolling if just a few questions are in the page. Satisficing was conceptualized in the context of survey research by Jon Krosnick (1991) to have three causes, one of which is respondent burden. Dillman, Tortora, and Bowker selected motory burden, additional mouse movements, as a cause. If either the necessary context or satisficing explanation is true, measurement error should be higher in the single question per page design.

I would like to argue to the contrary, that a single question per page design should yield less measurement error. The counter to the “necessary context” hypothesis is that the inferred similarity from placing the questions on the same page may limit respondents’ true variability across similar attitudes, inducing measurement error in the form of overly consistent responses. Another speculation is that presenting multiple questions in the same page may discourage respondents from exerting the same level of effort in responding through a perception of burden, compared to presenting only one question at a time. There is evidence for one form of burden (for a classification, see Bradburn, 1978), survey length, that presenting questions on the same page reduces actual time yet increases the perceived time to complete the survey (Toepoel et al., 2005). To the extent that grids of questions, are perceived as more burdensome, respondents may not provide as much effort in answering the questions.

These expectations contradict each other and need to be tested empirically. With the absence of an evaluative criterion, scale reliability coefficients like Cronbach's  $\alpha$  are endorsed, leading to decisions of placing multiple items on the same page.

### 2.3. Proposed Method

With this argument, the goal of estimating a coefficient of reliability in the common CTT-based approaches needs to be abandoned temporarily in order to be able to evaluate the impact of instrument design and administration protocol conditions on measurement error. The key should be not trying to estimate the relative proportions of true score variance to true score and measurement error variance, but rather the properties of what we think is true score variance in order to include the impact of both uncorrelated and correlated measurement error. The other change in thinking that is necessitated is that instead of computing reliability after data collection has been completed, experimental manipulations and outcomes need to be designed prior to the study that would allow the evaluation of the impact of various features of the instrument design and data collection protocol on uncorrelated and correlated measurement error combined. As Brennan (2001) notes: "In classical theory, the investigator effectively defines error through specifying a data collection design and adopting the classical test-theory model." Therefore, to address the inability of the current literature to distinguish between less uncorrelated measurement error and more correlated

measurement error for related attitude questions under different designs, a methodology will be employed that allows for different measurement error across items, and utilizes measures external to the experiment for evaluation of the measurement error under different designs.

From the above, it is clear that both uncorrelated and correlated measurement error can affect inter-item correlations. The two are confounded in most measurement designs, especially in the typical settings where reliability is compared between groups, such as males and females. However, if a design is employed in which respondents are randomly assigned to different measurement conditions, an external criterion could be used to evaluate the properties of the true score that has been obtained under each condition. To perform this evaluation, the original confirmatory factor framework proposed by Heise and by Joreskog (Heise, 1969; Heise, 1970; Joreskog, 1970; Joreskog, 1971) can be used, in which items are allowed to have different variances and reliabilities in measuring a construct. However, instead of looking at the reliability estimates which would be biased upward from correlated measurement error, the measured construct itself can be evaluated through its association to another construct with which it should be highly correlated.

To achieve this, a Structural Equation Modeling (SEM) approach is used, the conceptual model for which is shown in Figure 6 with notation from Classical Test Theory. Rewriting the Classical Test Theory model from equation 1 with the same notation, but as a linear regression:

$$x_i = \beta_x X + e_{xi} \quad [6]$$

And analogously for another construct,

$$y_i = \beta_y Y + e_{yi}$$

Where “x” and “y” can be survey questions, “X” and “Y” are the latent constructs that they measure (Figure 6).

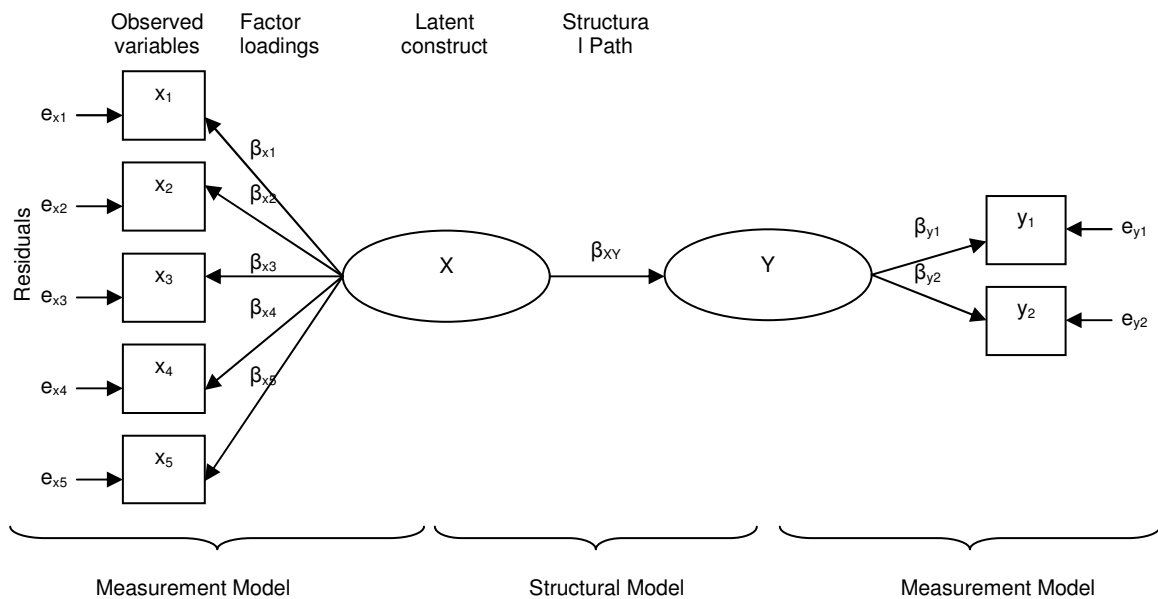


Figure 6. Schematic Representation of the Full (Hybrid) Structural Equation Model Used and Its Components

If construct Y has been measured independently, i.e., having measured  $y_1$  and  $y_2$  through the same design for all experimental conditions and not on the same pages as any  $x_i$ , then a method of evaluation of how well X is measured is through the criterion validity coefficient  $\beta_{xy}$ . If respondents are randomly assigned to different groups, each receiving different measurement conditions to the vector of x’s but the same conditions to the vector of y’s, then comparing the

$\beta_{XY}$  coefficients for the largest estimate would identify the best measurement conditions for the construct of interest, X.

It becomes apparent that this approach relies on finding a criterion construct (Y) for which the relationship with X can be justified and expected to be highest when both X and Y are measured well. The more measurement error is included in X, the more  $\beta_{XY}$  will be attenuated (e.g. Fuller, 1987).

The link between attitudes and behavior has received considerable attention (e.g., Schuman and Johnson, 1976), and has been used and tested in Structural Equation Models; for attitudes towards ecological preservation and self-reported behavior (Milfont and Duckitt, 2004), attitudes and verdicts in studying jury bias (Myers and Lecci, 1998), attitudes towards behaviors and behaviors (Weinstein, 1972), and in political participation (Alwin, 1973; Alwin, 1976).

The attitude-behavior link is useful for theoretical justification of building a two factor model, but statistically the direction does not matter – in a two factor model it does not make a difference whether a factor is labeled an outcome, as the link is a simple bivariate correlation between two latent constructs. More important is building two measurement models for two constructs that are expected to be related. Research in the field of public health provides justification for relationships between ubiquitous concepts that have behavioral outcomes. Self-reported consumption of fruit and vegetables has been examined with linear regression and structural equation models as the outcome from various attitudes and related behaviors, such as fat intake and eating at fast



food restaurants (Neumark-Sztainer et al., 2003; Satia, Galanko and Siega-Riz, 2004) and attitudes towards various aspects of a healthy diet (Neumark-Sztainer et al., 2003), while diet related questions and scales have been used in Confirmatory Factor Analysis before (e.g., Lane, Lane and Matheson, 2004). In fact, the highest loading factor on fruit and vegetables consumption in the Neumark-Sztainer and colleagues model (2003) was Taste Preferences, comprised of attitudes towards fruit, vegetables, healthy, and unhealthy diets. And this was despite the fact that it had lower Cronbach's  $\alpha$  coefficient among their scales.

In order to compare the validity coefficients across experimental groups, invariance for the measurement models for each latent factor has to be established; that the factors have the same structure by having similar regression coefficients with each item across the experimental groups.

I address the number of questions per page design alternative through the use of attitudinal questions in an experimentally varied survey layout design, and a behavioral construct measured elsewhere in the survey to help evaluate the measurement error properties of the different designs.

### 2.3.1. Approach Assumptions

There are two types of assumptions that are made in this approach: Theoretical and Statistical.

It can be argued that there is no true score for something that cannot be directly measured, while psychometric theories such as the CTT and the GT make the true score assumption. Instead of each person having a true score on a construct that a researcher is trying to identify, survey questions can rather be seen as measuring what the question implies to a particular respondent, given the context of measurement and point in time. Under this argument, placing questions in the same page changes their meaning and hence any difference between designs can not be discussed in terms of measurement error. This is a much broader theoretical issue and is beyond the scope of this study.

Another theoretical assumption is that the questions that are used to measure each construct are unidimensional; that they measure a single construct. It is the very methods that are used to establish unidimensionality that are criticized in this study as potentially misleading when their assumptions are not met, but for this purpose reasonable unidimensionality will be established using methods such as internal consistency and factor analysis. In addition, linear regression is used to replicate the major findings without the assumption of unidimensionality of the attitude items.

Moving yet further towards the statistical assumptions, the estimation technique that is employed, the variables are asserted to be measured on a continuous scale, with multivariate normal distributions. The latter is testable, while ordinal SEM is also possible. Unfortunately, the estimation of the ordinal SEM could not converge in this study and results could not be tested for sensitivity to this assumption.

The model in Figure 6 sets all parameters omitted from the figure to zero. This means that the correlations between the measurement errors of the observed indicators for X are not correlated with the errors of the observed indicators for Y. Relaxing this assumption renders the factor covariance inestimable. In the context of the survey experiment this assumption is not so stringent; these correlations could exist, they are only assumed to be equally correlated across experimental conditions.

### 3. Data and Methods

There are multiple steps in the analytic plan of this study:

- Identifying an experimental design that manipulates the layout of questions;
- Verifying equal means across experimental conditions, consistent with prior studies;
- Fitting a measurement model on the variables in the experiment, ensuring reasonable fit of a latent construct;
- Fitting a structural equations model that includes a second factor for evaluating measurement error under each design condition;
- Conducting a multi-group invariance analysis to ascertain that the factor structure is the same and to test whether the validity coefficient is significantly different across conditions;

- Fitting a linear regression to check sensitivity to some model assumptions.

A web survey experiment testing various design features of the data collection method was conducted from April 2 to April 23, 2002, by the University of Michigan and MSInteractive using members from two of Survey Sampling International's web survey panels. A total of 14,282 members were invited, of which 2,871 (20.1%) started and 2,568 (18.0%) completed the survey. Although this is not a probability sample of the general population, it is diverse in characteristics and the focus of this study is on the effects of randomized treatments.

Eight questions measuring attitude towards a healthy diet were asked in the survey, under one of three conditions: one page with all eight questions in a grid, two pages with four questions each (in grids), and eight separate pages. The questions were measured on a seven-point scale with "Agree" and "Disagree" endpoint labels and a "Not Sure" option. Question order and wording are presented in Table 16.<sup>13</sup> Respondents were randomly assigned to one of the three conditions upon reaching the first of the eight questions in the survey, yielding an approximately equal allocation: 877, 897, and 884 in the one, two, and eight pages, respectively. Two questions measuring healthy diet behavior were asked at the beginning of the survey on separate pages (21 questions prior to the attitude questions): "During the past month, how often did you take any VITAMIN OR MINERAL SUPPLEMENTS, such as vitamin pills, wafers, liquid vitamins, or minerals?" and "During the past month, how often did you eat

---

<sup>13</sup> One question is omitted in Table 16, which is explained further in the text.

FRUIT?” Both questions had six response options measuring weekly frequency from “Never” to “Every day.”

Of the 2,871 respondents who logged in to start the survey, 304 abandoned before the end, of which 152 left the survey prior to answering the first question. Only 9 respondents broke-off on one of the eight questions, and they were spread between the three conditions. However, treating the “Not Sure” response category as unit nonresponse yielded uneven loss in complete cases for the attitude and behavior questions in the three experimental conditions: 113, 78, and 60 with final sample sizes of 764, 819, and 824 in the one, two, and eight pages designs, respectively.

The eight items were not created for use in a latent construct, hence ensuring unidimensionality here is needed. The seventh of the eight questions, “My lifestyle makes it difficult to eat right,” was intentionally worded as a behavior rather than an indicator of an attitude. This was also apparent in the data by higher values of coefficient alpha when it is removed and lower absolute values of bivariate correlations (the mean correlation for this question with the other questions was  $|.13|$ , while the mean correlation among the rest of the items was  $|.38|$ ). Exploratory factor analysis also clearly showed that this question was measuring a second factor by itself – a necessary step as the questions were not designed for this type of analysis. Since analyses in this study necessitate unidimensionality of the scale to identify a single latent construct, this question was omitted. The group sample sizes increased slightly to 782, 832, and 832 complete cases in the one-, two-, and seven-page designs.

SEM provides a model-based approach to analyzing multivariate associations among variables, as well as latent structures, but as noted earlier it also involves necessary assumptions. The data were first examined for normality and using the criterion of an absolute value of 2, severe skewness and kurtosis were not found (highest skewness=0.67, kurtosis=1.66, relative multivariate normality=1.05 [LISREL 8.53]). To the extent plausible for testing, linearity of associations and homoscedasticity were not rejected, yet they are difficult to affirm in these data as the variables are measured on either six or seven point scales (commonly used as continuous variables in SEM). Unlike in methods such as multiple regression, measurement errors and correlations between them do not have to be assumed to be zero and are estimated.

Following the test for bias, the focus in the primary analyses is on differences in the covariance structures between the three conditions. The expectation is that items will have higher associations among each other when placed on the same page, as reported by prior studies, but we will continue to a substantive analysis of the properties of the covariances to evaluate these differences. By placing questions on the same page, it is hypothesized, enhanced context effects and/or perceived burden will worsen the identification of the constructs and hence decrease the association between the derived latent constructs (also known as validity coefficients), by introducing excessive response consistency, i.e. less variation associated with other related constructs. To the contrary, the major criticism against placing questions on separate pages is that measurement error will be induced through lack of necessary context,

hence the analytical model has to account for measurement error. The latent constructs, the strength of associations between them, and the measurement errors of their indicators will be simultaneously estimated using SEM.

Furthermore, model fit can be compared between groups to test specific parameter estimates for equality, e.g. whether the measurement errors in the three experimental groups are invariant.

First, a confirmatory factor analysis is performed on each experimental group to ensure that the seven measures of attitude towards a healthy diet fit a model in which they are indicators of one latent construct, as shown in the left half of Figure 6. Dietary attitude has been successfully modeled as a cause of dietary behavior in path analysis (e.g. Lewis, Sims and Shannon, 1989) and based on this association the full structural model presented in Figure 6 depicts this causal relationship between the attitude and behavior measurement models. In this framework, if placing items on separate pages improves data quality in terms of covariance properties, the path between the attitude and the behavior constructs in the structural part of the model will be higher for the two-page than the one-page design, and highest in the seven-page design. This is supported by the prior finding that multiple items on a page lead to higher inter-item consistency, and the current assertion that this consistency leads to degraded measures of a construct, hence associations between constructs will be attenuated. It is also likely to be reflected by model fit – improved quality of data should be reflected in better model fit (by indices such as the Standardized Root

Mean Square Residual, which is a measure of how well the observed covariances are reproduced by the model, i.e. the relative size of the residuals).

Regardless of whether the hypothesized difference is found, multi-group invariance analysis is performed to determine which parameters, if any, are significantly different between the experimental conditions. Using partitioning of the Chi-square test statistic ( $\chi^2$  difference tests), the three groups can be sequentially tested for differences in factor loadings, factor covariances, and measurement errors<sup>14</sup>. While the residuals of the observed variables and possibly the factor covariance (measurement invariance models use non-causal relationships between constructs, instead of paths) are expected to differ, the factor loadings of the seven attitude questions should remain invariant if they are indicators of the same latent construct in all three conditions. The contrary would imply that attitude towards a healthy diet attains a different substantive meaning under the three conditions and hence the path coefficients could not be meaningfully compared. That is, if the constructs are inherently different based on the layout design of the survey, not just in terms of variance/covariance properties, then comparing associations between them should be substituted by some other method of evaluation. The differences in estimates of the structural paths across the groups is examined in the full structural models (Figure 6), while this set of tests (invariance analysis) explores which parameters are in fact

---

<sup>14</sup> In this terminology, measurement error variance is also labeled residuals of the observed variables, residual variances, and disturbance terms, while the questions are also named observed variables, indicators, exogenous variables, and items, due to the difference in terminology and intended interpretations in SEM across disciplines.



significantly different between the three conditions. Figure 7 shows the measurement model used in the multi-group analysis.

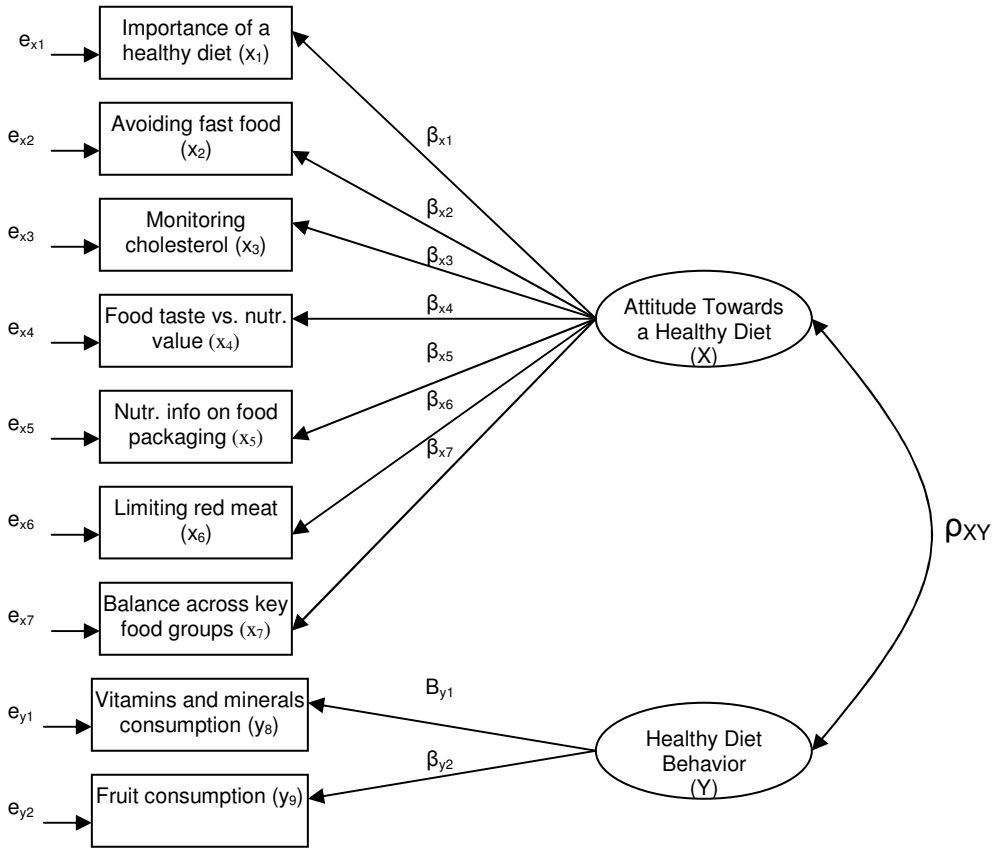


Figure 7. Measurement Model in the Multi-group Analysis

The models in Figure 6 and Figure 7 are functionally identical as in a two-factor model the structural path is simply a symmetrical covariance, or a correlation when the variances of the factors are set to 1. Figure 6 simply depicts how these models could be built into more complex structural equation models if additional factors are added.

In order to relax some of the model assumptions, as the latent factor structure imposes assumptions made in CTT by identifying the true score

through the common variance across similar items, a linear regression model is tested. Regressing the behavior items, which are combined to form a summative index, on the attitudinal questions for each layout condition can examine which condition leads to the highest proportion of variance explained in the behavioral items. In this model, the expected values of the behavior index are the same across conditions as they were not manipulated, and of interest is under which condition the (experimentally manipulated) predictors explain the largest proportion of variance.

#### 4. Results

The means for each of the seven attitudinal questions were compared across the three conditions to test for differential bias, as an indicator of contrast and assimilation context effects. As presented in Table 16, none of the 21 contrasts (7 questions x 3 conditions) was statistically significant at  $\alpha=.05$ , using Holm's Procedure to control for Type I error by sequentially adjusting the nominal  $\alpha$  for multiple comparisons (Holm, 1979). These results are in parallel to the findings by Couper et al. (2001) and Toepoel et al. (2005) that the means do not differ across the designs.

Table 16. Tests for Differences in Means between Groups for Each Question

Question	Contrast Between Groups	Means	Difference Between Means	Exact Significance (2-tailed)	Holm's Alpha
1: Maintaining a healthy diet is a priority in my life	1 page vs. 7 pages	3.08 vs. 3.19	-0.107	0.193	0.004
	1 page vs. 2 pages	3.08 vs. 3.13	-0.045	0.579	0.008
	2 pages vs. 7 pages	3.13 vs. 3.19	-0.061	0.447	0.006
2: I avoid fast food because it's not healthy	1 page vs. 7 pages	4.35 vs. 4.13	0.227	0.020	0.002
	1 page vs. 2 pages	4.35 vs. 4.17	0.188	0.054	0.003
	2 pages vs. 7 pages	4.17 vs. 4.13	0.038	0.689	0.010
3: I monitor my cholesterol level closely	1 page vs. 7 pages	4.90 vs. 5.06	-0.160	0.100	0.003
	1 page vs. 2 pages	4.90 vs. 4.96	-0.060	0.536	0.007
	2 pages vs. 7 pages	4.96 vs. 5.06	-0.100	0.297	0.005
4: How food tastes is more important to me than its nutritional value	1 page vs. 7 pages	3.81 vs. 3.71	0.103	0.241	0.004
	1 page vs. 2 pages	3.81 vs. 3.84	-0.026	0.771	0.013
	2 pages vs. 7 pages	3.84 vs. 3.71	0.129	0.137	0.004
5: I pay attention to nutritional information on food packaging	1 page vs. 7 pages	3.32 vs. 3.12	0.200	0.031	0.003
	1 page vs. 2 pages	3.32 vs. 3.14	0.175	0.057	0.003
	2 pages vs. 7 pages	3.14 vs. 3.12	0.024	0.791	0.017
6: I limit the amount of red meat in my diet	1 page vs. 7 pages	4.09 vs. 3.89	0.209	0.047	0.003
	1 page vs. 2 pages	4.09 vs. 3.86	0.235	0.025	0.003
	2 pages vs. 7 pages	3.86 vs. 3.89	-0.026	0.798	0.025
7: I try to balance my diet across the key food groups	1 page vs. 7 pages	3.68 vs. 3.59	0.094	0.285	0.005
	1 page vs. 2 pages	3.68 vs. 3.60	0.079	0.366	0.006
	2 pages vs. 7 pages	3.60 vs. 3.59	0.014	0.867	0.050

1 page - all questions on one page (n=782)

2 pages - questions 1-4 on first page and questions 5-7 on second page (n=832)

7 pages - all questions on seven separate pages (n=832)

Table 17 presents differences in the magnitudes of linear associations between groups. Comparing the covariances between the last question and each of the other questions (last column in each condition in Table 17), their values consistently decrease as the questions are separated on different pages (e.g. covariance between questions 1 and 7: 1.800 under the single-page design is greater than 1.649 in the two-page condition, which in turn is greater than 1.402 in the seven-page condition; covariance between questions 2 and 7: 1.554 is greater than 1.490, which in turn is greater than 1.281). In other words, responses to a question are more consistent with responses to other questions

as multiple questions per page designs are used (also confirmed by inter-item correlations – Cronbach’s  $\alpha$ ). Although this is evidence that there are context effects arising from placing questions together, as found in prior studies, it only motivates further analysis to determine which design is preferable. Furthermore, Box’s M test revealed a statistically significant difference between the variance-covariance matrices ( $p < .05$ ), supporting the need to model these differences.

Table 17. Covariances for All Questions under All Conditions

Condition	Question	2	3	4	5	6	7
Single page (n=782)	1	1.672	1.440	-1.014	1.746	1.284	1.800
	2		1.905	-1.013	1.576	1.682	1.554
	3			-0.900	1.661	1.634	1.560
	4				-0.797	-0.735	-1.041
	5					1.900	1.880
	6						1.561
Two pages (n=832)	1	1.652	1.395	-1.160	1.839	1.494	1.649
	2		1.499	-1.147	1.678	1.520	1.490
	3			-0.813	1.340	1.421	1.239
	4				-1.222	-1.189	-1.062
	5					1.769	1.807
	6						1.536
Seven pages (n=832)	1	1.368	1.073	-0.995	1.365	1.264	1.402
	2		1.546	-1.123	1.346	1.484	1.281
	3			-0.917	1.217	1.526	1.113
	4				-1.029	-1.024	-0.948
	5					1.474	1.348
	6						1.188

Box's M test for equal variance-covariance matrices  $p < 0.5$

Adequate fit of the measurement model was established for the attitude factor in each of the three conditions, the full structural model presented in Figure 6 was fitted to each group separately and the results are presented in Table 18. All models were tested using Maximum Likelihood estimation in LISREL 8.53 (Jöreskog and Sörbom, 1996).

Table 18. Full Structural Model Path Coefficients and Model Fit Indices by Experimental Condition

Design:	1 Page	2 Pages	7 Pages
Standardized Path Estimate between Attitude towards Healthy Diet and Healthy Diet Behavior Constructs	0.82	0.90	0.94
Minimum Fit Function Chi-square (26df)	105.94	91.99	66.74
Non-normed Fit Index (NNFI)	0.97	0.98	0.98
Comparative Fit Index (CFI)	0.98	0.98	0.99
Root Mean Square Error of Approximation (RMSEA)	0.063	0.054	0.044
Standardized Root Mean Square Residual (SRMR)	0.033	0.031	0.027

As described earlier, the expectation is that better quality data will allow better identification of latent constructs, hence related constructs will have stronger associations between each other. The magnitude of the association between the Attitude towards Healthy Diet and the Healthy Diet Behavior constructs was found to be consistently higher as questions are separated on different pages, with the standardized path coefficient (validity coefficient) increasing from 0.82 to 0.94 (Table 18). In the current framework this supports a recommendation of placing related attitude questions on separate pages as yielding less measurement error, at the possible cost of longer duration of the survey.

The other method of evaluating quality of data in this study is model fit. The  $\chi^2$  estimates are provided for comparison of model fit across the conditions but are not to be used as a measure of exact significance due to the  $\chi^2$  statistic sensitivity to sample size. Instead, fit indices are used to examine different aspects of model fit, as well as to adjust for parsimony and sample size. Formulas for the employed fit indices are provided in the Appendix. Both the

Non-normed Fit Index (NNFI) and Comparative Fit Index (CFI) contrast the difference between the actual model  $\chi^2$  estimate to the independence model<sup>15</sup>  $\chi^2$ , adjusting for the number of estimated parameters, and indicate a good model fit (>0.95). The Root Mean Square Error of Approximation (RMSEA) does not use a null model for comparison, but is based only on the model  $\chi^2$  estimate, adjusted for the number of parameters and sample size, and shows an acceptable model fit for the one-page design and good fit for the two- and seven-page conditions (acceptable fit is <0.10 and good fit is <0.06, Hu and Bentler, 1999). Of more importance here is the Standardized Root Mean Square Residual (SRMR) as this measure is the standardized difference between the observed covariance and predicted covariance matrix – it places an additional emphasis on the residuals (measurement error) in the evaluation. Although all three groups exhibit good model fit (<0.08, where 0.00 indicates perfectly reproduced covariances), it is apparent that model fit improves with the use of multiple page designs, i.e. the two-page design better reproduces the covariances than the one-page design, and the seven-page design reproduces the relationships better than either of the other designs. Hence differences were consistent (fit improved with increasing number of pages) and were even large enough, that RMSEA did not yield a good fit under the one-page condition.

In order to determine whether it is justifiable to compare these models and to determine which parameters are causing significant differences, if any, a multi-group analysis was conducted. Such invariance analysis uses the equivalent

---

<sup>15</sup> In the independence model (null model), all the covariances are zero.

CFA model where all latent factors are allowed to covary, represented by the curved arrow between Attitude and Behavior in Figure 7.

The model is then simultaneously applied to the covariance matrices of the groups in the experiment, estimating common parameters. In order to determine whether the groups differ significantly in a particular way (e.g. measurement error), separate sets of parameters are allowed to be estimated independently for each group and the multi-group model is re-estimated. The partitioning property of the  $\chi^2$  statistic is used by testing the difference between the  $\chi^2$  estimates of the two models with degrees of freedom equal to the number of parameters allowed to vary across groups (i.e. the difference between the degrees of freedom in the two models). The procedure could be viewed as a test of difference between nested models, where the difference in  $\chi^2$  and the difference in degrees of freedom are used for the new  $\chi^2$  test. A difference below the  $\alpha$ -level (of .05)<sup>16</sup> means that the model fit is significantly improved, i.e. that the “freed” parameters are not the same in the three groups. A common sequence of freeing parameters is to (1) start with all parameters invariant, (2) free factor loadings, (3) free factor covariances, and (4) free measurement error terms. In this analysis the factor loadings and measurement errors were allowed to vary only in the attitude measurement model as the experimental condition involved only the seven attitude questions. The analysis was repeated also by allowing the parameters in the behavior model to vary and the results were very similar, drawing the same conclusions.

---

<sup>16</sup> This method does not adjust the nominal alpha level for the multiple tests.

The results in Table 19 showed no significant difference in factor loadings, asserting that it is the same Attitude Towards Healthy Diet latent construct being measured in the three groups - placing the questions together on a page or on separate pages does not change the attitude that they measure. The difference in the covariance between the two factors was also found not significant. Since in the full structural models the path estimates were quite different (0.82, 0.90, and 0.94), it is possible that the covariance terms were not found significantly different as there are only two observed variables measuring Healthy Diet Behavior, i.e. the Behavior part of the model comprises a very small part of the covariance matrix (hence fewer covariances to be reproduced) and has a smaller impact on model fit. Alternatively, the factor covariance may become significant only after measurement errors are allowed to be different, which is the hypothesized cause. The difference in measurement errors between the three groups was found to be significant, as was indicated earlier by the SRMR estimates in the full structural models. This supports that it is indeed measurement error that is induced by placing questions on the same page and hence leads to worse model fit.



Table 19. Multi-group Invariance Analysis Using  $\chi^2$  Difference Tests

Model	$\chi^2$	d.f.	Change in $\chi^2$	Change in d.f.	P-value
All parameters invariant across groups	309.72	116			
Difference test			22.97	14	0.061
Allowing <b>factor loadings</b> to vary	286.75	102			
Difference test			1.83	2	0.401
Allowing factor loadings and <b>factor covariances</b> to vary	284.92	100			
Difference test			51.17	15	0.000
Allowing factor loadings, factor covariances, and <b>measurement errors</b> to vary	233.75	85			

Finally, to test model sensitivity to some of the assumptions, in particular assumptions requiring a latent factor structure to fit well and to be the same across conditions, a linear regression model was applied. The summative index of the two behavioral validation questions was regressed on the seven attitudinal questions, separately for each of the three layout conditions. The variance explained in the behavior index was highest for the one question per page condition (Adjusted  $R^2=.218$ ), then for the four questions per page (Adjusted  $R^2=.195$ ), and lowest for the eight question in the same page (Adjusted  $R^2=.186$ ).

## 5. Discussion and Conclusions

Consistent with previous findings, placing more questions per page increases inter-item correlations, yielding higher reliability estimates under Classical Test Theory. Under the proposed model, rather than lower measurement error, this shows to be the result of higher correlated measurement error.

First, a method was used that overcomes the typical assumption of no correlated measurement errors when evaluating data from attitudinal scales. It does not provide an unbiased estimate of the correlated measurement error, but the effect of uncorrelated and correlated errors is evaluated together.

Second, the strength of experimentation to help identify the specific causes was demonstrated in the context of measurement error where no gold standard is available. Unlike current literature on Reliability Generalization (Vacha-Haase, 1998), the same population could be used to test different conditions, and unlike past literature, the same items and at a single point in time can be used (Spearman, 1904; Guttman, 1953). This approach does require the researcher to select the design features that are believed to or have been found to affect measurement error the most, and to incorporate validation measures that are expected to be highly related to the manipulated scale in a manner that would not produce correlated error between the two constructs.

Thirdly, a test was performed comparing survey design conditions without the assumption of no correlated error. The results showed no significant differences in the means of the responses to the attitudinal questions by number of questions per page were found. Covariance matrices were significantly different, in a manner consistent with prior literature – covariances were higher when more questions are presented on the same page, reinforcing the need for a covariance approach to data quality evaluation. Using a latent variable approach, a theoretically simple model was fitted to the data. Placing the related attitudinal questions on separate pages produced data with preferable properties,

as relationships between the latent constructs in the structural model were even stronger and overall model fit was improved, indicated by better reproduction of the observed covariances (e.g. RMSEA was 0.063, 0.054, 0.044 in the one-page, two-page, and seven-page design, respectively). Multi-group invariance analysis revealed that the attitude questions were measuring the same latent constructs under all three conditions, but the designs differed significantly in terms of measurement error.

It can be deduced from these results that the processes that occur with multiple questions per screen generate additional measurement error. These processes could be a product of short-term memory (fewer intervening events like clicking “Next” buttons and changing pages) and excessive response consistency to related items (avoiding incongruent responses) resulting in enhanced context effects. Support was not found for the satisficing hypothesis forwarded by Dillman et al. (1993) to this problem as the single question per page yielded the least measurement error despite the need for additional clicking through pages. Similarly, no support was found for the claim that *necessary* context effects will be lost when questions are placed on multiple pages (Dillman and Bowker 2001) – on the contrary, indices like SRMR detected less residual variance when separating questions on different pages.

Rather, the findings are consistent with the expectations under the perceived burden hypothesis – placing questions on the same page (in a grid) not only creates the perception of a longer duration of the survey (Toepoel et al., 2005), but also has adverse effects on the measurement properties of the data.

There are processes such as context effects that differentially affect data collected via single and multiple questions per page web-based survey designs, far beyond simple reliability estimates and bivariate associations. The example presented utilized only two latent constructs and showed significant deterioration of measurement properties when placing multiple questions per page, under the model. Yet if both factors were measured by multiple questions per page designs, the substantive associations could be again underestimated or even overestimated as the increased measurement errors become correlated between the factors (measurement error is induced by the same process). Hence there is a detrimental effect on the data and the results could be unpredictable.

There are two causes of context effects that could be separated in a future experiment. First is the effect of the time delay from clicking on the “Next” button to the displaying of the following page, as information quickly dissipates from short-term memory. The other is the visual separation of the questions, i.e., do context effects arise from simply being able to see the questions together or also from the format – matrix versus separate questions, all on the same page.

This is only the first step – evaluating different layout designs and identifying the preferable conditions. The causes for these measurement differences have still to be identified in order to anticipate the optimal survey designs for any given population, method of data collection, etc.

## 6. Appendix

$$CFI = \frac{d(\text{null model}) - d(\text{proposed model})}{d(\text{null model})}, \text{ where } d = \chi^2 - df$$

$$NNFI = \frac{d(\text{null model}) - d(\text{proposed model})}{d(\text{null model}) - 1}, \text{ where } d = \chi^2 / df$$

$$RMSEA = \sqrt{\frac{\chi^2 - df}{(n-1)df}}$$

SRMR is the square root of the averaged standardized squared residuals or differences between observed and predicted variances and covariances.

For a complete computational description, see the LISREL© web page at:

<http://www.ssicentral.com/lisrel/fitmeas.htm>.

## Chapter 5. Causes of Correlated Measurement Error by Layout Design

### 1. Introduction

Literature on context effects has demonstrated that placing related attitude questions together could affect responses, typically increasing inter-item correlations (Couper et al., 2001; Toepoel et al., 2005). The previous chapter employed a structural equation model approach to address this, testing a model in which the higher inter-item correlations are the result of higher correlated measurement error rather than less measurement error.

However, the causal mechanism producing these differences remains unknown. While the layout choice is only one example of how survey design can affect correlated measurement error, understanding the causes in this design alternative through experimentation can help predict effects for other design changes.

There are different mechanisms that could be producing the correlated measurement error. First, proximate questions may be *perceived* as being multiple measures of the same construct, inducing more similar interpretations of the items. Second, when no actions are needed to get to the next question, the same material may be retrieved from *working memory* in answering all the

questions. Third, respondents may be affected by the *perceived burden* and pay less attention to the individual questions when they are presented on the same page and in the same grid. These hypotheses are tested in an experiment and results presented.

## 2. Background and Significance

The effects of different possible questionnaire designs of the same survey have been recognized for a long time (e.g., Bradburn and Mason, 1964; Dillman, 1978; Sudman and Bradburn, 1982; Sanchez, 1992; Jenkins and Dillman, 1995; Smith, 1995). Some design features have somewhat straight-forward explanations, such as the use of arrows for skip instructions for reducing errors of omission and commission (Dillman, 1978), and the order of general and specific questions resulting in contrast and assimilation effects (Schwarz et al., 1991). Other design features, such as the number of questions presented per page, may have multiple causes for differences across designs.

The number and order of questions on a page is a common design decision that needs to be made in self-administered surveys, and especially in web surveys. This decision can cause differences in associations between related questions have been found when varying the number and layout of questions per page (e.g. Couper et al., 2001; Fuchs, 2001), with associations being higher when the questions are presented together in a grid. But even after differences are found, it is difficult to evaluate which design is preferable.

Chapter 4 used a model-based approach to evaluate which design is preferable, concluding that placing the questions together on the same page yields higher inter-item correlated measurement error, and separating the questions on different pages provides data with better predictive ability.

While this finding helps future survey designs facing the same dilemma, not understanding why these differences arise limits the ability to extrapolate this knowledge to different circumstances and to other design decisions. For example, if the questions are placed together in the same page but not in a grid, would there be adverse effects on the data as found in the grid design, compared to different pages? The answer to this question affects even more extreme web survey design decisions, such as whether to place the entire survey in a single page. Should we expect that the results are worst in this design, or about the same as in a paging design as long as grids are not used? Experimental studies of paging vs. scrolling web survey design have found no implications for unit nonresponse (Vehovar, Manfreda and Batagelj, 2000; Peytchev et al., 2004), making measurement error an even more decisive factor. Understanding the mechanisms creating higher correlated measurement error when questions are in a grid on the same page could also provide methods for removing this effect by other means.

There are two types of explanations for effects of the number of questions presented together, found in past literature. One uses context effects, brought into survey methodology by psychologists, and the explanation speaks to how questions can be interpreted differently and how different information can be



made available by their presentation together. Context effects can include many hypotheses and outcomes, but can be parsed out in the framework of the survey response process model (Cannell et al., 1977; Cannell et al., 1981; Tourangeau, 1984; Strack and Martin, 1987; 1987), as effects on question comprehension and information retrieval.

The other type of explanation involves the perceived size and difficulty of the task, which have been referred to as “respondent burden.” Hence both approaches are multidimensional. An objective goal is to first identify which type of explanation is dominant, and then to tease apart the more specific causes.

### 2.1. Context Effects: Hypotheses on Comprehension and Retrieval

In paper self-administered questionnaires (SAQ's), context effects have been found to be stronger when questions are proximate (e.g. Schwarz et al., 1991). This is measured by first asking a general and then a specific question (including the response from the general into the specific, into an assimilation effect) and in another group asking the specific first (excluding the response to the specific question from the general, into a contrast effect). These effects are identified by finding lower correlations between the two items in the latter case (e.g. Schwarz and Hippler, 1995). Schwarz and Hippler (Sudman et al., 1996, p.123) compared the effect of the order of a question on satisfaction with life and another on satisfaction with marriage on correlations, by varying whether the questions are in a single box or separated in two boxes. They found that when

questions were presented visually together, thus emphasizing their relatedness, resulted in contrast effects, while separating them resulted in assimilation effects. That is, the questions are comprehended differently depending on their visual presentation – when visually boxed together, apparently meaning: “Apart from your marital satisfaction, how satisfied are you with life?” However, when not boxed together, it is likely retrieval that comes into play as information retrieved about marital satisfaction is more easily accessed when responding to life satisfaction, increasing instead of decreasing the correlation between the two questions.

In a web survey replication of this experiment, where the questions were separated by placing them in different pages, Fuchs (2001) found higher context effects in the multiple questions per page. Here context effects were operationalized as the difference between contrast and assimilation effects, and found them to be larger when the questions were in the same page. Ulf-Dietrich Reips (2002) replicated this experiment with a different set of questions in a web survey, but the analytic method allowed only the reporting on a statistically significant difference. A web experiment conducted by Couper, Traugott, and Lamias (2001) extended to specific-specific questions by placing multiple questions per construct together or in separate pages. They found that placing sets of questions in the same screen instead of separately, results in higher inter-item correlations. The authors recognized, however, that while establishing that associations between questions change as a result of manipulating the questionnaire design, it does not allow any guidance on which is preferable. This

could be achieved if the cognitive causes for these differences are revealed and also linked to evaluative criteria.

If placing similar questions together in the same page increases their inter-item correlations because relevant information has been retrieved for a previous question on the page and is still more available, then benefits from this design can be anticipated for questions that require a lot of information retrieval. Yet if placing questions together increases the perception of the questions being related, i.e., they are comprehended as asking more similar things, this design may limit the amount of information captured by the different questions. While these two hypotheses predict the same outcome on inter-item correlations, it is apparent that the exact mechanism causing this outcome has great implications for when it is desirable and when it is not.

To help answer these questions, two sets of hypotheses are formalized:

*Perception Hypothesis:*

Deducing from the observed phenomenon in a mail survey experiment (Schwarz and Hippler in Sudman et al., 1996, p.123), it can be argued that it is because questions are perceived as being more related when they are visually bound together in the same box (i.e. placed in a grid), and that effect will diminish when they are listed separately still on the same page. If this mechanism is true, giving respondents instructions that the questions measure the same construct

should reduce the effect of separating them (i.e., listing them instead of placing them in a grid, while still in the same page).

Hypotheses if a mechanism of *perception* is dominant:

H<sub>1</sub>: Under no instructions, placing questions in a grid on the same page should produce stronger context effects than either placing them in separate pages or in the same page but as separate questions.

H<sub>2</sub>: Asking respondents to treat the questions as independent when they are in a grid should decrease context effects and become comparable to the other groups (comparable covariance structure).

H<sub>3</sub>: When questions are on separate pages with explicit instructions to treat the questions as related, the results should be closer to the grid condition (no instructions).

H<sub>4</sub>: Associations between questions from separate constructs will increase when the questions are intermixed.

*Working Memory Hypothesis:*

Alternatively, it may not be the perceived relation between items that affects measurement, but it is working memory – placing the questions in

different screens, so that answering each question involves clicking on navigation buttons, page loading from the server, i.e., more time and actions between questions. Residues from explicit retrieval, creating recency effects, or from implicit retrieval lead to priming (Baddeley and Hitch, 1993). It would be expected under these processes that the fewer intervening cognitive and motory tasks and less time lapse, the stronger the effect of the retrieval residues. This should also be evident in associations between questions when they can be read without having to scroll and those when scrolling is necessary. If this is the main causal mechanism, implications might be drawn for deciding whether to use a design such as a paging survey with only single questions per page vs. having the entire questionnaire in a single scrolling page.

Hypotheses if mechanism of *short-term memory* is dominant:

H<sub>1</sub>: Single question per page should be the most different from the other layout designs.

H<sub>2</sub>: Placing questions in a grid or in a list in the same page should lead to comparable results.

H<sub>3</sub>: Instructions to treat the questions as related in the single question per page condition should not be different from having no instructions.

H<sub>4</sub>: Instructions to treat the questions as independent in the grid condition should not be different from having no instructions in the grid layout.

H<sub>5</sub>: Associations between questions from separate constructs will NOT increase when the questions are intermixed.

## 2.2. Burden: Hypotheses on Perceived and Experienced, Amount and Type of Demand

Respondents may not only comprehend and retrieve information differently based on survey design, but that their approach to the survey tasks can also be altered. That is, the more we demand from respondents, the less likely it will be that respondents will dedicate resources to the survey. This is a rather general statement, as the amount of the demand can be operationalized in various ways, and the resource commitment by the respondents can also be viewed in very different ways, from whether they respond to the survey to how they respond to survey questions.

One possible classification of respondent burden is that posed by Bradburn (1978), which presented it as four components: Length, Respondent effort, Respondent stress (i.e., sensitive and threatening content), and Frequency of being interviewed. From the survey response process model perspective, effort and stress can all be viewed as the ***type of burden***, i.e., questions that are more demanding in terms of comprehension, retrieval, and in the case of stress,

in terms of judgment processes. Different questions place varying types of demands on different respondents.

The length of the survey can be viewed as contributing to the **amount of burden**. While this dimension seems straightforward, some mixed results have been found on the effect of the length of the survey because it has been poorly defined. In a review on the effect of questionnaire length on response rates, Bogen (1996) finds that while a few experimental studies support that longer questionnaires decrease response rates, another study supports the opposite. The latter study varied length not by adding questions, but by adding more pages. We should consider using the term *amount of burden* to keep operationalizations of burden such as questionnaire length relevant to the dosing of the response task.

What we observe and can possibly code about the questions and questionnaires is not necessarily what is interpreted by the respondents. For example, two surveys with the exact same type of questions (i.e., type of burden) and same number of questions (i.e., amount of burden) can achieve very different response rates. Survey methodologists have long recommended the use of white space in self-administered paper surveys “to make the questionnaire look easier” (Sudman and Bradburn, 1982, p.239). In an early experiment laying out the exact same questions on 3, 6, and 9 pages yielded the highest response rates in the design with the most white space (Champion and Sear, 1969). None of the questions are altered, nor are any questions dropped, yet formatting the questionnaire with more white space we expect to lower the **perceived burden**

by the respondents. Because it is a difficult dimension of burden to quantify, it has been used in the literature but never called “burden.”

Finally, arguments have been posed that when placing questions later in the questionnaire may lead to respondents answering the questions “in a more perfunctory manner” (Bradburn and Mason, 1964, p.58). While this has been called a “fatigue effect,” it is implied to be a function of the other types of respondent burden to a particular point in the survey, or ***experienced burden***. It may also interact with the other dimensions of burden – a question requiring a lot of recall may be found more burdensome late in the survey. Similarly, a respondent may find a survey more burdensome if he has recently responded to a different survey request. The fourth factor suggested by Bradburn (1978), the frequency of being interviewed, would also fall here.

The difficulty in studying respondent burden is not just from its multidimensional nature, but also from the outcomes used to evaluate it. Not surprisingly, measures and effects of respondent burden have been driven by the specific purpose of evaluating it. One motivation came legislatively, when in 1976 the Commission on Paperwork Reduction was formed and the “consequences of this across the board reduction fell disproportionately on sample surveys undertaken for research and evaluation purposes” (Bradburn, 1978). The other motivation comes from the researchers’ genuine concern to minimize respondent burden as they fear that it adversely affects response rates and the responses provided by the respondents. Most of the literature on burden focuses on how it affects response rates, and specifically how questionnaire



length affects response rates (for a review, see Bogen, 1996; also Treat, 1996; Apodaca, Lea and Edwards, 1998; Galvin, Sae-Ung and King, 2000) – in the current framework, the effect of the *amount of burden* on unit-nonresponse. Statistical methods were being developed on reducing the length of the questionnaires also as early as the 1970's (Ford, Hocking and Coleman, 1978; Raghunathan and Grizzle, 1995; Ghosh and Vogt, 2000). Although unmeasurable except in the case of sampling by the same agency, *experienced burden* indicated by repeated surveying of the same sample elements also received attention, with development of sampling methods that minimize respondent burden through disproportionate sampling (Tortora and Crank, 1978; Creel et al., 2002), but this has limited implications for the majority of sample surveys.

The effect of burden has also been studied on response quality. One would expect that asking a respondent to recall the occurrence of a behavior for an unreasonably long past period would place unrealistic demands on recall leading to measurement error. However, simply asking more questions on the same topics and later in the survey has also been found to affect responses, indicated by more identical responses to different questions (Herzog and Bachman, 1981). Later in the survey means that the respondents should have experienced more cumulative burden, while more questions in each set of questions could be related to perceived burden, depending on how the questions were presented.

While in a particular study there may not be any freedom in changing the questions, constraining the type and the amount of burden, there is typically more room for changes in the layout and order of the questionnaire – possibly affecting the perceived and the experienced burden. More specifically, would changing the design by placing questions on separate pages rather than on the same page decrease the perceived burden by chunking the task into smaller pieces, and would placing the questions later in the survey affect responses through burden experienced to that point. Continuing this logic, placing questions on the same page should have an even greater detrimental effect when they come late in the survey as burden is cumulated for the respondent.

While researchers in computer-human interaction share the design decision of number of tasks per page, it is still left to speculation, such as: “Tasks that seem well-suited for this approach [separating tasks on different pages] tend to be either branched or very long and tedious” (Tidwell, 2005). While survey practitioners do use paging designs when questions depend on answers to previous questions, paging designs to avoid perceptions of burden have neither been proposed nor tested. These are very different explanations compared to those based on the survey response process model.

One relevant study randomly assigned 1, 4, 10, or 40 questions per page (Toepoel et al., 2005). They found that placing questions on the same page took respondents less time, consistent with other studies, but that respondents reported that it took more time than those assigned to fewer questions per page.

One likely explanation is that respondents found the multiple questions per page (in a grid) to seem more burdensome regardless that it took them less time.

*Perceived Burden Hypothesis:*

The third competing mechanism for the effect of placing questions in the same page is that of perceived burden. That is, the task should seem less burdensome when the respondent is presented a single question in the page to answer, as opposed to multiple questions, and this should seem even more demanding when a grid of questions and response options are presented. In terms of measurement error, the effect of perceived burden should be much larger later in the survey as respondents would have experienced more burden by that point – hence varying the placement of the questions within the questionnaire would detect this cause. In addition, if respondents feel more burdened, there should be less evidence of effortful responding – more infrequent changing of responses, more item-missing data, and taking less time on the questions. In order to separate this potential cause from the perception of the questions being related, explicit instructions could be used to tell the respondent to treat the questions as unrelated and see if respondents select the same responses at the same rate.

Hypotheses if mechanism of *perceived burden* is dominant:

H<sub>1</sub>: The effect of placing the questions on the same page, and further into a grid is from perceived burden, it should interact with the effect of location in the questionnaire as a form of experienced burden.

H<sub>2</sub>: Compared to presenting questions separately, respondents will take less time to answer the questions than when questions are placed together in a list, and even less time when presented in a grid.

H<sub>3</sub>: Compared to presenting questions separately, fewer respondents will change their responses when questions are placed together in a list, and even fewer when presented in a grid.

### 3. Data and Methods

An experiment was designed to test the competing causal mechanisms for the dissimilar measurement properties from different layout designs, and embedded in a paging web survey with other experiments. The survey was on lifestyles and health and was administered in December 2004 and January 2005. Respondents were recruited from two sources: members of the Survey Sampling International (SSI) web panel who were sent invitations, and America On-Line (AOL) users who were invited through popup messages (“river-sampling”). AOL respondents typically have less computer experience and have done fewer

surveys than SSI panel members. The median time for taking the survey was 18 minutes.

The experiment consisted of four fully crossed factors: Layout, Instructions, Topic, and Location, resulting in a 3x3x2x2 design (Table 20).

Table 20. Experimental Design (except Early vs. Late Location in the Survey): Layout, Instructions, and Topic

	Single per page	All <i>listed</i> in 1 page	All in a <i>grid</i> in 1 page
Treat as related	By topic/Mixed	By topic/Mixed	By topic/Mixed
No instructions	By topic/Mixed	By topic/Mixed	By topic/Mixed
Treat independently	By topic/Mixed	By topic/Mixed	By topic/Mixed

Eight questions were asked – 4 on diet and 4 on exercise, presented in the Appendix. In the layout manipulation, the questions were placed either all on separate pages (Figure 8b), on the same page but each question listed separately with its own response options (Figure 8c), or on the same page and in a grid (Figure 8d). Note that the separate pages condition is the paging extreme of the paging vs. scrolling continuum in web survey design, while the 8 questions listed on the same page is a scrolling design as all respondents had to scroll – respondents’ browser window sizes were recorded and most did not see beyond the first three questions before scrolling.

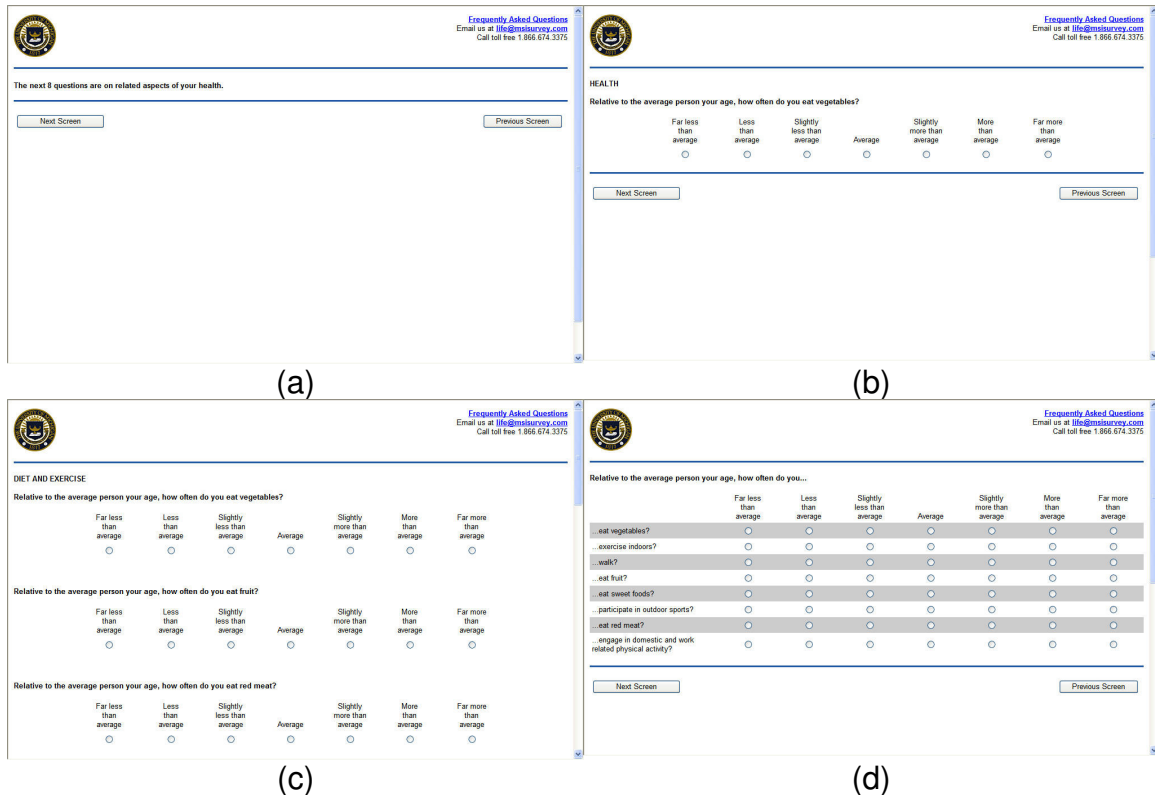


Figure 8. (a) Related Instructions, (b) Different Pages Layout, Related Instructions, (c) Same Page Separate Question Layout, Independent Instructions, Ordered Topics, and (d) Same Page Grid Layout, No Instructions, Mixed Topics.

Another experimental factor was instructions – respondents were instructed on the page prior to the questions (Figure 8a) that: “The next 8 questions are on related aspects of your health” (related), “Of the next 8 questions, 4 ask about your diet and 4 ask about how much you exercise” (independent), or no instructions. To reinforce this, under the related condition each page with questions had a subtitle “Health” (Figure 8b), under the independent condition the subtitle was “Diet and exercise” (Figure 8c), and no subtitle in the neutral condition.

The topic manipulation randomly assigned respondents to one of two conditions – the 4 diet questions were followed by the 4 exercise questions (by topic), or the 8 questions were intermixed (mixed).

The fourth manipulation varied the placement of the questions in the survey - early in the survey vs. late in the survey, a necessary manipulation to address the perceived burden hypothesis. In addition, location of the questions in the survey may interact with the survey design features that are in the previous experimental conditions – an undesired effect that from necessity is typically assumed away in surveys. That is, the effect of instructions may be smaller when the questions are toward the end of the survey as less attention is devoted to them, while the effect of the grid layout may be increased as suggested by a cumulating burden hypothesis, to the extent that different types of respondent burden interact.

For half of the respondents for whom the questions appeared closer to the end of the survey, a follow-up question was asked about how closely related the respondent found the previous 8 questions.

The diet and exercise topics for the questions were chosen as they are the input and output of energy in the body and are therefore major predictors of people's Body Mass Index (BMI) – an arithmetic combination of weight and height. Hence the better personal opinions on both dietary and exercise behavior are measured, the better respondents' BMI can be predicted. Therefore at the end of the survey respondents were asked about their height and weight in order to construct this criterion measure.

#### 4. Results

Among the 2,831 who started the survey, 2,587 completed the instrument. Responses to one of the questions on diet and another on exercise were inadvertently lost due to a programming error, leaving 3 questions on diet and 3 on exercise.

Among all the manipulations and two-way interactions, layout had a significant effect on the means of 3 of the 6 questions (Wilk's Lambda,  $F(12,4928)=2.78, p<.05$ ), while the location in the survey had an effect on 5 of the means (Wilk's Lambda,  $F(6,2464)=23.06, p<.05$ ). For the item for which the main effects were not significant, there was an interaction between layout and location. In the previous experiment (Chapter 4) means were not significantly different when the questions were in a grid or on different pages, and this experiment somewhat replicates this finding – when only these two layout conditions are compared and accounting for multiple tests, only one of the six questions has different means. The current finding further shows that some layout and design characteristics can also have effects on the point estimates of related self report measures.



#### 4.1. Hypotheses on Comprehension and Retrieval

Consistent with the finding in Chapter 4 and prior research, the variance-covariance matrices across the groups were significantly different (Box's M test,  $p < .05$ ). In addition, Table 21 shows that the Cronbach  $\alpha$  coefficient estimates were also consistent with the previous studies – higher when the questions were in a grid than in different pages (0.580 for diet and 0.739 for exercise vs. 0.564 and 0.687). In addition, they were higher when the questions were asked earlier in the survey, when respondents were instructed to treat the questions as related, and when the questions were ordered by topic. This merely indicates higher consistency of responses across questions and cannot be interpreted as which designs are preferable.

Table 21. Cronbach's Coefficient Alpha Estimates by Condition for Diet and Exercise Questions, and Percent of Respondents Leaving at Least One Question Unanswered.

Factor	Level	Diet Questions (3)	Exercise Questions (3)	Both Sets (6)	Percent with Item-missing	Sample Size
Location	Early	0.596	0.725	0.696	3.2%	1281
	Late	0.554	0.695	0.701	4.4%	1306
Layout*	Different Pages	0.564	0.687	0.696	2.2%	828
	Same Page, Listed	0.613	0.707	0.708	3.9%	906
	Same Page, Grid	0.580	0.739	0.707	5.3%	853
Instructions*	Related	0.598	0.709	0.709	4.3%	853
	Independent	0.577	0.732	0.714	2.5%	887
	None	0.586	0.690	0.686	4.6%	847
Topic	By Topic	0.591	0.728	0.707	4.1%	1299
	Intermixed	0.582	0.693	0.702	3.5%	1288
Overall		0.587	0.711	0.704	3.8%	2587

\* Significantly different in Percent with Item-missing,  $\alpha = .05$

In terms of internal consistency, the key difference between the perceived relatedness and the short-term memory hypothesis is that the former predicts differences between the grid design and the other two, while the short-term memory predicts that as long as the questions are on the same page, consistency should be about the same. Neither hypothesis seems supported by internal consistency estimates, as the three layout conditions are about as different from one another, while under both perceived relatedness and working memory hypotheses the expectation is for two conditions to be more alike.

There was some indication of conditions that resulted in higher internal consistency estimates to be also associated with higher percent of respondents leaving questions unanswered, shown in the last two columns in Table 1. Particularly notable are the two significant factors – layout and instructions. Consistent with Tourangeau, Couper, and Conrad (2004), although not with Couper, Traugott, and Lamias (2001), when questions were on different pages as opposed to a grid, there were fewer respondents skipping questions. Furthermore, when the questions were on the same page but not in a grid, the rate was somewhat in between. With respect to instructions, when respondents were informed that from the following questions half are on one topic and the rest on another topic, a lower percent of respondents left questions unanswered.

Internal consistency was examined in a more formal manner – computing variability across the 6 questions for each respondent. These items measured two constructs, diet and exercise, and even within the diet construct they were relatively heterogeneous – the question on sweet foods is very different from

consumption of fruits and vegetables, also exhibited in loadings in factor analysis. Asserting that lower variability across questions measuring different constructs is undesirable based on the first study (Chapter 4), standard deviations of responses from each respondent were regressed on the experimental manipulations, their interactions, and some covariates – computer and survey experience and demographic characteristics. The effects of the layout design, the location in the survey, and thematic ordering of the questions were significant predictors of the variability in responses (Figure 9). There was more variation across the 6 questions when they came earlier in the survey and when they were ordered by topic. They were also more variable when placed in the same page but as separate questions, and even more so when in different pages, compared to the grid format. There was also an indication of interaction between layout and location in the survey, as it was marginally significant ( $p=.070$ ) and necessitates replication – the variability reducing effect of the grid being more pronounced later in the survey.

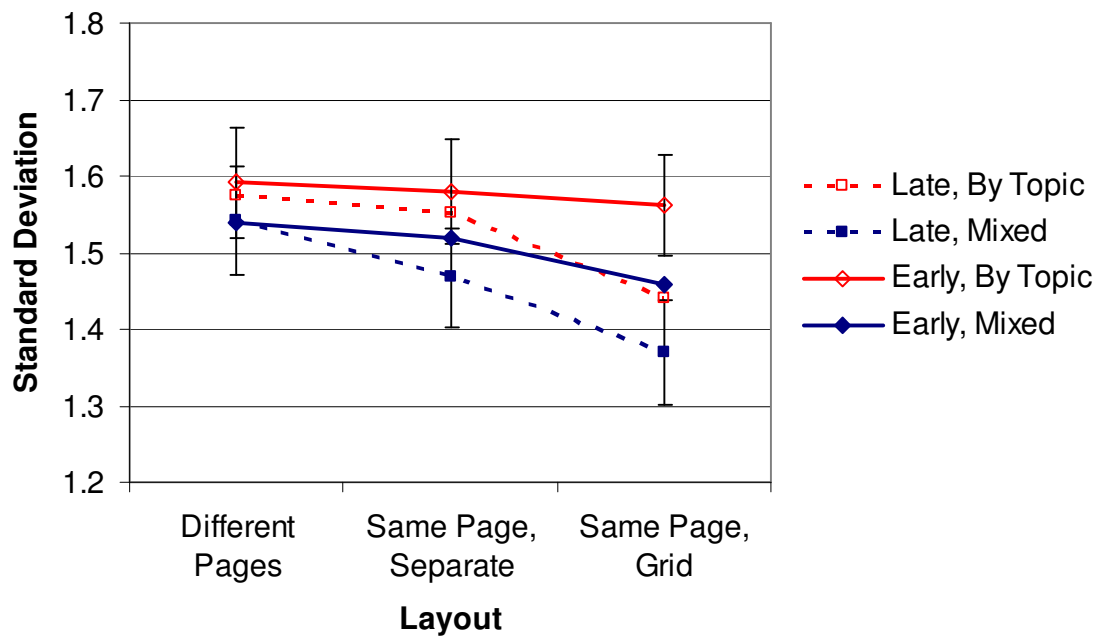


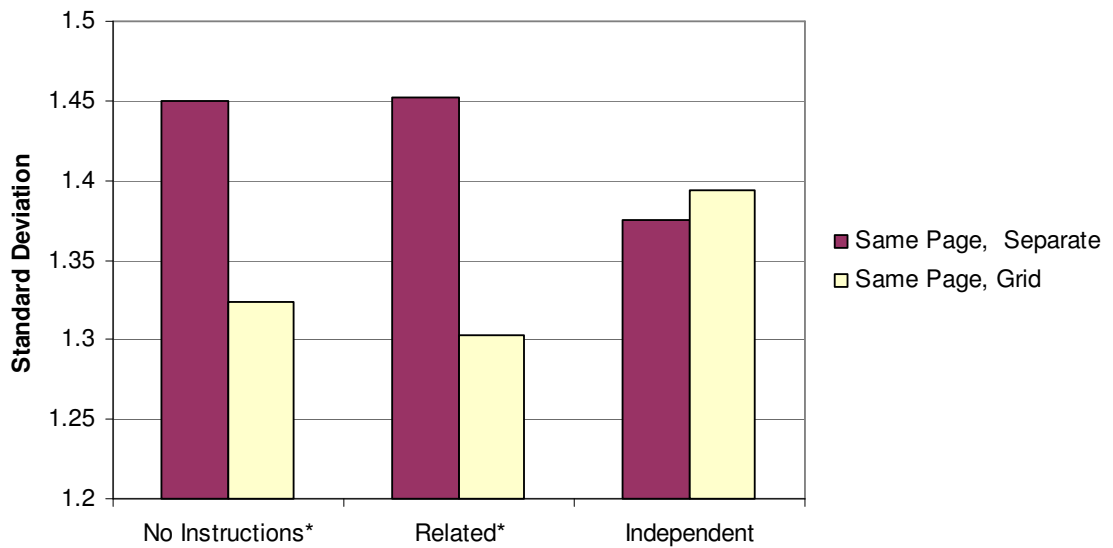
Figure 9. Variability of Responses Across the Six Questions by Layout Design, Location in Survey, and Order of Questions (Least Squares Adjusted Means).

Here as well, no two layout conditions seem similar in order to find support for either the perceived relatedness or the short-term memory hypothesis. However, under the perceived burden hypothesis, layout is expected to interact with other forms of burden – in this experiment it is experienced burden operationalized by early vs. late in the survey. Indeed, there was a significant interaction between layout design and location in survey, supporting the hypothesis for perceived burden.

Age, education, and gender were controlled for in the model, as they are related to diet and exercise and could assist in the model by removing a large proportion of the error variance. Among the covariates, age had a significant effect in the model above, with a linear relationship: the older the respondent, the higher the variability across questions. Although plausible explanations can be

found in the literature on aging in terms of differences in encoding of information, question comprehension, retrieval, etc., and in terms of the substantive answers, such as eating healthier as age progresses but exercising less, there was the opposite effect of age in the first study. None of the covariates, including age, interacted with the experimental manipulations.

Under the hypothesis that respondents perceive the questions as more related when they are on the same page, it would be expected that when no information about the relatedness of the questions is given, the grid design will elicit more consistent responses. Conversely, when presented with instructions that there are two separate constructs, there will be no difference between the grid and same page, but listed questions design. Similarly, there will be no difference when there are instructions that the same construct is being measured (i.e. that the questions are related). There was mixed support for the perception of relatedness hypothesis, as although under no instructions there was less variability in the grid condition, that also occurred when the related constructs instructions were used (Figure 10). As expected, there was no difference between the layout designs when the independent instructions were used.



\* Significant at  $\alpha=.05$

Figure 10. Variability of Responses Across Questions in the Same Page Layout Designs, by Instructions on Relatedness of Questions (Least Squares Adjusted Means for Gender, Age, and Education).

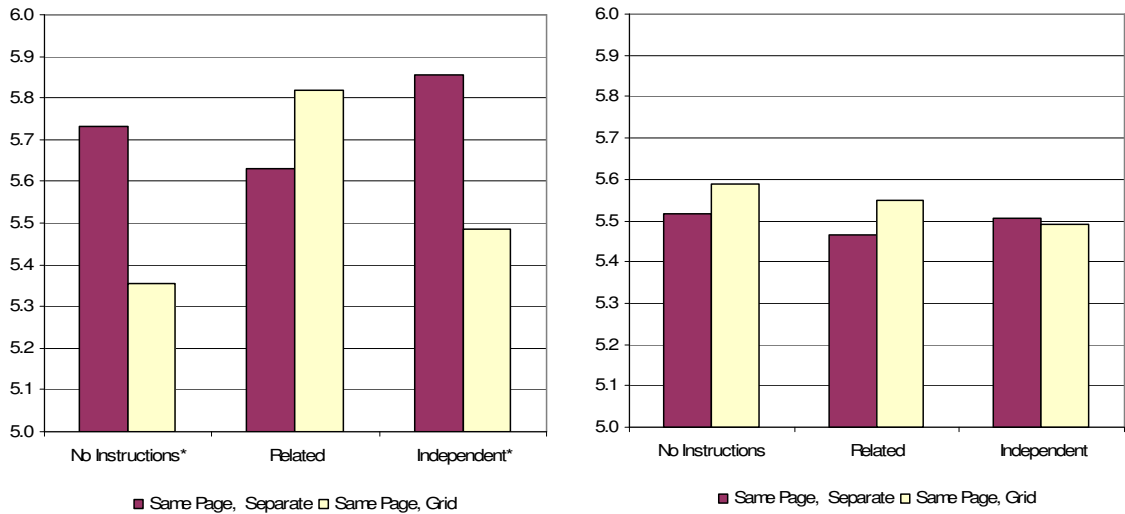
In order to evaluate these differences in variability using a criterion exogenous to the layout manipulation, BMI was regressed on the 6 questions and demographic covariates, for each combination of layout and instructions. The bottom two rows in Table 22 show that the biggest difference in proportion of variation in BMI explained is under the related instructions – when the questions were in a grid and informed that the questions are on the same topic (8.1%) vs. when listed separately in the same page (17.8%). This provides some limited support for the perceived relatedness hypothesis, as questions placed in a grid yield less variability and have lower predictive ability, and this is exacerbated when respondents are told explicitly that the questions are on the same topic. If so, a conclusion to a researcher is to keep questions separate and not to use a

common header on a single topic, as that can induce lack of variability and predictive ability of the questions.

Table 22. Percent of Variability in BMI Explained by the Questions, by Layout and Instructions.

	No Instructions	Related	Independent
Different Pages	29.9%	16.1%	21.2%
Same Page, Listed	22.1%	17.8%	22.6%
Same Page, Grid	25.4%	8.1%	20.8%

Under the perceived relatedness hypothesis, respondents are expected to find questions in a grid as more related. Ordering the questions by topic should increase that effect. The random half of respondents who received the questions late in the survey were asked an additional question on how closely related they found the preceding 8 questions. The reports from respondents were somewhat contradictory to the hypothesized and to results from the variability in responses – when the questions were in a grid, ordered by topic, and without any instructions, respondents found the questions less related (Figure 11a).



\* Significant at  $\alpha=.05$

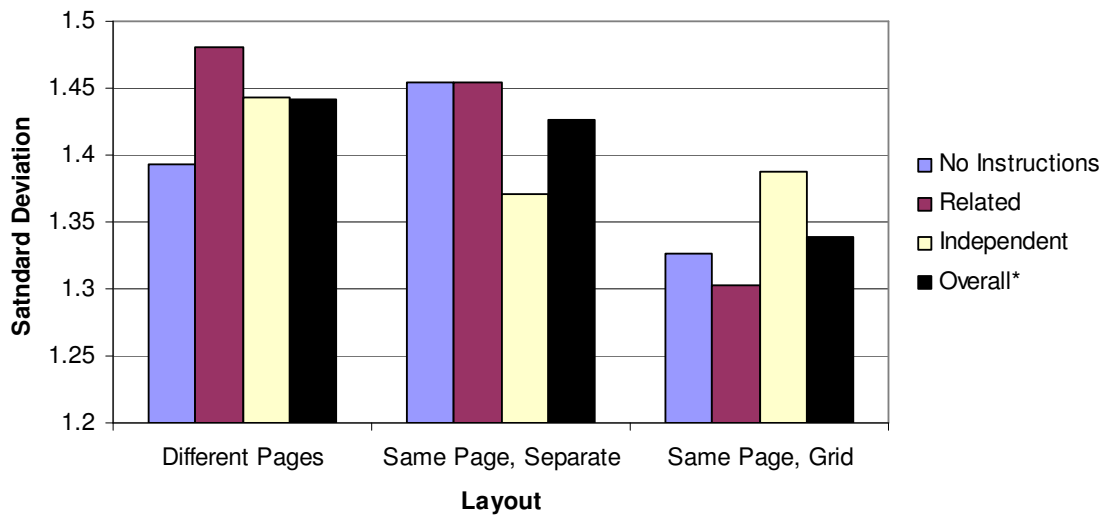
(a) By Topic

(b) Mixed

Figure 11. Responses to How Closely Related the Questions Were Perceived By Layout Design and Instructions on Relatedness of Questions When Questions are (a) Ordered By Topic and (b) Mixed. (Scale 1-7, 1=Completely unrelated, 7=Completely related)

Under the short-term memory mechanism, responses should be most different when the questions are in different pages from the other two layout designs, while the latter two should be similar. This was not supported, as the grid resulted in far more consistent responses than the other two conditions, represented by the black bars in Figure 12. Another expectation under short-term memory was that instructions (No instructions, Related, or Independent) would not have an effect within each layout design. No differences by instructions within layout conditions were found. However, this alone does not provide sufficient evidence as the instructions manipulation may have been too subtle for this purpose.





\* Significant at  $\alpha=.05$

Figure 12. Variability of Responses by Layout and Instructions on Relatedness of Questions (Least Squares Adjusted Means).

A much stronger test of the short-term memory hypothesis is to look at differences within a page by layout. In the grid layout all the questions were visible simultaneously by almost all respondents, while in the same page but separate, respondents could see 3-4 questions at the same time as scrolling was necessary. Under this hypothesis, there should be the same correlations between the first two variables in both layout versions, as they were visible together, but different associations between the questions when they were far apart – hence not visible together in the “scrolling” version. All plausible combinations were examined but none were significantly different across the two layout versions.

#### 4.2. Hypothesis on Perceived Burden

The third rival hypothesis was that of perceived burden and fatigue. Under this hypothesis, the multiple questions in a page, and especially the grid of radio buttons would have a negative impact on the measures. This was evident in Table 23 where the proportion of variance explained in BMI was highest when the questions were in different pages, followed by same page but separately listed questions. This was also expected to interact with location in the survey – the beneficial effect of placing the questions in different pages would be highest when the questions are earlier in the survey, as fatigue and loss of motivation can take place throughout the survey. There was some support for this, as  $R^2$  was highest for different pages early in the survey (24.2%), followed by same page but listed (17.4%), grid (13.5%), and the advantage for the different pages was lost when the items appeared closer to the end of the survey (14.2%), shown in Table 23.

Table 23. Percent of Variability in BMI Explained by the Questions, by Layout and Location in the Survey.

	Early	Late
Different Pages	24.2%	14.2%
Same Page, Listed	17.4%	18.6%
Same Page, Grid	13.5%	14.2%

The model included age and gender and interactions between them and the experimental manipulations.

Another test of perceived burden as the mechanism was to examine whether respondents exerted less effort into responding when the questions are in the same page, and even more so when they are in a grid. One such indicator

is the percent of respondents in each condition that changed responses to any of the 6 questions for which client-side paradata was available. This was confirmed as 27.4% changed responses in the different pages design, 23.7% in the listed in the same page, and 14.9% in the grid condition (Figure 13). To check for the unlikely possibility that changing responses is the result of not more, but less effortful responding, the time spent on the questions was also derived, which followed the same trend and was lowest for the grid condition (noting that the different pages design involves more interaction with the server).

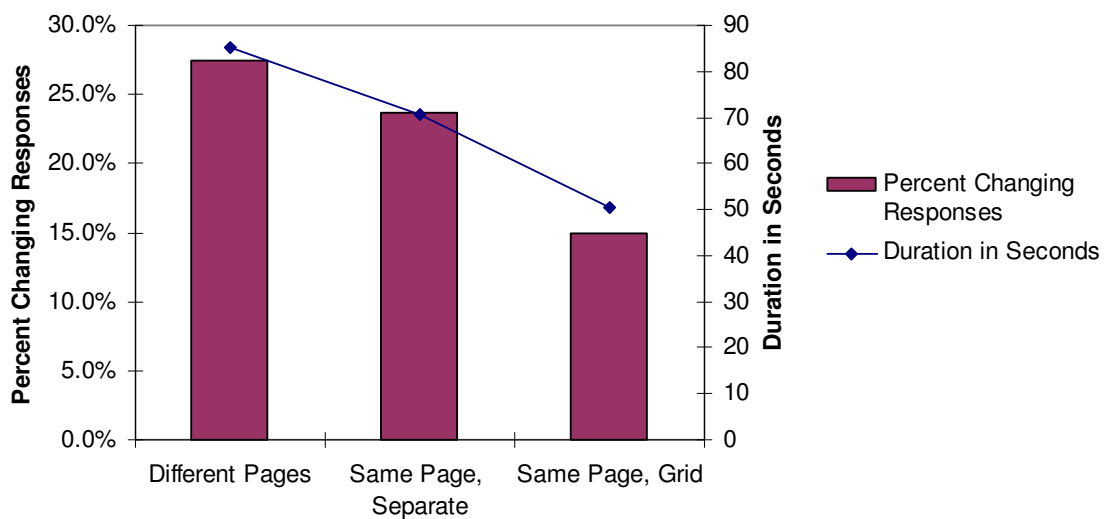


Figure 13. Percent Changing Responses and Time in Seconds by Layout Design.

This is further supported by lower data quality in the separate page listed design and even lower in the grid layout, as the item-missing data for the six questions was 2.2% in the different pages, 3.9% in the same page listed, and 5.3% in the grid (Table 1).

## 5. Discussion and Conclusions

There was mixed evidence in support of the hypothesis that respondents see questions as more related when they are in grid, as it had an impact only for the predictive ability of a criterion measure (BMI), and that impact was larger when combined with instructions that the questions are related. However, the other outcome measures did not support any of the perceived relatedness hypotheses.

None of the hypotheses for short-term memory impacts were confirmed.

The strongest support was found for the perceived burden hypothesis, as respondents provided higher rate of item-missing data, took less time to answer the questions, and fewer respondents changed responses, as questions were placed on the same page and in addition laid out in a grid.

Regardless of whether the point estimates for particular related questions are affected by layout, inter-item associations are consistently affected – inter-item correlations are higher when questions are on the same page. Chapter 4 showed evidence that the higher inter-item consistency is the result of correlated measurement error. The present study tested three likely causes, by manipulating design features. These features included the layout of the questions within a page, the number of questions in a page, whether the questions appear early or late in the questionnaire, topical ordering vs. intermixing of the questions, and instructions about the related/unrelated nature of the questions. The variability of, and associations among the questions was

affected significantly by the manipulated design features. Specifically, variability across questions measuring different constructs was reduced when: questions were placed in the same page, additionally placed in a grid, intermixed instead of organized by topic, and appeared later in the questionnaire. Lower variability across responses (higher internal consistency) was also associated with higher measurement error and lower validity in the previous experiment, necessitating formal tests of the underlying mechanisms through which these design features affect the quality of the measurement.

There was not sufficient support for the perceived-relatedness hypothesis in terms of the variability across responses, relationship with the criterion measure (Body Mass Index), and the debriefing question that asked directly how related the respondent found the questions to be in each condition. Instructions to the respondents on how related the questions are did have an effect on these criteria, but not in the expected pattern under this hypothesis.

Support for the working memory hypothesis was not found either, as instead of the different pages condition being the most different among the three layout designs, it was the grid layout that was different from the other two. In the layout designs that placed the questions on the same page, there was a unique difference that should have resulted in higher inter-item associations in the grid layout – all the questions were visible at the same time, while in the listed questions in the same page condition only about three of the questions could be seen simultaneously. However, no differences were found in the strength of the

associations for the questions that were not visible at the same time in the listed (scrolling) condition.

There was strong support for the third hypothesis – that of perceived burden. Asking respondents a single question at a time is likely to be perceived as a relatively simple task at each page, while at the other extreme, a grid with many questions can be demanding. Another expectation was that as respondents carried on through the survey, fatigue (i.e., experienced burden) would cause all layout designs to be viewed as more burdensome. The latter effect was evident in the proportion of variation explained in BMI – the highest estimate was in the different screens condition when the questions appeared early in the survey and the lowest was for the grid condition. The difference between the layout designs did diminish when presented late in the survey. The respondents also used less effort in responding (less time to the exact same questions apart from layout manipulations) when the questions were in the same screen and especially when they were in a grid, indicated by significantly fewer respondents changing their responses, more item-missing data, and taking less time to answer them.

The implications from these findings can help us design better surveys that reduce measurement error; as many researchers place multiple questions in a page and use grids to alleviate respondent burden, they are likely increasing the perceived burden. The implications of how we lay out questions are surprisingly far-reaching. What makes a survey design seem difficult or not deserving effort by the respondent has direct implications for measurement error.

In this particular case, placing questions on the same topic in the same page led to seemingly less effortful responding, and even more so when the questions were in a grid. Future work needs to identify what respondents perceive as difficult.

## 6. Appendix

Questions used in the study:

Relative to the average person your age, how often do you...

... vegetables?

... fruit?

... eat red meat?\*\*\*

... sweet foods?\*

... walk?

... exercise indoors?

... participate in outdoor sports?\*\*\*

... engage in domestic and work related physical activity?

Scale: Far less than average, Less than average, Slightly less than average, Average, Slightly more than average, More than average, Far more than average.

\* Recoded by reversing the scale for analyses.

\*\* Excluded from analyses due to a programming error overwriting responses.

The next question will help us understand your answers better.

How closely related did you find the last 8 questions (in the previous 8 screens)?

How closely related did you find the 8 questions on the previous screen?

Very closely related

Related

Somewhat related

Neither related nor unrelated

Somewhat unrelated

Unrelated

Completely unrelated

What is your height?

Feet \_\_\_\_\_

Inches \_\_\_\_\_

What is your weight?

\_\_\_\_\_ pounds.



## Chapter 6. Conclusion

This dissertation addressed questions about participation decisions and measurement error as parts of a unified framework for response behavior in web surveys. The framework is intended to assist with causal thinking about response behavior by linking factors with decisions and subsequently with response outcomes, analogously to the framework for cooperation in household surveys proposed by Groves and Couper (1998). However, the proposed framework for web survey cooperation and responding includes multiple response outcomes, including breakoffs in addition to unit nonresponse. Furthermore, measurement error and data quality are also included as the factors that affect breakoff decisions are likely to also affect the response process (Cannell et al., 1977; Cannell et al., 1981; Tourangeau, 1984; Strack and Martin, 1987; 1987). The scope of this framework is too large for any single work to address. I focused on four related topics within the framework that I find to be gaps in the literature. The first study set out to demonstrate that breakoffs and nonrespondents are different in expected ways, based on common and unique influences on each participation decision. The second study examined causes of breakoff in web surveys, in part motivated by the survey response process model for answering questions. In order to study the effect of causes of breakoff and

nonresponse on measurement error, we first need to be able to evaluate different survey designs in terms of measurement error. The third study compared two survey layout designs using a method that did not require true values to be known for respondents, and in addition, relaxed the commonly made assumption of absence of correlated measurement error. The last study followed up on these findings to identify the causes producing different measurement error in each survey design.

There were four main findings, listed in the order the presentation of the studies:

1. Breakoffs are different from unit nonrespondents.

Breakoff is a relatively understudied response behavior. If breakoffs are the same respondents who tend to also be nonrespondents, and similarly, breakoff is predicted by the same correlates that predict nonresponse, it would be acceptable to use the same methods to reduce and to adjust for both breakoffs and nonrespondents. To the extent that breakoffs are different from nonrespondents, this response outcome should be studied separately and the causal mechanisms producing breakoff should receive separate attention, and particularly because breakoffs occur at relatively high rates in web surveys (Musch and Reips, 2000; Lozar-Manfreda and Vehovar, 2002). To address this question,

respondent behavior was linked for students invited to five different surveys administered between 2001 and 2005.

Unlike nonresponse, breakoff behavior is not consistent across surveys, even after controlling for differences in response rates across surveys and any potential differences in sample composition across pairs of surveys.

The respondent and survey characteristics associated with nonresponse are less predictive of breakoff than of nonresponse.

2. Breakoffs are a function of different cognitive demands from question and page characteristics within the survey questionnaire.

Survey and respondent covariates of nonresponse are not as predictive of breakoff and even more notably, breakoffs do not have a higher propensity to be breakoffs in another survey. Identification of the causal mechanism producing breakoffs is needed in order to anticipate the consequences, to be able to reduce breakoffs, and to create effective adjustments. Unlike unit nonrespondents, breakoffs have been exposed to some part of the survey content. Coding question and page characteristics within the questionnaire into cognitive demands that fit into the survey response process model provides the ability to test a generalizable mechanism producing breakoff, which would also be common to that producing measurement error. Consistent with the

hypothesized cognitive mechanism, sample members with lower cognitive ability (i.e., older) and lower cognitive sophistication (less educated) should be more likely to break off. Two surveys with the same study design and sample source were combined. The surveys had a relatively large degree of variability in question and page characteristics, and some degree of random assignment of order of sets of questions.

Pages with questions that require more comprehension, judgment, or response formatting elicit significantly higher rates of breakoff, controlling for respondent characteristics and survey design. While respondents with lower cognitive sophistication (indicated by level of education) are more likely to breakoff, such relationship is not found for lower cognitive ability (indicated by age).

In addition, there is some limited evidence for particular breakoffs to be induced by a similar decision process as unit nonrespondents. The surveys included a few section introduction pages to help transition between topics. Such pages have no questions or demands from respondents, other than the topic change and implied request to start another part of the survey, both of which carry analogies to initial survey requests. These pages had significantly higher breakoff rates. Higher breakoff rates on section introductions are also consistent with findings from telephone surveys (Groves and Kahn, 1979).

3. Placing questions on the same page yields higher correlated measurement error.

One survey design decision that has to be made in web surveys is whether to group similar questions on the same page or keep them on separate pages. Previous findings show that placing questions on the same page increases the inter-item correlations (Couper et al., 2001). Despite the differences, there is no guidance on preferable design. The evaluation of the effect of different designs on measurement error is hindered by the need for estimation methods for measurement error in the absence of true values. Such estimation is particularly problematic due to the likely presence of not just measurement error, but also correlated measurement error across questions, as much of the survey measurement literature discusses how previous questions affect responses to subsequent ones. Commonly used methods for estimating the reliability of the measurement of constructs are based on inter-item correlations, such as Cronbach's Coefficient  $\alpha$ , and assume uncorrelated measurement errors.

Using a latent variable approach with the concurrent use of criterion measures allow us to relax the assumption of uncorrelated measurement error and to compare layout designs.

When questions are placed on the same page inter-item correlations are higher and yield higher reliability estimates, but that is a result of correlated measurement error.

4. Placing questions on the same page makes the respondent's task more burdensome, leading to inferior responding.

Identification of survey designs that yield less measurement error is necessary, yet it does not reveal the causal mechanism – why placing questions on the same page yields more (correlated) measurement error. A follow-up experiment was conducted to manipulate different design features that comprise the various layout formats.

The experiment found that the difference between single versus multiple questions per page is likely because the respondent's task seems more burdensome when multiple questions are displayed simultaneously, rather than the questions being interpreted as more related, or respondents being able to recall previous responses from short-term memory more easily.

A researcher needs to be cautious about what is assumed to be common knowledge. While questions on the same topic that share a response scale are often placed in grids on the same page, possibly because it makes it easier and faster for the respondent, and yields less

item-missing data, it can actually be perceived as more burdensome and yields higher measurement error.

There are some inherent limitations to these studies, as well as directions for future work based on these findings.

There seem to be no reason to expect that the general population will behave in a completely different manner from the student population to negate the findings on consistency and relationship between response behaviors: 1) nonrespondents tend to be nonrespondents in subsequent surveys, yet breakoffs are not more likely to be breakoffs; and 2) nonrespondents are not more likely to be breakoffs, nor are breakoffs more likely to be nonrespondents. A strength of these surveys is that they are probability samples of a non-self selected population, as opposed to volunteer web survey panels that may exhibit a different relationship between respondent characteristics and unit nonresponse. However, the respondent and survey characteristics associated with the nonresponse and breakoff propensities in the five Student Life and Living Learning Program studies are limited, in addition to sample composition limitations. A significant limitation of this study is that the substantive data had been destroyed in accordance with the Institutional Review Board, while studying bias in substantive survey statistics, such as means, variances, and associations for key survey variables is much needed to understand the survey behaviors of interest.

The causes of breakoff have received less attention than unit nonresponse, yet breakoffs can be a relatively large part of those who start the survey. Considering that in many web surveys unit nonresponse is not identifiable or does not have a meaningful definition, such as in volunteer web survey panels, it is surprising that breakoff has not received more attention. We suggest that the causal models for breakoffs should test causes that have already been identified for nonresponse and for measurement error over the past decades. This dissertation addresses only part of what can be tested as mechanisms producing breakoff, and there is much more that can be imported from the causes identified for nonresponse and measurement error that under our framework are expected to overlap. Of lesser importance but nonetheless important to note are the practical limitations of this study, as even greater variability in question and page characteristics with random assignment of placement in the questionnaire would be desirable.

An ideal study that addresses breakoff in web surveys would incorporate all design features discussed above into a single study. Multiple surveys would be administered to the same target population with at least some overlap in sample across surveys in order to better understand the consistency and the survey and respondent covariates that explain the consistency of each response behavior. While consistency is not found for breakoff in the study reported here, this result needs to be replicated on a more general target population that will have greater variability in respondent characteristics, and a survey design that will have greater variability of features. The within subject design would also



permit the evaluation of bias in key survey statistics from each nonresponse behavior, at least for respondents who completed at least one survey. While better understanding of the survey design features would inform future designs, attributes such as the announced length and incentive structure of the survey can also be used to inform the design of nonresponse follow-up surveys.

Understanding associations with respondent covariates may in turn be more valuable for post-survey adjustment, although surveys of special populations that exhibit different behavior may take that into account at the design stage as well. As noted earlier, the processes producing breakoffs have received little attention and it is the combination of the three sets of factors in the proposed framework that need to be studied simultaneously: survey design features that vary across surveys, respondent characteristics and predispositions, and within survey page and question characteristics.

The latter needs to be addressed with much attention in an experimental design. Page and question characteristics need to be randomized not for just sets of pages, but in such a manner that there is no confounding in the order of characteristics while maintaining a flow in the questionnaire. For example, comprehension intensive questions should not always follow a section introduction and there should still be sufficient context for the questions. While the flow is maintained in the second study presented here, randomization of placement of pages could be more extensive if the study is designed for analysis of response behavior. An additional improvement for future studies is to consider the possible design features that could cause different response behaviors. This

alludes to the design of the fourth study on causes of measurement error, as the number of questions per page may affect the breakoff rate, yet it could be one specific component of that design. For example, it could be that questions placed in a grid on the same page produce higher or lower rates of breakoff, while questions placed on the same page but not in a grid may not alter the breakoff rate per question. Ideally, such manipulations would be informed by hypotheses that are either based on theory, empirically justified by findings from nonresponse or measurement error, or both.

Although the causal mechanisms can be shared to a large degree with breakoff, studying measurement error poses a different set of problems, especially in the absence of validation data.

The method used to evaluate different survey designs addresses a previously unaddressed estimation problem for survey designers. Whether simple linear regression or structural equation modeling is used, the fundamental premise needs validation – the survey design that leads to the best explanation of the variance in a criterion measure is the one that has the least measurement error. This is a testable hypothesis that requires a study with validation measures for the experimentally manipulated questions. Similarly, the identification of the causal mechanisms producing measurement differences between two specific layout designs, single versus multiple questions per page, can benefit from improving the criterion measures from variability, item-missing data, changing responses, and time, to person-level estimates of measurement error.

There are many other design decisions that are not empirically informed by measurement error criteria. This approach may be useful to future studies evaluating designs for superiority over alternatives instead of stating mean or distributional differences, or data quality at best. Each evaluation method makes certain assumptions. Other approaches to evaluating measurement error across designs are needed, as they could be based on different sets of assumptions and would help ascertain the veracity of findings.

In summary, this set of studies addresses only part of the factors and causal processes involved in response behavior in web surveys. As greater understanding is gained about the processes that lead to each response outcome, the nexus between them should also receive more attention and I hope the proposed framework will assist with this thinking. More generally, to improve the properties of a statistic of interest, consideration must be given to multiple competing outcomes that include errors of observation and nonobservation, and have overlapping causal mechanisms.

## References

- AAPOR (2006). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Lenexa, Kansas, AAPOR. **4th edition**.
- Alwin, D. F. (1973). "Making Inferences from Attitude-Behavior Correlations." Sociometry **36**(2): 253-278.
- Alwin, D. F. (1976). "Attitude Scales as Congeneric Tests - Re-Examination of an Attitude-Behavior Model." Sociometry **39**(4): 377-383.
- Andrews, F. M. and A. R. Herzog (1986). "Respondent Age and Survey Measurement Error." Journal of the American Statistical Association **81**: 403-410.
- Apodaca, R., S. Lea and B. Edwards (1998). The Effect of Longitudinal Burden on Survey Participation. American Statistical Association Proceedings of the Survey Research Methods Section.
- Baddeley, A. D. and G. Hitch (1993). "The Recency Effect - Implicit Learning with Explicit Retrieval." Memory & Cognition **21**(2): 146-155.
- Bayer, L. R. and R. K. Thomas (2004). A Comparison of Sliding Scales with Other Scale Types in Online Surveys. Paper presented at the RC33 International Conference on Social Science Methodology. Amsterdam.

- Beatty, P. and D. Herrmann (2002). To Answer or Not to Answer: Decision Process Related to Survey Item Nonresponse. Survey Nonresponse. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York, Wiley: 71-85.
- Bell, D. S., C. M. Mangione and C. E. Kahn (2001). "Randomized Testing of Alternative Survey Formats Using Anonymous Volunteers on the World Wide Web." Journal Of The American Medical Informatics Association **8**(6): 616-620.
- Berk, M. L., N. A. Mathiowetz, E. P. Ward and A. A. White (1987). "The Effect of Prepaid and Promised Incentives: Results of a Controlled Experiment." Journal of Official Statistics **3**(4): 449-457.
- Blumberg, S. J., L. Olson, M. R. Frankel, L. Osborn, K. P. Srinath and P. Giambo (2005). "Design and Operation of the National Survey of Children's Health, 2003." National Center for Health Statistics, Vital Health Statistics **43**(1).
- Bogen, K. (1996). The Effect of Questionnaire Length on Response Rates -- a Review of the Literature. American Statistical Association Proceedings of the Survey Research Methods Section.
- Bosnjak, M. and T. L. Tuten (2001). "Classifying Response Behaviors in Web-Based Surveys." Journal of Computer Mediated Communication **6**(3).
- Bosnjak, M. and T. L. Tuten (2003). "Prepaid and Promised Incentives in Web Surveys - an Experiment." Social Science Computer Review **21**(2): 208-217.

- Bradburn, N. M. (1978). Respondent Burden. American Statistical Association Proceedings of the Survey Research Methods Section.
- Bradburn, N. M. and W. M. Mason (1964). "The Effect of Question Order on Responses." Journal of Marketing Research 1(4): 57-61.
- Brennan, M. and J. Hoek (1992). "The Behaviour of Responders, Non-Responders and Refusers across Surveys." Public Opinion Quarterly 52(4): 530-535.
- Brennan, R. L. (2001). "Some Problems, Pitfalls, and Paradoxes in Educational Measurement." Educational Measurement: Issues and Practice 20(4): 6-18.
- Busemeyer, J. R. and J. T. Townsend (1993). "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment." 100(3): 432.
- Campbell, A., P. E. Converse and W. L. Rogers (1976). The Quality of American Life. New York, Sage.
- Cannell, C., P. Miller and L. Oksenberg (1981). "Research on Interviewing Techniques." Sociological Methodology: 389-437.
- Cannell, C. F., K. H. Marquis and A. Laurent (1977). A Summary of Studies of Interviewing Methodology, Vital Health Stat. No. 69.: 77-1343.
- CASRO (1982). A Special Report of the Casro Task Force on Completion Rates. Port Jefferson, New York, CASRO.
- Catania, J. A., D. Binson, J. Canchola, L. M. Pollack and W. Hauck (1996). "Effects of Interviewer Gender, Interviewer Choice, and Item Wording on

- Responses to Questions Concerning Sexual Behavior." Public Opinion Quarterly **60**(3): 345-375.
- Champion, D. J. and A. M. Sear (1969). "Questionnaire Response Rate: A Methodological Analysis." Social Forces **47**: 335-339.
- Cialdini, R. B. (1993). Influence: The Psychology of Persuasion. New York, Morrow.
- Conrad, F. G., M. P. Couper, R. Tourangeau and A. Peytchev (2005). Impact of Progress Feedback on Task Completion: First Impressions Matter. CHI '05 extended abstracts on Human factors in computing systems, Portland, OR, ACM Press.
- Conrad, F. G., M. P. Couper, R. Tourangeau and A. Peytchev (In press). "Effectiveness of Progress Indicators in Web Surveys." Journal of Official Statistics.
- Couper, M. P., R. Tourangeau, F. G. Conrad and E. Singer (2006). "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment." Social Science Computer Review **24**(2): 227-245.
- Couper, M. P., M. W. Traugott and M. J. Lamias (2001). "Web Survey Design and Administration." Public Opinion Quarterly **65**(2): 230-253.
- Crawford, S. D., M. P. Couper and M. J. Lamias (2001). "Web Surveys - Perceptions of Burden." Social Science Computer Review **19**(2): 146-162.
- Creel, D., D. Jang, D. Kasprzyk and T. Williams (2002). Permanent Random Number Technique to Minimize Response Burden in Repeated Surveys.

American Statistical Association Proceedings of the Survey Research Methods Section.

Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests."

Psychometrika **16**(3): 297-334.

Cronbach, L. J. (2004). "My Current Thoughts on Coefficient Alpha and

Successor Procedures." Educational and Psychological Measurement **64**:

391-418.

Cronbach, L. J., N. Rajaratnam and G. C. Gleser (1963). "Theory of

Generalizability - a Liberalization of Reliability Theory." British Journal Of

Statistical Psychology **16**(2): 137-163.

Deutskens, E., K. de Ruyter, M. Wetzels and P. Oosterveld (2004). "Response

Rate and Response Quality of Internet-Based Surveys: An Experimental

Study." Marketing Letters **15**(1): 21.

Dillman, D. A. (1978). Mail and Telephone Surveys. New York, Wiley.

Dillman, D. A. and D. K. Bowker (2001). The Web Questionnaire Challenge to

Survey Methodologists. Dimensions of Internet Science. U.-D. Reips and

M. Bosnjak. Lengerich, Pabst Science Publishers: 159-178.

Dillman, D. A., M. D. Sinclair and J. R. Clark (1993). "Effects of Questionnaire

Length, Respondent-Friendly Design, and a Difficult Question on

Response Rates for Occupant-Addressed Census Mail Surveys." Public

Opinion Quarterly **57**(3): 289-304.

Dillman, D. A., R. D. Tortora and D. Bowker (1998). Principles for Constructing

Web Surveys. SESRC Technical Report 98-50, Pullman, Washington.



- Ford, B. L., R. R. Hocking and A. Coleman (1978). Reducing Respondent Burden on an Agricultural Survey. American Statistical Association Proceedings of the Survey Research Methods Section.
- Fox, S. (2005). Digital Divisions: There Are Clear Differences among Those with Broadband Connections, Dial-up Connections, and No Connections at All to the Internet, <http://www.pewinternet.org/>. 2006.
- Frick, A., M. T. Bächtiger and U.-D. Reips (2001). Financial Incentives, Personal Information, and Drop out in Online Studies. Dimensions of Internet Science. U.-D. Reips and M. Bosnjak. Lengerich, Germany, Pabst Science: 209-219.
- Fuchs, M. (2001). Screen Design in a Web Survey. Paper presented at the annual meeting of the American Association for Public Opinion Research, Montréal, Quebec.
- Fuller, W. A. (1987). Measurement Error Models. New York, Wiley.
- Galesic, M. (2006). "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." Journal of Official Statistics **22**(2): 313-328.
- Galvin, L. N., S. Sae-Ung and K. King (2000). Effect of Interview Length and Proxy Interviews on Attrition to the Survey of Income and Program Participation. American Statistical Association Proceedings of the Survey Research Methods Section.

- Ghosh, D. and A. Vogt (2000). Determining the Optimal Split for a Lengthy Questionnaire. American Statistical Association Proceedings of the Survey Research Methods Section.
- Goyder, J. (1987). The Silent Minority: Nonrespondents on Sample Surveys, Polity Press/Westview Press.
- Goyder, J. C. (1982). "Further Evidence on Factors Affecting Response Rates to Mailed Questionnaires." American Sociological Review **47**(4): 550-553.
- Gräf, L. (2002). Assessing Internet Questionnaires: The Online Pretest Lab. Online Social Sciences. B. Batinic, U.-D. Reips and M. Bosnjak. Seattle, Hogrefe & Huber Publishers: 69-80.
- Gray, R., P. Campanelli, K. Deepchand and P. Prescott-Clarke (1996). "Exploring Survey Non-Response: The Effect of Attrition on a Follow-up of the 1984-85 Health and Life Style Survey." The Statistician **45**(2): 163-183.
- Groves, R. and R. L. Kahn (1979). Surveys by Telephone. New York, Academic Press.
- Groves, R. M., R. B. Cialdini and M. P. Couper (1992). "Understanding the Decision to Participate in a Survey." Public Opinion Quarterly **56**(4): 475-495.
- Groves, R. M. and M. P. Couper (1998). Nonresponse in Household Interview Surveys. New York, Wiley.
- Groves, R. M. and M. P. Couper (1998). Nonresponse in Household Interview Surveys. New York, John Wiley and Sons.

- Groves, R. M., S. Presser and S. Dipko (2004). "The Role of Topic Interest in Survey Participation Decisions." Public Opinion Quarterly **68**(1): 2-31.
- Groves, R. M., E. Singer and A. Corning (2000). "Leverage-Saliency Theory of Survey Participation - Description and an Illustration." Public Opinion Quarterly **64**(3): 299-308.
- Guttman, L. (1953). "Reliability Formulas That Do Not Assume Experimental Independence." Psychometrika **18**(3): 225-239.
- Heberlein, T. A. and R. Baumgartner (1978). "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature." American Sociological Review **43**(4): 447-462.
- Heise, D. R. (1969). "Separating Reliability and Stability in Test-Retest Correlation." American Sociological Review **34**(1): 93-101.
- Heise, D. R., and G.W. Bohrnstedt (1970). Validity, Invalidity, and Reliability. Sociological Methodology. E. F. Borgatta and G. W. Bohrnstedt. San Francisco, Jossey-Bass: 104-129.
- Herzog, A. R. and J. G. Bachman (1981). "Effects of Questionnaire Length on Response Quality." Public Opinion Quarterly **45**(4): 549-559.
- Herzog, A. R. and L. Dielman (1985). "Age Differences in Response Accuracy for Factual Survey Questions." Journal of Gerontology **40**: 350-357.
- Hogg, A. and J. Miller (2003). Watch out for Droupouts.
- Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure." Scandinavian Journal of Statistics **6**: 65-70.

- Hu, L. and P. M. Bentler (1999). "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." Structural Equation Modeling **6**: 1-55.
- Interactive Marketing Research Organization, M. R. A. (2006). Imro Guidelines for Best Practices in Online Sample and Panel Management. **2006**.
- Jenkins, C. R. and D. A. Dillman (1995). Towards a Theory of Self-Administered Questionnaire Design. Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin. New York, Wiley: 165-196.
- Joreskog, K. G. (1970). "A General Method for Analysis of Covariance Structures." Biometrika **57**(2): 239-&.
- Joreskog, K. G. (1971). "Statistical Analysis of Sets of Congeneric Tests." Psychometrika **36**(2): 109-133.
- Jöreskog, K. G. and D. Sörbom (1996). Lisrel 8 User's Reference Guide. Chicago, Scientific Software International.
- Kalton, G. and D. Kasprzyk (1986). "The Treatment of Missing Survey Data." Survey Methodology **12**: 1-16.
- Kaplowitz, M. D., T. D. Hadlock and R. Levine (2004). "A Comparison of Web and Mail Survey Response Rates." Public Opinion Quarterly **68**(1): 94-101.
- Katz, L. (1964). "Effects of Differential Monetary Gain and Loss on Sequential Two-Choice Behavior." Journal of Experimental Psychology **68**: 245-249.

- Knauper, B. (1999). "The Impact of Age and Education on Response Order Effects in Attitude Measurement." Public Opinion Quarterly **63**(3): 347-370.
- Knauper, B., R. F. Belli, D. H. Hill and A. R. Herzog (1997). "Question Difficulty and Respondents' Cognitive Ability: The Effect of Data Quality." Journal of Official Statistics **13**: 181-99.
- Komaroff, E. (1997). "Effect of Simultaneous Violations of Essential Tau-Equivalence and Uncorrelated Error on Coefficient Alpha." Applied Psychological Measurement **21**(4): 337-348.
- Krosnick, J. A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." Applied Cognitive Psychology **5**: 213-236.
- Krosnick, J. A. and D. F. Alwin (1987). "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." Public Opinion Quarterly **51**(2): 201-219.
- Lane, H. J., A. M. Lane and H. Matheson (2004). "Validity of the Eating Attitude Test among Exercisers." Journal Of Sports Science And Medicine **3**(4): 244-253.
- Lewis, C. J., L. S. Sims and B. Shannon (1989). "Examination of Specific Nutrition-Health Behaviors Using a Social Cognitive Model." Journal of the American Dietetic Association **89**(2): 194-199.
- Little, R. J. A. (1982). "Models for Nonresponse in Sample Surveys." Journal of the American Statistical Association **77**(378): 237-250.

- Lozar-Manfreda, K. and V. Vehovar (2002). Survey Design Features Influencing Response Rates in Web Surveys. ICIS.
- MacElroy, B. (2000). Variables Influencing Dropout Rates in Web-Based Surveys.
- McGuire, W. J. (1960). A Syllogistic Analysis of Cognitive Relationships. Attitude Organization and Change; an Analysis of Consistency among Attitude Components. M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. Abelson and J. Brehm. New Haven, CT, Yale University Press: 65-111.
- Milfont, T. L. and J. Duckitt (2004). "The Structure of Environmental Attitudes: A First- and Second-Order Confirmatory Factor Analysis." Journal Of Environmental Psychology **24**(3): 289-303.
- Mooney, G. M., B. Rogers and D. Trunzo (2003). Examining the Effect of Error Prompting on Item Nonresponse and Survey Nonresponse in Web Surveys. American Association for Public Opinion Research, 58th Annual Conference, Nashville, Tennessee.
- Musch, J. and U.-D. Reips (2000). A Brief History of Web Experimenting. Psychological Experiments on the Internet. M. H. Birnbaum. San Diego, CA, Academic Press: 61-85.
- Myers, B. and L. Lecci (1998). "Revising the Factor Structure of the Juror Bias Scale: A Method for the Empirical Validation of Theoretical Constructs." Law And Human Behavior **22**(2): 239-256.

- Neumark-Sztainer, D., M. Wall, C. Perry and M. Story (2003). "Correlates of Fruit and Vegetable Intake among Adolescents - Findings from Project Eat." Preventive Medicine **37**(3): 198-208.
- Novick, M. R. (1966). "The Axioms and Principal Results of Classical Test Theory." Journal of Mathematical Psychology **3**: 1-18.
- Nunnally, J. C. and I. H. Bernstein (1994). Psychometric Theory. New York, McGraw-Hill.
- O'Neil, K. M., S. D. Penrod and B. H. Bornstein (2003). "Web-Based Research: Methodological Variables' Effects on Dropout and Sample Characteristics., 35. 217-236." Behavior Research Methods, Instruments, and Computers **35**(2): 217-236.
- Olson, K., M. C. Murray, J. Cardoni, M. Battaglia and J. Lepkowski (2004). Breaking Off in the Middle: Partial Interviews in a Large Telephone Survey. Paper presented at the Midwest Association for Public Opinion Research Annual Conference, Chicago, IL.
- Osborn, L., S. J. Blumberg and L. Olson (2000). Assessing a New Data Collection Methodology for Obtaining Sensitive Data Using an Rdd Telephone Survey. Paper Presented at the 55th Annual Conference of the American Association for Public Opinion Research & World Association for Public Opinion Research, Portland, Oregon.
- Peytchev, A., M. P. Couper, S. McCabe and S. Crawford (2004). Web Survey Design: Paging Vs. Scrolling. Paper presented at the American

Association for Public Opinion Research 59th Annual Conference, May, Phoenix.

Peytchev, A., M. P. Couper, S. E. McCabe and S. Crawford (2006). "Web Survey Design: Paging Vs. Scrolling." Public Opinion Quarterly **70**(4).

Raghunathan, T. E. and J. E. Grizzle (1995). "A Split Questionnaire Survey Design." Journal Of The American Statistical Association **90**(429): 54-63.

Ramos, M., B. M. Sedivi and E. M. Sweet (1998). Computerized Self-Administered Questionnaires. Computer Assisted Survey Information Collection. M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II and J. M. O'Reilly. New York, John Wiley & Sons, Inc.: 389-408.

Raykov, T. (1997). "Estimation of Composite Reliability for Congeneric Measures." Applied Psychological Measurement **21**(2): 173-184.

Raykov, T. (2001). "Bias of Coefficient Alpha for Fixed Congeneric Measures with Correlated Errors." Applied Psychological Measurement **25**(1): 69-76.

Reips, U.-D. (2002). Context Effects in Web Surveys. Online Social Sciences. B. Batinic, U.-D. Reips and M. Bosnjak. Seattle, Hogrefe & Huber Publishers: 69-80.

Sanchez, M. E. (1992). "Effect of Questionnaire Design on the Quality of Survey Data." Public Opinion Quarterly **56**: 206-217.

Satia, J. A., J. A. Galanko and A. M. Siega-Riz (2004). "Eating at Fast-Food Restaurants Is Associated with Dietary Intake, Demographic,



- Psychosocial and Behavioural Factors among African Americans in North Carolina." Public Health Nutrition 7(8): 1089-1096.
- Schaie, K. W. (1994). "The Course of Adult Intellectual Development." American Psychologist 49: 304-313.
- Schuman, H. and M. P. Johnson (1976). "Attitudes and Behavior." Annual Review Of Sociology 2: 161-207.
- Schwarz, N. and H.-J. Hippler (1995). "Subsequent Questions May Influence Answers to Preceding Questions in Mail Surveys." Public Opinion Quarterly 59: 93-97.
- Schwarz, N., F. Strack and H. Mai (1991). "Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis." Public Opinion Quarterly 55: 3-23.
- Schwarz, S. and U.-D. Reips (2001). Cgi Versus Javascript: A Web Experiment on the Reversed Hindsight Bias. Dimensions of Internet Science. U.-D. Reips and M. Bosnjak. Lengerich, Germany, Pabst Science: 75-90.
- Sigelman, L. (1981). "Question-Order Effects on Presidential Popularity." Public Opinion Quarterly 45(2): 199-207.
- Simon, H. A. (1957). Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York, Wiley.
- Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York, Wiley: 163-177.

- Singer, E., R. M. Groves and A. D. Corning (1999). "Differential Incentives - Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation." Public Opinion Quarterly **63**(2): 251-260.
- Singer, E., J. Van Hoewyk and M. P. Maher (2000). "Experiments with Incentives in Telephone Surveys." Public Opinion Quarterly **64**(2): 171-188.
- Smith, T. W. (1983). "An Experimental Comparison between Clustered and Scattered Scale Items." Social Psychology Quarterly **46**: 163-168.
- Smith, T. W. (1995). Little Things Matter; a Sampler of How Differences in Questionnaire Format Can Affect Survey Responses. Proceedings of the American Statistical Association, Survey Research Methods Section.
- Spearman, C. E. (1904). "'General Intelligence" Objectively Determined and Measured." American Journal of Psychology **5**: 201-293.
- Spearman, C. E. (1904). "Proof and Measurement of Association between Two Things." American Journal of Psychology **15**: 72-101.
- Strack, F. and L. L. Martin (1987). Thinking, Judging and Communicating: A Process Account of Context Effects in Attitude Surveys. Social Information Processing and Survey Methodology. H.-J. Hippler, N. Schwarz and S. Sudman. New York, Springer-Verlag.
- Stussman, B. J., B. Taylor and H. Riddick (2003). Partials and Break-Offs in the National Health Interview Survey, 2002. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

- Sudman, S. and N. M. Bradburn (1982). Asking Questions. San Francisco, Jossey-Bass, Inc.
- Sudman, S., N. M. Bradburn and N. Schwarz (1996). Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco, Jossey-Bass Publishers.
- Tarkkonen, L. and K. Vehkalahti (2005). "Measurement Errors in Multivariate Measurement Scales." Journal Of Multivariate Analysis **96**(1): 172-189.
- Tidwell, J. (2005). Designing Interfaces: Patterns for Effective Interaction Design, O'Reilly Media.
- Toepoel, V., M. Das and A. v. Soest (2005). Design of Web Questionnaires: A Test for Number of Items Per Screen, Tilburg University. **2006**.
- Tortora, R. and K. Crank (1978). The Use of Unequal Probability Sampling to Reduce Respondent Burden. American Statistical Association Proceedings of the Survey Research Methods Section.
- Tourangeau, R. (1984). Cognitive Science and Cognitive Methods. Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines. T. Jabine, M. L. Straff, J. M. Tanur and R. Tourangeau. Washington, DC, National Academy Press: 73-100.
- Tourangeau, R. (1987). Attitude Measurement: A Cognitive Perspective. Social Information Processing and Survey Methodology. H.-J. Hippler, N. Schwarz and S. Sudman. New York, Springer-Verlag: 149–162.

- Tourangeau, R., M. P. Couper and F. Conrad (2004). "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." Public Opinion Quarterly **68**(3): 368-393.
- Tourangeau, R., L. J. Rips and K. A. Rasinski (2000). The Psychology of Survey Response. Cambridge, U.K.; New York, Cambridge University Press.
- Treat, J. B. (1996). The Effect of Questionnaire Length on Response. American Statistical Association Proceedings of the Survey Research Methods Section.
- Trouteaud, A. R. (2004). "How You Ask Counts: A Test of Internet-Related Components of Response Rates to a Web-Based Survey." Social Science Computer Review **22**(3): 385-392.
- Vacha-Haase, T. (1998). "Reliability Generalization: Exploring Variance in Measurement Error Affecting Score Reliability across Studies." Educational And Psychological Measurement **58**(1): 6-20.
- Vehovar, V., K. L. Manfreda and Z. Batagelj (2000). Design Issues in Web Surveys. Proceedings of the Survey Research Methods Section of the American Statistical Association.
- Weinstein, A. G. (1972). "Predicting Behavior from Attitudes." Public Opinion Quarterly **36**(3): 355-360.
- Zimmerman, D. W., B. D. Zumbo and C. Lalonde (1993). "Coefficient Alpha as an Estimate of Test Reliability under Violation of 2 Assumptions." Educational And Psychological Measurement **53**(1): 33-49.