University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Survey Research and Methodology program (SRAM) - Dissertations & Theses

Survey Research And Methodology Program

8-1-2012

Numeric Estimation and Response Options: An Examination of the Measurement Properties of Numeric and Vague Quantifier Responses

Mohammad T. Al Baghal
University of Nebraska-Lincoln, mtalbaghal@hotmail.com

Follow this and additional works at: http://digitalcommons.unl.edu/sramdiss

Part of the Other Social and Behavioral Sciences Commons, and the Quantitative, Qualitative, Comparative, and Historical Methodologies Commons

Al Baghal, Mohammad T., "Numeric Estimation and Response Options: An Examination of the Measurement Properties of Numeric and Vague Quantifier Responses" (2012). *Survey Research and Methodology program (SRAM) - Dissertations & Theses.* Paper 5.
http://digitalcommons.unl.edu/sramdiss/5

This Article is brought to you for free and open access by the Survey Research And Methodology Program at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Survey Research and Methodology program (SRAM) - Dissertations & Theses by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

NUMERIC ESTIMATION AND RESPONSE OPTIONS: AN EXAMINATION OF THE MEASUREMENT PROPERTIES OF NUMERIC AND VAGUE QUANTIFIER RESPONSES

By

Mohammad Tarek Al Baghal

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Survey Research & Methodology

Under the Supervision of Professor Robert Belli

Lincoln, Nebraska

August, 2012

NUMERIC ESTIMATION AND RESPONSE OPTIONS: AN EXAMINATION OF THE MEASUREMENT PROPERTIES OF NUMERIC AND VAGUE QUANTIFIER RESPONSES

Mohammad Tarek Al Baghal, Ph.D.

University of Nebraska, 2012

Advisor: Robert Belli

Many survey questions ask respondents to provide responses that contain quantitative information. These questions are often asked requiring open ended numeric responses, while others have been asked using vague quantifier scales. How these questions are asked, particularly in terms of the response format, can have an important impact on the data. Therefore, the response format is of particular importance for ensuring that any use of the data contains the best possible information. Generally, survey researchers have argued against the use of vague quantifier scales. This dissertation compares various measurement properties between numeric open ended and vague quantifier responses, using three studies containing questions with both formats. The first study examines uses new experimental data to compare accuracy between the measures; the second and third use existing data to compare predictive validity of the two formats, with one examining behavioral reports, the other examining subjective probabilities. All three studies examine the logical consistency between measures, and the potential correlates related to improved measurement properties. Importantly, these studies examine the influence of numeracy, a potentially important but rarely examined variable. The results of the three studies indicate that vague quantifiers may have better

measurement properties than numeric open ended responses, contrary to many researchers' arguments. Studies 2 and 3 are most clear about this increased strength; in both of the studies, using a number of tests, the predictive validity of vague quantifiers was consistently greater than that of numeric open ended responses, regardless of numeracy level. Study 1 shows that at that generally, vague quantifiers result in more accurate data than numeric, but this finding depends on other factors, such as numeracy. Therefore, numeracy was infrequently found to be important, but at times did have an impact on accuracy. Further, in the three studies, it was found that the two formats were logically consistent when translations between the questions were directly asked for, but inconsistency occurred when there was not a direct translation.

Acknowledgements

I would like to acknowledge the Indiana University Center for Postsecondary Research for their permission to use their experimental data set used in Study 2. I would also like to thank the members of my committee, who have been helpful and supportive in this process. Specifically, I thank my advisor, Robert Belli, Allan McCutcheon, Jolene Smyth, and Brian Bornstein.

I would also like to thank those who have helped me in other phases of the dissertation process, including Lynn Phillips, Nick Ruther, Mathew Stange, and Bryan Parkhurst.

Finally, I would like to thank my wife, Mubeena Siddiqi for all of the necessary support and kind words needed to make it through this project.

Table of Contents

Introduction 1

Dissertation Structure 5

Research Objective 5

Objective 1: Accuracy of Different Response Formats 6

Objective 2: Predictive Validity of Different Response Formats 7

Objective 3: Logical Consistency between Measures 7

Objective 4: Circumstances of Differential Performance Different Response Formats 8

Problem Statement and Significance 8

Objective 1: Accuracy of Different Response Formats 12

Objective 2: Predictive Validity of Different Response Formats 12

Objective 3: Logical Consistency between Measures 13

Objective 4: Circumstances of Differential Performance Different Response Formats 14

Literature Review 15

Outline of the Literature Review 15

Numeracy Introduction 16

Cognition of Numeric Information 17

Understanding of Numeric Information 18

Frequency Estimation Introduction 21

Strategies for Recall of Frequencies 22

Impact of Task on Strategy Selection and Frequency Estimation 26

Content of the Frequency 27

Content of the Request 29

Measurement of Frequencies 30

Accuracy 35

Psychology of Subjective Probabilities and Risk Perception
Introduction 38

Numeric Estimates of Risk Perception 41

Vague Quantifier Measures of Risk Perceptions 46

Summary and Gaps in the Literature 49

Study 1 58

Data and Methods 58

Hypotheses 65

Results 66

Numeracy 66

Logical Consistency 67

Accuracy 69

Discussions and Conclusions 91

Study 2 99

Data and Methods 99

Hypotheses 105

Results 106

Data Management 106

Effects of Clustering 108

Logical Consistency 109

Predictive Validity 118

Discussions and Conclusions 132

Study 3 137

Data and Methods 137

Hypotheses 141

Results 142

Measures of Risk 142

Logical Consistency 144

Predictive Validity 149

Discussion and Conclusions 161

Summary 165

References 174

List of Tables

Table 1 Mean Absolute Risk Estimates of Lung Career Risk for Smokers 44

Table 1.1 Numeric Translations for Vague Quantifier Scale 68

Table 1.2 Regression on Respondents' Slopes 72

Table 1.3 Regression on Respondents' Correlations 74

Table 1.4 Signed Differences at Levels of Actual Frequency, by Form and Context 75

Table 1.5 Hierarchical Linear Model of Signed Differences on Response and Respondent Characteristics 80

Table 1.6 Predicted Mean Signed Error by Context, Form, and Numeracy 81

Table 1.7 Absolute Differences at Levels of Actual Frequency, by Form and Context 86

Table 1.8 Hierarchical Linear Model of Absolute Differences on Response and Respondent Characteristics 88

Table 1.9 Predicted Mean Absolute Error by Context, Form, and Numeracy 89

Table 2.1 Active-Collaborative Learning Scale (ACLS) and Student-Faculty Interaction Scale (SFIS) Means 107

Table 2.2 Number of Times Asked Question in Class by Vague Quantifier Response 110

Table 2.3 Number of Times Made a Class Presentation by Vague Quantifier Response 110

Table 2.4 Number of Times Worked with Other Students during Class by Vague Quantifier Response 111

Table 2.5 Number of Times Worked with Other Students Outside Class by Vague Quantifier Response 111

Table 2.6 Number of Times Tutored or Taught Other Students by Vague Quantifier Response 112

Table 2.7 Number of Times Participated in a Community-based Project by Vague Quantifier Response 112

Table 2.8 Number of Times Discussed Ideas from Classes with Others Outside of Class by Vague Quantifier Response 113

Table 2.9 Number of Times Discussed Grades or Assignments with an Instructor by Vague Quantifier Response 113

Table 2.10 Number of Times Talked About Career Plans with a Faculty Member or Advisor by Vague Quantifier Response 114

Table 2.11 Number of Times Received Prompt Feedback from Faculty on Academic Performance by Vague Quantifier Response 114

Table 2.12 Number of Times Worked with Faculty Members on Activities Other than Coursework by Vague Quantifier Response 115

Table 2.13 Number of Times Discussed Ideas from Classes with Faculty Members Outside of Class by Vague Quantifier Response 115

Table 2.14 Correlation of Active-Collaborative Learning Scales with Grades 120

Table 2.15 Correlation of Active-Collaborative Learning Scales with College Experience Rating 120

Table 2.16 Correlation of Active-Collaborative Learning Scales with Same College Preference 120

Table 2.17 Correlation of Student-Faculty Interaction Scales with Grades 121

Table 2.18 Correlation of Student-Faculty Interaction Scales with College Experience Rating 121

Table 2.19 Correlation of Student-Faculty Interaction Scales with Same College Preference 121

Table 2.20 Logistic Regression Indicators of ACLS on Grades 125

Table 2.21 Logistic Regression Indicators of ACLS on College Experience Rating 125

Table 2.22 Logistic Regression Indicators of ACLS on Same College Preference 126

Table 2.23 Logistic Regression Indicators of SFIS on Grades 126

Table 2.24 Logistic Regression Indicators of SFIS on College Experience Rating 126

Table 2.25 Logistic Regression Indicators of SFIS on Same College Preference 127

Table 2.26 Scale Coefficients and Odds Ratios of ACLS on Grades 129

Table 2.27 Scale Coefficients and Odds Ratios of ACLS on College Experience Rating 129

Table 2.28 Scale Coefficients and Odds Ratios of ACLS on Same College Preference 130

Table 2.29 Scale Coefficients and Odds Ratios of SFIS on Grades 130

Table 2.30 Scale Coefficients and Odds Ratios of SFIS on College Experience Rating 131

Table 2.31 Scale Coefficients and Odds Ratios of SFIS on Same College Preference 131

Table 3.1 Risk Measure Tendencies 143

Table 3.2 Absolute Risk Means by Vague Quantifier Response 145

Table 3.3 Correlation of Risk Measures with Smoking Status 150

Table 3.4 Logistic Regression Models Fit Indicators by Age 154

Table 3.5 Logistic Regression Models Fit Indicators by Education (Numeracy) 155

Table 3.6 Logistic Regression Models Fit Indicators by Total Sample 155

Table 3.7 Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking by Age 159

Table 3.8 Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking by Education (Numeracy) 159

Table 3.9 Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking by Education (Numeracy) 160

List of Figures

Figure 1 Example Question for Absolute Frequency Intended by Vague Quantifier..... 101

Figure 2 Example Screen of How Vague Quantifier Questions Asked 208

List of Appendices

Appendix 1 Questions about Smoking Risks used in Literature Review 190

Appendix 2 Target Words and Exemplars for Use in Study 1 193

Appendix 3 Numeracy test from Galesic and Garcia-Retamero (2010)..... 203

Appendix 4 Questions Used from the NSSE 206

Appendix 5 Questions Used from Annenberg 2 Surveys 210

Introduction

Surveys collect data on a wide range of information that is important for not only academic research but a range of other areas, such as marketing and policy decisions. For example, questions about television use may affect the decisions a business makes and how to invest money. Asking about how many times a person takes pain medication may affect drug policy. Questions about subjective beliefs, such as the risks of smoking can be used to make decisions about public health campaigns. Importantly, all of these questions ask respondents to provide responses that contain numeric information. Responses to these queries can be affected by the way the question design allows an answer. How these questions are asked, particularly in terms of the response format, can have an important impact on the data, including the response distribution and the overall quality of the responses. This quality can be expressed in different ways, most importantly as the level of measurement error in each response format (Biemer and Lyberg 2003, Groves 1989, Lessler and Kalsbeek 1992). As such, the response format is of particular importance for ensuring that any use of the data contains the best possible information on which to make decisions.

Survey questions provide methods for respondents to give their answers, either in their own words (open ended questions) or through response alternatives provided by the researcher (closed ended questions) (Converse and Presser 1986). The choice between open and closed ended and the choice of which alternatives to offer in a closed ended question can have distinct effects on the response distributions (Schuman and Presser 1981). That there is an impact of question format on response distributions is likely true for all survey questions and question types, including questions about quantities

regardless of whether the questions ask about frequencies, such as for behaviors, or for expressions of subjective probabilities, such as those for level of perceived risk.

Three main response options have been developed and used in requesting quantitative information from respondents: open ended, numeric scales, or vague quantifier scales (Tourangeau et al. 2000). The open ended approach, which has been used most frequently by survey designers, leaves the response formulation and reporting to the respondent, whereby, for numeric questions, the respondent generally responds with one number. For scale responses, respondents choose a scale point most closely associated with their formulated answer. For objective measures, such as frequencies, the response options can include scale points for distinct, individual values, (e.g. 1 time) or for ranges of values (e.g. 1 to 5 times). Subjective probabilities may also be elicited using open ended questions, in which the expected probability is directly asked for (e.g. Viscusi 1990). Both open ended and numeric scale options presume that the respondent has some numeric understanding and representations of the requested information in numeric form in order to respond (Schwarz et al. 1985). However, some have argued against the use of numeric scales, as it may bias respondent answers as the scale provides not only a measurement device but also an informative component as well (Schwarz et al. 1985).

The last response format frequently used is vague quantifier scales. These scales use no numeric values directly, but rather use verbal phrases frequently used in natural language to describe numeric data (Sanford et al. 1994, 1996). These scales provide options that are, as the name suggests, inherently vague. For example, scale options may include words and phrases such as very often, somewhat often, and not very often. As such, there is often a large variation in the numeric meaning assigned to vague quantifiers

(e.g. Wallsten et al. 1985). The scales also have relative meaning, such as where on the scale a respondent believes they are in comparison to similar others (Schaeffer 1991). As such, it has been argued against the usage of such scales when it is possible to use numeric response options (especially open ended) instead (Beyth-Marom 1982, Schaeffer 1991, Tourangeau et al. 2000).

In considering the formatting of questions that ask for numeric information, an issue that deserves additional research is to better determine how quantitative information is represented cognitively. Matching response options to cognitive structure appears to be a possible best practice in survey methodology (Tourangeau et al. 2000). Included in this cognitive structure is the way that quantitative information is used when judging the frequency of the occurrence of events and in providing subjective probabilities such as assessing risks. Hence, if people think about quantitative information in vague ways, asking them to quantify using vague quantifier scales may be more appropriate in comparison to other formats. For example, given that the risk of death for a surgery is 1 in 1000, what may matter more than the numeric representation is the vague representation of whether this risk is “a little” or “a lot”. Similar vague representations may also underlie judgments of behavioral frequencies.

As presented in the literature review below, there are reasons to believe that vague quantifiers are indeed better measures than numeric indicators. First, it is not clear that people are able to think numerically in a range of instances. In general, there is a lack of numeracy (numeric literacy) in the population (e.g. Galesic and Garcia-Retamero 2010). Theories such as “fuzzy-trace” and other dual-process theories suggest that people frequently rely on vague, intuitive representations of numeric information rather than on

the verbatim representation of the numbers (Reyna and Brainerd 2008). Correlations between subjective beliefs and behaviors have been found to be higher when using vague quantifier scales than numeric responses (Windschitl and Wells 1996). Further, it has been noted it is more cognitively burdensome to ask about numeric information than vague quantifiers (Bradburn and Miles 1979).

Still, there is a dearth of research on whether vague quantifiers or numeric responses perform better in regards to validity. Studies examining frequencies have focused on the meanings and variations in these meanings people have placed on vague quantifiers, rather than which format performs better in predicting accuracy. Further, these studies have not examined which format is more strongly related to other variables of interest, i.e. predictive validity. In regards to subjective probabilities, there has been some work on the predictive validity of the different measures, but these studies have been limited to bivariate measures of association, and have been conducted on nonrandom and small samples (i.e. college students in a class).

The goal of the research described in this dissertation is directed to filling some of these gaps in the research on vague quantifiers and numeric responses. In particular, the ultimate goal is to provide evidence as to which format of response to quantitative queries, numeric open-ended versus vague quantifier, provides better data and in which circumstances. I will use three different data sets, all of which contain responses to similar questions using both response formats, numeric open ended responses and vague quantifier responses. Two of these data sets are related to frequencies; one is related to subjective probabilities. Two of the data sets come from national samples, while one is an experiment that has been designed specifically for this dissertation.

Dissertation Structure

This dissertation is organized into six sections. The first section briefly lists the four major research objectives of this investigation. The objectives are followed by the problem statement and significance of the research, including a brief overview of the gaps in the literature, and a brief description of problems and significance of each of the four objectives. The second section is a review of the existing literature regarding the areas of numeracy, frequency estimation, and the subjective probability estimation, concluding with a summary and a more extended discussion of the gaps in the literature in regards to the current research. Following the review, there are three separate sections outlining three studies. Three data sets are needed to look at the various aspects of numeric estimation and facets of data quality, and each section describes the data and methods, the hypotheses, the results from the analyses, and the discussion and conclusion for the respective study. The first two data sets examine frequency estimation and response formats, while the third data set examines subjective probability estimation. The final section presents a summary of the research including recommendations for survey research.

Research Objective

The main objective of this research is to examine which type of response formats, numeric open ended or vague quantifier scales, have better measurement properties for questions requesting numeric data, that is, which has the least measurement error. Numeric scales are not examined by the three studies for several reasons. First, both numeric scales and numeric open ended responses require some numeric understanding and cognitive representation of the information requested by the survey question.

However, only one of these response forms is needed to compare responses requiring a numeric representation versus those that do not. Second, previous research has found numeric scales to introduce systematic biases in respondents' behavioral reports. Specifically, Schwarz et al. (1985) suggest that numeric open ended response formats are preferable to numeric scales, as numeric scales "may introduce systematic bias in respondents' behavioral reports and related judgments, because response scales are not only measurement devices but serve informative functions as well" (Schwarz et al. 1985 p. 394). Third, of practical importance, much of the data available that allow comparisons to be made only have data on numeric open ended responses and vague quantifier responses, barring comparison with numeric scales. For these reasons, all of the data examined will compare only numeric open ended and vague quantifier responses.

This main objective entails first, which format, numeric open ended or vague quantifier scales, is more accurate, and second, which has higher predictive validity. A third objective is to see if the semantic representations of numbers and the actual numeric translation of these are logically consistent. Finally, these objectives include an examination of generalizability, that is, in which circumstances these findings hold. These objectives are divided across frequency and subjective probability data. The research objectives are presented more expressly below:

Objective 1: Accuracy of Different Response Formats

The aim of objective 1 is to examine whether one response format, numeric open ended or vague quantifier scales, is more accurate than the other in relation to verifiable data. Since by definition vague quantifiers do not convey an exact meaning, numeric values must be assigned to gauge accuracy, as in Lu et al. (2008). If one is more accurate

than the other this possibly suggests the cognitive structure of quantitative data and measurement properties of the different response formats. That is, more accurate data may suggest that the quantitative information is stored in a way more closely resembling one response format (numeric versus vague), and in addition, accuracy is suggestive of which response format has better measurement properties. Since accuracy needs factual supportive data, accuracy data relies on frequency data.

Objective 2: Predictive Validity of Different Response Formats

The aim of objective 2 is to determine which response format, numeric open ended or vague quantifier scales**,** if either, is stronger in terms of predictive validity, i.e. the relationship of the target response to other theoretically related variables. In many instances it is the relationships between variables that may be of interest rather than the simple univariate statistics, and different question formulations have found differential predictive validity (e.g. Chang and Krosnick 2003). In comparison to accuracy measures, assigning numeric values when examining predictive validity is less necessary, and predictive validity can be tested using both frequency and subjective probability data.

Objective 3: Logical Consistency Between Measures

The aim of objective 3 is to inspect the numeric translations of vague quantifiers and identify whether the numeric responses and vague quantifiers given are consistent. It may be that there is a logical incoherence in the translated numeric meaning and the vague quantifier given. In such cases it is necessary to posit whether one measure rather than the other is more likely to be more consistent with cognitive-semantic processes. The examination of logical consistency is closely related to determining face validity.

Objective 4: Circumstances of Differential Performance of Different Response Formats

The aim of objective 4 is to investigate in what cases the accuracy and predictive validity differs for the varying response formats, and in which direction. It is likely that in some cases the accuracy and/or predictive validity is stronger for one measure, but that the other response format has better measurement properties for a different measure.

Problem Statement and Significance

The reporting of a response to a survey question, given different response formats, is an issue mainly of measurement. As such, in the Total Survey Error paradigm, the source of error mainly of interest when dealing with response formats is measurement error (Biemer and Lyberg 2003, Groves 1989, Lessler and Kalsbeek 1992). Measurement error refers to the difference between the true value and the survey estimate (Lessler and Kalsbeek 1992). This measurement error can arise from a number of sources, including the interviewer, the questionnaire mode, the questionnaire, and the respondent. Although all may interact with response formats, only the questionnaire and the respondent are of interest in the current dissertation (although it will be noted when other sources may be a cause of error).

Measurement error due to the questionnaire in this case would be due to the provision of inadequate or difficult response options, in which the response options do not match the information stored in memory or is difficult for the respondents to match to the information requested. This measurement error can occur due to a mismatch in the format and the natural way that respondents' store information, or if asking a question in one form or the other increases cognitive difficulty. From the respondent side of the

equation, the error occurs due to several different sources, which arise from the response process taken during survey response. The response process includes the steps of encoding of the requested information, comprehension of the question, retrieval of the requested information, judgment of the relevance and completeness of memories, and giving a response, which may require editing (Biemer and Lyberg 2003, Tourangeau et al. 2000).

The first step of the response process is how the respondent initially encodes the information that will be requested by the questionnaire. For quantitative information, initially the information may not be encoded at all, only some of the information is encoded, some estimate is used, all information is encoded, or only vague notions are stored (see below in literature review). At the second stage, respondents need to understand the question and how to respond given the response format, but may not, including the meaning of vague terms (Groves 1989). At the third stage, numeric information must be recalled in some format, generally based on the way the information was encoded. However, it may not be the case that information is recalled as it was initially stored, as information may be forgotten or only vague representations remain (see literature review). The fourth stage, judgment, requires assessment of the recalled information for completeness. An error can occur, for example, when the recalled information is judged complete when it is not. The response step requires that the numeric information recalled and judged complete is reported given the response format. However, if the response format is not matched to how the information is encoded or recalled, then the mismatch can create error. Finally, respondents may skip some of these steps entirely (i.e. satisficing), which can increase error as well (Krosnick 1991).

Given all of these potential sources of error, it is important to identify response formats that lead to the least amount of error as possible. Numeric open ended responses or vague quantifiers are two alternatives which are often used in measuring information. As noted above and more fully below in the literature review, both response formats have theory and findings suggesting one may be preferable, and are important in the measurement properties the question, especially in terms of validity. Further, although there is evidence people may naturally store quantitative information as vague quantifiers, practitioners at times recommend against using vague quantifier response options in questions. This avoidance may be in part due to a lack of research showing the comparative measurement properties of the numeric and vague quantifier response formats. As of now, there are few, if any, studies that examine the comparative measurement properties of the two response formats. The goal of this dissertation is to fill this gap in the literature, and assess the validity of the different response formats via accuracy and predictive validity, and the instances that the different response formats perform better on these assessments. By examining different forms of validity and the contexts in which the measures display better or worse measurement properties, the results are hoped to triangulate to some conclusions about which measures should be used and when.

Logical consistency of responses to the different formats has been studied more than other aspects of the measurement properties of response formats, such as accuracy and predictive validity; however, some important gaps still exist in the literature, particularly in the survey methodology literature. First, is the examination of whether numeric translations of vague quantifier are linguistically logical. For example, it should

be examined if the response to the vague quantifier response option “never” is always equal to zero. In addition, it follows that respondents should always give a translation for quantifiers at higher points on the scales significantly greater than points lower on the scale. However, this logical ordering has not been studied completely, and is another gap in the literature that this research proposes to address.

There are a number of other gaps in the literature that exist, in particular regarding the respondent and other characteristics that may interact with response formats to affect the differences in measurement properties. Most prominently is a lack of the examination of numeracy on responses to questions about quantitative information, and how numeracy interacts with the different response formats to affect measurement properties, including accuracy, predictive validity, and logical consistency of responses. Another gap in the literature is the examination of how aspects of memory for frequency interact with the response formats used in the question. These aspects of memory include the actual frequency of the event, and the context memory. For subjective probability estimation in particular, there has been a lack of studies examining diverse and representative populations, which limit examination of some important contrasts, in particular the possible effect of maturity (age) on estimation in terms of its predictive validity. Adults and youths may perceive risks in different ways, such that adults use vague, intuitive information, while youths rely more on numeric, objective data (Reyna and Farley 2006). The difference in what information is used in risk perceptions in turn may affect the measurement properties of the different response formats. However, the difference between ages has not been examined systematically, and is another gap in the literature the research discussed in this dissertation attempts to address.

Objective 1: Accuracy of Different Response Formats

Accuracy is an aspect of validity in that both are concerned with the differences between the true value and the estimated value (although validity entails other aspects as well) (Groves 1989, Lessler and Kalsbeek 1992). It is also one of the main components of the quality of a survey (Biemer and Lyberg 2003). The expected value of the survey estimate can be put into equation format as follows:

$$y = \mu + \varepsilon$$

where y is the survey estimate, μ is the true value, and ε is the error. Accuracy can be defined most simply where $\varepsilon = 0$, and the survey estimate is equal to the true value. More broadly, it can be defined as to where the value of ε is minimized. Therefore the goal is to choose measures, including response formats, in which the value of ε is minimized. As of now, few studies have examined accuracy in this way.

Objective 2: Predictive Validity of Different Response Formats

Predictive validity, as the name suggests, is another form of validity on which to assess the different measurement formats. Unlike accuracy, which compares the measure to the true value, i.e. a gold standard, predictive validity is concerned with the relationship of the target measure with some other variable it should theoretically be related to, a criterion (Kumar 2005). A high level of relationship (e.g. correlation) between the target and the criterion suggests a higher level of validity, as the variable is able to predict what it should theoretically predict. As such, the stronger the relationship between one target and the criterion compared to other target variables suggests better measurement properties for the target with the higher relationship. Predictive validity tests have been used to examine measurement properties of competing question formats

(Chang and Krosnick 2003). However, few studies have made comparisons between numeric measures and vague quantifiers in terms of predictive validity, and in most studies in which predictive validity has been examined, the sample is small and not randomly selected from the wider population.

Objective 3: Logical Consistency Between Measures

If one is asked similar questions with different response formats that map onto the same underlying dimension, it should follow that there is a natural consistency between the two responses. That is, higher responses on one response scale should be related to higher responses on the other; similarly, lower responses on one should relate to lower response on the other response scale. Thus the numeric open-ended response should be greater for responses in which the highest point on the vague quantifier scale was given than all other numeric open ended responses in which lower vague quantifier scale points were selected.

Further, responses should be related in a semantic-logical format. Vague quantifiers are natural part of everyday language (Sanford et al. 1996). Conversely, a large number of people have problems with numeric responses (Galesic and Garcia-Retamero 2010, Reyna and Brainerd 2008). As such, it may be that numeric answers are not consistent with the natural language usage of quantifiers. For example, if the response “never” is given on a vague quantifier scale and the answer “5” given on the numeric open ended response, there is an apparent inconsistency on the common understanding of what “never” means. A lack of consistency suggests that it is possible that one measure has better measurement properties than the other.

Objective 4: Circumstances of Differential Performance of Different Response Formats

It may be that either numeric open ended measures or vague quantifiers have better measurement properties overall, or that there are different circumstances when one performs better than the other. First, one measure may be more accurate, but, however unlikely, have lower predictive validity. Second, it may be the case that one measure has better measurement properties for frequency data than subjective probabilities. Third, it may depend on the actual level of frequency that affects cognition of the numeric information and hence the measurement properties of the two response formats. For example, greater frequencies may be more likely to be stored as vague quantifiers relative to smaller frequencies (Conrad et al. 1998).

Fourth, greater context of encoding and memory may lead respondents to be more likely to have numeric information stored in one format versus another. For instance, Brown (1997) found that those with greater context memory (i.e., memory based on greater distinctiveness among target events) were more likely to use a strategy of enumeration to answer numeric open ended questions than when context memory was low (i.e., memory based on greater similarity among target events). Fifth, the way information is stored and preferred to be recalled and reported may depend on the level of the individual's numeracy. More numerate individuals may be more likely to have numeric information stored as numbers, and more comfortable recalling and reporting numbers, which may lead to differential measurement properties for the different response formats across numeracy levels. Similarly, characteristics that may be related to numeracy or other individuating factors may lead to different outcomes of the response

formats. These characteristics include demographic variables such as age, gender, race, and education. Further, it is possible that the adults are more likely to rely on the vague, gist-based information compared to adolescents, who rely more on verbatim information (Reyna and Farley 2006). As such, there may be even greater differential functioning of the response formats across age, and when possible, will also be examined.

Literature Review

Outline of the Literature Review

As noted, responses to queries can be affected by the response formats that are given as answer choices. Most studies of frequencies use numeric responses as the outcome, as do many studies of subjective probabilities. The literature review examines the pertinent areas related to the way people think about quantitative information (in both numeric and vague terms), how this information is recalled, and how response formats can affect responses. As such, the literature review is divided into three main sections: numeracy, frequency estimation, and subjective probability estimation (risk perceptions). Within each, the topics are divided according to different aspects of the literature. Each of these three sections begins with a short introduction on the topic. Numeracy, underlying the way people understand numbers, and hence both frequency estimation and subjective probabilities begins, followed by the literature review on frequency estimation and concluding with the literature on subjective probabilities. Studies 1 and 2 (see below) deal with frequency estimation and more related to the frequency estimation literature, while Study 3 deals with subjective probabilities, and is linked more closely to the subjective probability literature. All three studies examine numeracy's impact on

response validity. The literature review concludes with a summary and a discussion of the current gaps in the literature as it relates to these topics.

Numeracy Introduction

People use and encounter numbers on a regular basis. Numbers are part of our general information environment, that they include such aspects as prices for goods, speed limits, and instructions (e.g. for baking, construction), to name just a few instances. This numeric information is important for understanding the environment and in decision making. The decisions made using numeric information can be of paramount importance, including financial management and health care. Researchers studying these and other areas, including most, if not all, social sciences, often ask subjects for numeric information on a variety of topics, including frequency estimation, risk estimation, and monetary decisions and information. Even the response scales respondents are required to respond to are often anchored and labeled using numeric information, e.g. 0 to 10, -5 to +5.

People's ability to think numerically is assumed when such data is requested and frequently when decision making is involved. Data from surveys can impact important policy decisions (e.g. public education initiatives), marketing strategies, or social research. Similarly, numeric information is frequently presented for people to make financial, public policy, and health care decisions. As such, it is important to examine the basic assumption of capability for numerical thinking, and if this assumption does not hold, then the decisions made and data from surveys will be suboptimal at best. Understanding how numerical data is understood, produced and reported is needed to potentially increase the efficacy of decision-making and data coming from surveys.

Generally, numerical thinking is related to both cognitive understanding and memory (Brown and Siegler 1993). Numbers must be understood in the proper metric and meaning for useful processing of the information. Memory is important in both understanding the numbers presented, as it often provides the metric needed, but also frequently in recall of personal numeric data. How such data is encoded, stored, and its accessibility and availability at later points are important issues in understanding memory broadly. Further, when the data requested from subjects is numeric (e.g. behavior frequency), the tabulation of the memories is also of interest. As such, the focus of this work is the cognition and memory aspects of numeric information. Although previous research on cognition of numeric information is expansive, a basic summary of major points can be suggestive of numeric thinking.

Cognition of numeric information

fMRI studies have shown a link between exact arithmetic to language centers of the brain (Dehaene et al. 1999). As such, it is necessarily a human-based structure, incompatible with infants and not completely compatible for youths, who have less language (and memory) ability. Conversely, this same study found that approximate mathematics shows language independence, relying more on numerical magnitudes and spatial processes. The authors suggest that mathematical intuition may rely on an interaction of these cognitive systems.

Once a basic understanding of numbers is obtained, it is reasonable to question how these numbers are mentally represented. It has been posited that numbers are represented on a mental number line (Dehaene 1997, Nuerk et al. 2004). Differentiation of numbers on this line can be made using several identified processes. The first is

magnitude differences (Moyer and Landauer 1967, Nuerk et al. 2004, Turconi et al. 2006). As such, as the magnitude of the difference between the compared numbers increases, the ease of differentiation increases. An additional differentiation process is that of serial-search, in which serial position is accessed in judgments (Turconi et al. 2006).

In processing and producing numeric information, additional information is also used. Estimation of numeric information first requires knowledge of the metric and information about the distribution of the domain of interest (Brown and Siegler 1993). For example, estimating driving speed requires information of both the metric (e.g. miles per hour) and the distribution of normal speeds (e.g. 35 M.P.H. in cities, 70 on highways). These sources of information are obviously related, but are conceptually distinct. Further, error can arise from either one source of information separately, or from both. In addition, lacking information from either necessary source increases the reliance on heuristics in judgments; when greater domain-specific information is available from both sources, such information is emphasized over heuristics in making judgments (Brown and Siegler 1993).

Understanding of numeric information

The above research only indicates that people do in fact have some cognitive processes for numbers. This research, however, does not indicate that people in general are able to understand numbers is a meaningful way. Indeed, much research suggests that people do not fully understand numeric information. The National Literacy Survey indicates that children lack minimum math skills needed to use numbers in printed materials (Kirsch et al. 2002). Even among highly educated samples, only 32% correctly

answered eight numeracy skill questions (Lipkus et al. 2001). Nearly 20% of these samples answered the most direct risk question incorrectly (i.e. “Which represents the largest risk? 1%, 5% or 10%?”). Studies using similar numeracy scales have found similar results (Schwartz et al. 1997, Peters et al. 2006, Galesic et al. 2009). In the largest identified study of numeracy, Galesic and Garcia-Retamero (2010) conducted surveys using nine questions for objective numeracy (also used in this dissertation in Study 1). This research found that people could answer only about two-thirds of the questions correctly. Importantly, the findings also show a strong correlation with education, with more educated people being more highly numerate.

Studies in survey research also indicate the inability of people to use numbers as expected. A number of respondents do not answer with logically consistent responses to subjective probability measures using numeric scales, particularly among those with lower cognitive abilities (Belli et al. 1999). People have also been found to incorrectly select greater magnitude probabilities when frequency and base rates differed (e.g. 24.14 out of 100 versus 2414 out of 10000) by choosing cases where the higher total frequency was greater, neglecting the base rate (Yamagishi 1997). Similarly, when asking about equivalent risks, but in a different manner (e.g. frequencies vs. percentages), people have been found to give different answers (Windschitl 2002).

Another finding in the survey literature is the influence that numeric labels placed on scales has on response (e.g. varying levels of number of hours of television watched at the end points of the scale). Studies have found that using different values produces differing estimates of frequency or duration (Schwarz et al. 1985; Wright et al. 1994). People use the information conveyed in the numeric scale points to infer an “average” for

the population, and to place themselves on the scale accordingly. The authors of these studies suggest that numeric scales possibly be avoided; instead, numeric open ended responses should be used.

In addition, asking a belief question using differing end points of the scale produces different results. For example, asking the same attitude question using a 1 - 7 scale consistently produces different results when using a -3 to 3 scale (Tourangeau et al. 2000). The difference in results is evidently due to people's aversion to negative numbers. Further, Schwarz et al (1991) found that the choice of numeric labels can facilitate or dilute polarity implications of endpoints provided to respondents. For example, negative values may emphasize the differences more on a bipolar scale, whereas the lack of negative numbers provides the impression that the underlying scale is unipolar. The authors examined the polarity issue in a number of studies in which respondents were randomly assigned to different conditions; both were asked the same questions, but one group was asked to respond on a 0-10 scale while the other was asked to answer on a -5 to +5 scale. Since each scale contained the same number of points, the differences suggest that people infer scale polarity based on the numeric end-points added to the scale.

If people do not understand numbers, how then do they understand the numeric information that they consistently encounter? It appears that individuals understand, and prefer to communicate, numeric information in vague and intuitive ways, rather than the concrete manner numbers entail (Peters et al. 2006, Reyna and Brainerd 2008). This understanding and preference does not mean that people do not use rational (i.e. numeric) reasoning when making decisions; rather these are made using dual-processes, relying on

both intuitive and/or rational processes. One prominent dual-process model is fuzzy-trace theory (Reyna and Brainerd 1991). This theory assumes that people rely on their memories for the vague “gist” of information they have stored in making decisions, even when they can recall verbatim (e.g. numeric) information. Unlike similar dual-process theories (e.g. Peters et al. 2006), use of vague gist information is a more advanced form of information processing, rather than inferior to relying on verbatim information (Reyna and Brainerd 2008). Given that the theory argues gist processing is an advanced process, fuzzy trace theory suggests that gist processing develops into adulthood, whereas youth attempt to rely more on numeric information (Reyna and Farley 2006).

Frequency Estimation Introduction

Survey researchers are frequently interested in asking questions about behavioral frequencies, such as how many times someone went to the doctor or how many hours of television has been watched. These questions require that respondents not only access autobiographical memory but also to make decisions on how to process these memories and how to report them. The outcomes of these responses can have impact on important policy decisions (e.g. public education initiatives), marketing strategies, or social research. Understanding how this information is produced and reported is needed to potentially increase accuracy.

More generally, estimating frequencies of any kind, including those for behavior, is partly an issue of memory. How such data is encoded, stored, and this information’s accessibility and availability at later points are important issues in understanding memory broadly. Further, since the data requested is numeric, the tabulation of the memories is also of interest. At issue is whether people recall and count individual episodes of the

event of interest, keep a running tally of events as they occur, or some other process is involved in memory recall. An additional issue is the impact that the requested memory task has on processes and outcomes. This request includes a number of factors, including how the question is asked as well as the frequency to be recalled. In some cases accuracy of the report can be assessed, giving a measure to the strength of such memories.

Strategies for Recall of Frequencies

Until relatively recently, frequency estimation was conceptualized by researchers as the respondent recalling and enumerating individual events of the target (Blair and Burton 1987, Bradburn et al. 1987). Enumeration strategies potentially led to errors of two kinds only – omission and commission (Bradburn et al. 1987). However, later studies examined different strategies that may be used. In one of the earliest works on the topic, Blair and Burton (1987) found a more diverse set of strategies than originally proposed. Although some use of episodic enumeration did occur, the most frequently used strategy was that of rate estimations. People would recall some rate of occurrence (e.g. “I purchase gasoline twice a week”) and then convert it into a frequency for the time frame requested (e.g. “so I purchased gasoline 8 times this month”). They also found a substantial number of uses of what they termed “other processes”. These other processes included combinations of both rates and enumerations and adjustments to rates/enumerations. For example, some people used enumeration within subdomains, such as rate of going out for lunch or dinner in estimating the frequency that one ate at a restaurant.

Others may have used a rate and adjusted using an enumeration. Thus, someone may have enumerated for a shorter time frame, e.g. a week, and then created a rate for a

longer, inclusive, time frame, e.g. a month. The opposite, using a rate and then adjusting with enumeration also occurs. Finally, several people reported using direct estimates of frequencies, i.e. guesses or statements with no knowledge how they arrived at such a number. Interestingly, Blair and Burton (1987) found no evidence that people used an availability heuristic (Tversky and Kahneman 1974), whereby the ease that events could be recalled would be the basis for frequency estimation.

Following this early work, additional research refined the understanding of the strategies used by individuals in frequency estimation. Enumeration of events continued to be found as an often used strategy, as did rate-based estimation (Brown 1995, 2008, Brown and Sinclair 1998, Burton and Blair 1991, Conrad et al. 1993, 1998, Menon 1993, Tourangeau et al. 2000). Episode enumeration generally was used about one-quarter to less than one-half of the time, with other strategies being used the majority of the time (Blair and Burton 1987, Brown and Sinclair 1998, Burton and Blair 1991, Conrad et al. 1993, 1998, Tourangeau et al. 2000). Greater understanding of potential strategies was also gained with these studies, including differences in types and in information used for these.

Conrad et al. (1998) provide a taxonomy of strategies which encompass the majority and most used strategies identified to date.[1] In this study, respondents were asked about ten behavioral frequencies occurring in daily life, and were told to think about it quietly until an answer was decided upon. Then respondents were asked to explain how they came up with the answer, i.e. a retrospective recall. Responses to the

---

[1] See Conrad et al. (1998) Fig 3 p. 361 for a diagram of the taxonomy

survey were tape-recorded. The retrospectively recalled strategies were categorized into a taxonomy developed on the different techniques described, and recalled strategies coded by two groups of two coders.

Based on the findings, a number of strategies were identified. First, strategies are divided along enumeration and non-enumeration lines. Enumeration strategies entail either the recall of individual episodes (episodic enumeration) or the use of a few recalled events to estimate a rate for rate-based frequencies (rate estimation). Enumeration strategies are cognitively more difficult, as individual episodes are recalled. This increased difficulty is evidenced by the greater time required to respond using an enumeration strategy (Brown 1995, Conrad et al. 1998). Consistent with the finding that enumeration strategies were used less often than others, in part due to the increased cognitive burden, there are more different types of non-enumeration strategies in the taxonomy than enumeration strategies.

Within non-enumeration strategies, an individual could use either memory assessment or a form of direct retrieval. Memory assessment is essentially usage of the availability heuristic (Tversky and Kahneman 1974). The ease that events are recalled would be the basis for frequency estimation. Direct retrieval relies on either quantitative or qualitative information. Using quantitative information, a respondent can either retrieve a stored rate or retrieve the stored rate and adjust it based on some additional understandings (e.g. “I go to the store every week, plus twice more this month for extras”). Finally, using qualitative information, what the authors termed a “general impression” could be formed. These general impressions are mostly conversions of information stored as vague quantities into acceptable numeric reports.

These vague quantities are interesting in that they are not numeric and must be converted into a numeric response. Questioning about activities done in the past month, 18% of responses were formed based on strategies relying on vague quantities to infer a specific frequency report (Conrad et al. 1993, 1998). This percentage is also likely low compared to the actual percentage of those relying on vague quantifiers for several reasons. First, the coding strategy used in the study contributed to an undercount. For example, “filling my gas tank several times a week” is coded as rate retrieval in the study, although “several” is a vague quantity.

Second, some respondents may also have inferred what would be an “optimal” strategy according to the researcher, and reported on more quantitative processes to come to a response. While providing a retrospective justification for their response, individuals may have wanted to appear as though they made a high level of effort, rather than just relying on some vague quantity. Even so, 18% is a substantial percent, and was the second largest category coded, indicating that at least some information is naturally stored as vague quantities.

If numeric data is stored as vague quantifiers, as suggested by Conrad et al. (1998), differences found in response distribution to response scales with different values attached to the end points of the scales would be explained. A respondent would map their vague quantity data on to the scale provided. The necessary conversion from the general impression to an acceptable numeric response is aided by the response options. For example, someone who has the information requested stored as “a lot” may select the highest scale point, regardless of its value, which is also consistent with the theory put forth that the middle option being the typical population value (Schwarz and Bienias

1990). Although response scales may be mainly a task related variable (see below), one could conceptualize reliance on the form of the question itself as a strategy, and is evidence that strategies and tasks are closely interrelated.

A recent theory that has been put forth is that people at times use a metastrategy in frequency estimation (Brown 2008). A metastrategy is where a set of strategies is selected prior to the set of tasks and employed throughout the entire set. Therefore, a respondent may use non-enumeration strategies on some questions even when enumeration is possible on them all. Enumeration strategies are more concrete and deliver numeric responses consistent with the task. However, they are more cognitively taxing than non-enumeration strategies, and so a mixed strategy may be possible. People will try to balance accuracy and ease at the aggregate level, rather than considering each individual information request.

Impact of Task on Strategy Selection and Frequency Estimation

There are two classes of task related variables to be considered. The first is related to the content of the task. That is, variables relating to the frequency itself. A second is related to issues of the structure of the request. Generally this structure involves how a question is asked or other manipulations in data collection. Although largely based on memory, frequency requests are also based on other important methodological considerations. Studies examining these facets have used differing methodologies, including laboratory and survey research. Still, the findings provide generally consistent results indicating the impacts these task variables have on strategy selection and frequency estimation.

Content of the Frequency

Given the limitations of memory, it may not be surprising to find that the accessibility of episodic details impact strategy selection. Several studies have found that increases in frequency (number of episodes) reduce the likelihood one uses episodic enumeration (Blair and Burton 1987, Brown 1995, Burton and Blair 1991, Conrad et al. 1993, 1998, Means and Loftus 1991). Brown (1995) conducted an experiment in which respondents were given lists of target words to study. The number of times each target word appeared on the word list was varied, appearing 2, 4, 8, 12 or 16 times, and was a within subjects factor. In addition, each target word was coupled with a "context" word, an event instance for each of the target word. For example, Chicago would be an event instance, i.e. context word, for the target word City. Since the experiment was in part to examine the effect of context memory, two experimental groups of respondents were created. In one, each target word was paired with the same context word at every presentation (e.g. City-Chicago, City-Chicago, etc.). In the other, each presentation of the target word was coupled with a different context word (e.g. City-Chicago, City-Paris, City-London, etc.). The different context word was assigned randomly to each presentation of the target word. The presentation of same or different context words was a between subjects factor. During the frequency test, which asked about how frequently each target word occurred (using an open ended numeric response option), respondents described their process concurrently about how they came to their answer about the frequency. Brown (1995) found that enumeration decreased when the actual frequency increased, and was near non-existent when the context word was the same for each target word. In addition, general impressions, similar to Conrad et al.'s (1998) finding,

increased as the actual frequency increased. When the context word was always the same, Brown (1995) found that most responses among those in the same context condition were made using uninformative strategies, that is, no rationale could be given.

In addition, the reference period (although partly under the control of the researcher) also has been found to impact strategy selection, with longer periods leading to less enumeration strategies (Blair and Burton 1987, Burton and Blair 1991). Although number of events and time frame will be related, Blair and Burton (1987) found independent effects for each. As the number of events moves to the highest levels of frequency within a sample, or the time of an event increases in duration, this data is more likely to be stored as vague quantities. In the Conrad et al. (1998) study, the highest behavioral frequency responses were reported to be inferred from these vague quantities, which were labeled as “general impressions”. Similarly, Brown (1995) found in a laboratory study of word frequencies that higher frequencies led subjects to select general impressions as their strategy of choice. Although not coded in the exact same manner, two studies by Blair and Burton (1987, Burton and Blair 1991) found that the greatest reported frequencies corresponded to strategies they labeled as “other”. This other category included several strategies, including what other studies have labeled as general impressions.

One final noteworthy aspect regarding the frequency itself that affects strategy and estimating is whether the event relates to oneself or to another person (or group). An individual should have more information about events that they are personally involved in, including greater episodic detail. Therefore, when the requested frequency relates to the individual, enumeration will be used more than when requesting information about

some other person (Bickart et al. 1990, Menon et al. 1995, Schwarz and Bienias 1990). Further, these studies show that when answering as a proxy for other people, individuals first estimate the frequency for themselves, and then adjust their responses based on additional knowledge of the other person. The information provided by their personal frequency estimates are also weighted more heavily when making estimates about another who is more similar than more personally distant (Menon et al. 1995).

Content of the Request

The manner in which information is requested frequently affects the cognitive processes that produce a response to this request (Tourangeau et al. 2000). One type of request that is also not immune to the influence of the presentation and content of the request are those for frequency estimates. As noted earlier, closed-ended responses with frequency categories can impact the frequency estimate provided by the respondent (Gaskell et al. 1994, Schwarz et al. 1985).

Theoretically, if respondents gave their actual frequency accurately, then no difference should occur between low and high frequency conditions. The percentage saying “more than two and one half hours” in the low frequency condition should equal those saying any amount more than the smallest option in the high frequency condition. Similarly, the percentage responding “up to two and one half hours” in the high frequency condition should equal that of those choosing any except the greatest option in the low frequency condition. This expected equality is far from the case, however. Although no one selected “more than two and one half hours” in the low frequency condition, nearly thirty-percent chose some value greater than this in the high frequency condition. Clearly, the high frequency condition led people to state higher frequency

estimates, and the low frequency condition led to lower frequency estimates. Interestingly, the open ended responses led to a distribution that fell between these two.

If frequencies are recalled using vague quantifiers or other general impressions, the recollection as vague quantities also helps explain differences in numeric response scales with different values attached to the points of the scales. A respondent would map their vague quantity data on to the scale provided. If people do select scale points based on this stored vague quantity, it is also consistent with the theory put forth that the respondents infer the average population frequency (Schwarz et al. 1985, Schwarz and Bienias 1990). Using the middle response option as an anchor of the average, people can then say they are comparatively more or less than average, consistent with understandings of how people use vague quantifiers (Schaeffer 1991).

Measurement of Frequencies

The measurement of numeric responses has been generally done by asking for some numerical answer, in one of two ways. First, the request is made by a question with an open-ended response in which a numeric answer is provided (e.g. Blair and Burton 1989, Brown 1995, 1997, 2008, Burton and Blair 1991 Conrad et al. 1997, Hasher and Zacks 1984). The second way has been to ask for frequencies using a scale labeled with numeric quantities, generally involving ranges (e.g. Knauper et al. 2004, Menon et al. 1995, Schwarz et al. 1985, Schwarz and Bienas 1990). However, it has been noted there are difficulties with using a numeric scale, with suggestions for avoidance in favor of using numeric open ended responses (Schwarz et al. 1985). In addition, some have instead used vague quantifiers, which place verbal labels onto scale points meant to represent some number (Bradburn and Miles 1979, Lu et al. 2008, Schaeffer 1991,

Wanke 2002). The general lack of use of vague quantifiers may be due in part to the complexities of their use and the potential problems associated with vague quantifier response scales, as identified in research on the topic.

For example, vague quantifiers have linguistic meaning beyond simply the numeric quantity these are conveying. First, as the name suggests, the quantity expressed is done so in a vague manner, which may lead to a range of numeric values that is possibly expressed by the verbal label (Budescu and Wallsten 1985). Indeed, a number of studies examining the numeric translation of vague quantifiers find large between-subject variation (Beyth-Marom 1982, Biehl and Halpern-Felsher 2001, Budescu and Wallsten 1985, Clarke et al. 1992, Lichtenstein and Newman 1967). Biehl and Halpern-Felsher(2001) find that this variability tends to be greater among younger respondents compared to older respondents. Although many studies suggest strong internal consistency of vague quantifier use within individuals (e.g. Budescu and Wallsten 1985, Lichtenstein and Newman 1967), one study suggests that there is also possible within-subject variation, with people giving inconsistent numeric translations of vague quantities (Clarke et al 1992). However, it should be noted that many of these translations were context independent, and that vague quantifiers are naturally used in context (Windschitl and Wells 1996). Still, at least in one case, when the vague quantifiers are placed in context, there is greater between subject variation (Beyth-Marom 1982). It should be noted however, that in this instance, the context was asking about subjective probabilities rather than behavioral frequencies.

Further, quantifiers focus attention on the reference set or complement set of the quantity (e.g. the subject vs. other) (Sanford et al. 1994, 1996, Teigen and Brun 2003).

The reference set is focused on by positive quantifiers (e.g. a few, some), while the complement set is focused by negative quantifiers (e.g. few, not many) (Sanford et al. 1996). This attention focusing of vague quantifiers occurs in both understanding and production of sentences. Similarly, optimism for an outcome is greater when the vague quantifier used is positive and there is greater surprise when the outcome occurred (did not occur) when a negative (positive) quantifier is used to described the chance of outcome, even when numeric translations of the quantifier are the same (Teigen and Brun 2003).

Studies examining the use of vague quantifiers in survey questions regarding behavioral frequencies have also identified several potential issues. One is the differential use of the vague quantifier scale, i.e. different interpretation of meanings (Bradburn and Miles 1979). This different interpretation of meanings is consistent with the large between-subject variation that has been noted previously. The Bradburn and Miles (1979) study first asked respondents about the frequency that they felt excited and bored using a vague quantifier scale, and then later asked how many times the respondent meant by the vague quantifier they responded with. For example, if the respondents said they were excited about things “very often”, they were asked how many times “very often” meant. This study also found that the differences in numeric translation between scale points are not equidistant, as would be expected (e.g. very often vs. not very often were not equidistant). Recent research, however, has shown that differences between subjects (students) in numeric translation are relatively small (Nelson-Laird et al. 2008).

Reanalyzing this same data as Bradburn and Miles (1979), Schaeffer (1991) argues that use of vague quantifiers is based on not only actual frequency but the relative

frequency of similar people, such as those with similar demographics. That is, when responding that a behavior occurred “often”, this response is based not only on whether the event occurred often, but also if it did so compared to what is expected of similar individuals. For example, more educated respondents gave higher value meanings for “pretty often” and “very often” than did those with less education (Schaeffer 1991). Additional studies further show the importance of referent groups (Wanke 2002).In addition, the type of behavior asked about is also compared to similar behaviors, such as to other leisure or cultural activities (Wanke 2002). This comparison may be in large part due to the respondent following conversational norms and being a cooperative communicator, altering responses based on information provided by the survey (Wanke 2002).

An additional finding suggests that the context of the vague quantifier matters; what “often” means for one behavior is different than another behavior (Bradburn and Miles 1979, Nelson-Laird et al. 2008). For example, Bradburn and Miles (1979) find that being bored receives a lower numeric meaning of “not too often” compared to that of being excited. It should be pointed out that the Bradburn and Miles (1979) data, also used by Schaeffer (1991), asks about emotional states, which may be different in terms of recall strategies (Brown et al. 2007). As such, this differential use of recall strategies may also affect the numeric translations of the vague quantifiers. However, studies examining vague quantifier responses to questions about other activities produce findings consistent with these (e.g. Nelson-Laird et al. 2008, Wanke 2002), suggesting that asking questions using vague quantifiers may be found in both emotional state and activity frequency settings.

Based on such results, it is often suggested to avoid vague quantifiers when possible (Beyth-Marom 1982, Schaeffer 1991, Tourangeau et al. 2000). Tourangeau et al. (2000) suggest that the only meaning that may be derived from vague quantifiers, given these issues, is the ordinal position of the response, i.e. sometimes is more than never and less than always. However, the potential difficulty in answering numerically is also recognized. Numeric data are affected by ease the events are remembered, how often people think of it, how well instances are imagined, the spacing of events, recall techniques used, the feelings at time of occurrence, response scale used, and true frequency (Schaeffer 1991). Bradburn and Miles (1979) suggest that it may not always be possible to ask for exact estimates of many behaviors and subjective states due to the cognitive burden. Indeed, in the Bradburn and Miles (1979) study, interviewers said that many respondents had difficulty translating vague quantifiers into numeric values. Further, numeric information may be more difficult to think about and recall, and possibly stored inaccurately, although vague quantifiers may entail more error due to its comparative nature (Schaeffer 1991). In a study using behavior coding, numeric responses produced significantly more problems than did vague quantifiers (Johnson et al. 2006). It is also possible that people use vague notions plus some set of heuristics to compute numeric data, which is consistent with more missing data for numeric frequencies (Schaeffer 1991). In addition, studies on subjective probabilities suggest that like vague quantifiers, numeric data may be affected by the context of the question, contrary to general understandings (Windschitl and Weber 1999).

In a recent dissertation, Marincic (2011) examined vague quantifiers and numeric open ended responses (translations) using the National Survey of Student Engagement

(NSSE) (see Study 2). She examined how vague quantifiers were quantified and whether there were differences in interpretation; whether any individual differences in interpretation was due to individual characteristics; and whether latent models used for numeric data also fit for those when vague quantifiers are used. The author used a number of multilevel and latent variable models to assess these questions. Using a reduced set of the items available, significant differences in regards to interpretation were found at the individual level, and some evidence of differences in items (when items were analyzed simultaneously). Additionally, the analysis found that unit of time for the vague quantifier (e.g. daily, weekly, etc.) significantly influenced interpretation.

In terms of what caused the differences in interpretation, the largest contributing factor was that of respondent level engagement in a behavior, with more engaged respondents giving larger estimates for higher frequency vague quantifiers than less engaged respondents. Other variable had an effect, such gender, race, and school size, but all of these led to much smaller differences. Further, using numeric responses as the "gold standard" model, Marincic (2011) found some support (depending on the scenario used) that latent factors extracted from the different measures were different models. Finally, the author applied a number of factor mixture models, with the hypothesis that that model would best reflect the numeric data. However, limitations in the data did not allow for full examination of this hypothesis.

Accuracy

Of course, much of the research on frequency estimation is conducted by behavioral researchers who are interested in the accuracy of frequency estimates, especially given the implications noted in the introduction. Assessing the accuracy of

such frequency estimates, however, is quite difficult. There must be some knowledge of what the actual frequency is, which limits studies of accuracy to laboratory settings, where the researcher controls the actual frequency, or in instances where other sources are available to obtain actual frequency.

In terms of strategy selection, in experiments examining frequency estimates of words presented, individuals tended to overestimate frequencies when using a rate-based strategy and underestimate when enumerating (Brown 1995). Absolute error tended to be smaller in this study for those using enumeration. This difference was more pronounced as actual frequency increased. The finding that enumeration led to smaller error is consistent with findings of a study regarding frequency using bank services, which was confirmed through bank records (Burton and Blair 1991). However, several other studies have found that rate-based strategies have led to more accurate frequency estimates (Menon 1993, 1997).

These differences can be explained by the type of frequency requested. When the event is regular and similar, then accuracy is increased by using rate-based strategies (Menon 1993, 1997). When the event is irregular and dissimilar, then enumeration will likely lead to greater accuracy compared to rate-based strategies. Still, the errors are greater than when rate-based strategies are used for regular and similar behaviors. When events are regular or similar (but not both), respondents at times try to enumerate, leading to increased errors.

How the question is asked may also affect accuracy. Asking about the behavior during the typical week led to higher predictive validity for theoretically related constructs than did asking about behaviors during the past week (Chang and Krosnick

2003). Recent research indicates that eliciting vague quantities or general impressions may be at least as accurate as asking for numeric scale responses (Lu et al. 2008). The authors obtained data on medication adherence through an electronic system that monitored the number of times the medication bottle was opened, using this opening as an indicator that the medication was taken. They then asked those using this system about their medication adherence over three days, seven days, and one month. For the one month time frame, they asked respondents about adherence using three different scales, a six-point vague quantifier scale (from “none of the time” to “all of the time”, an eleven-point percent scale (0, 10, 20 …100), and a six-point scale rating adherence (from “very poor” to “excellent”). The authors then assigned values to the six-point rating and vague quantifier scales based on a percentage scale, i.e. 0, 20, 40, 60, 80, and 100.

They found that self-reported adherence to HIV medication treatment conformed to actual adherence when general impressions were obtained on a vague quantifier and a rating scale. The vague quantifier scale performed as well compared to recorded data as did the numeric percentage scale response, with the rating scale performing better than the numeric scale. These findings conform to the finding that larger frequencies are estimated from general impressions (Conrad et al. 1998). The better performance of the rating scale and similar performance of the vague quantifier scale with the numeric response options may also be due to the respondents’ limited numeric capabilities (Lu et al. 2008). Further, this study examined the predictive validity of the different response scales, by correlating the responses to the observed HIV RNA in the respondent. The results show that there are no significant differences between response scales in terms of predictive validity.

Psychology of Subjective Probabilities and Risk Perception Introduction

People are confronted with subjective probabilities on a regular basis. These probabilities are most often conceptualized as the risk an event will (or will not) occur (Tversky and Kahneman 1981, Tourangeau et al. 2000). How people perceive and understand this risk has been studied in various ways, with varying conclusions. Studies have concluded that experts in the topical areas understand risk differently than lay people (Slovic 1987). Unlike experts of risk, who base risk levels on actual probabilities of occurrence or annual fatalities (when dealing with health risks), lay people base their risk perceptions not only on actual risk but also on the catastrophic potential (i.e. the possible outcomes) and the future threat (Slovic 1987, 1998). Further, some studies suggest that people are accurate about some risks while quite wrong about others (Dominitz and Manski 1997, Benjamin et al. 1997). Importantly, several theories have developed on the ways people perceive risks and make decisions about these risks.

Early in the study of risks it was noted that people did not always make rational or error free decisions regarding risks (Tversky and Kahneman 1974). Three heuristics were identified that made complex numeric problems, such as those dealing with risks, easier but that often led to these irrational and erroneous decisions. These heuristics are that of representativeness, in which probabilities are evaluated by the degree which the recalled information is representative of the target outcome; availability, in which judgments are made on the ease of recall; and anchor and adjust, by which an initial value (which may come from suggestion or recall) is anchored upon and adjusted this initial value for judgment (Tversky and Kahneman 1974). Kahneman and Tversky later also posited that decisions diverge from expected utility due to both the weighting of prospects and frames

(whether decisions framed as gains or losses, or status quo) (Kahneman and Tversky 1979, Tversky and Kahneman 1981). Called Prospect Theory, the theory states that there are decision weights which are combined with values (utilities) to judge between decisions.

As noted by this research, at some level, people rely on their intuitions when making decisions. Later theories of processing risk also reflect the notion that people rely on intuition, with a number of theories positing dual processes to understand and make decisions on risk (e.g. Peters et al. 2006, Reyna 2004, Reyna and Farley 2006, Reyna and Brainerd 2008, Windschitl and Weber 1999). Although these theories differ slightly, all suggest that one of the processes relies more on numeric, rational and rule based thinking, while the other relies on affective, experiential, associative and intuitive based thinking. Much research suggests people rely on intuitive decision processes in regards to risk (Reyna and Brainerd 2008, Slovic 1987, Yamagishi 1997). Even so, some of this research suggests that flawed and incorrect decisions are made when the relying on the affective-experiential (intuitive) based process rather than the rational and rule based process (e.g. Peters et al. 2006). Still, it is also noted that the individual, in order to make sound decisions on numeric information, must have accurate information, then make calculations and inferences, remember the information, and finally be able to weigh factors to make a decision (Peters et al. 2007).

Rule based reasoning, however, depends on retrieval cues that elicit principals stored in long term memory. Fuzzy trace theory also suggests that in addition to verbatim and gist based memories (encoded and processed in parallel), people retrieve what they know about ratios when the knowledge is cued, but applying that knowledge is interfered

with when classes of objects/events overlap or are nested in one another (Reyna and Brainerd 2008). As a result of this confusion, people often focus on salient gist, i.e. comparisons between numerators at expense of denominators. Denominator neglect explains overestimation of some risks and underestimation of others, because people neglect the base rates, focusing only on numerator (Reyna 2004). Generally, intuition has important place in fuzzy trace theory, and is considered an advanced form of reasoning because of developmental evidence. As such, adults, who are more likely to rely on gist based information, are better prepared to make decisions about risks compared to adolescents, who rely more on verbatim information (Reyna and Farley 2006).

Another, although not necessarily contradictory process, that has been suggested is that people process risk through a Bayesian updating process (Viscusi 1990, 1992, 2002). As such, people have preconceived stored beliefs about a perceived risk. New information about the risk is often encountered, such as through personal experience or through coverage in the media or informational environment. This new information is used to update the individuals' risk beliefs, weighting the new information and stored beliefs according to the strength of each. Thus, beliefs that have long been held and have been developed on a large number of information points are less likely to be shifted much in the face of new information. In comparison, newly formed beliefs based on a small number of information points are more likely to be altered by new information obtained.

These different theories suggest potential ways to measure risk, either through numeric or vague quantifier measures. Indeed, some of the theorists have argued for use of one measure over the other. Specifically, Viscusi (2002) argues for the use of numeric measures instead of vague quantifier measures. However, if risk is represented in gist

manner, then risk may be better measured using vague quantifier scales. A number of studies have examined risk numerically and through vague quantifier measures, and are described in more detail in the following sections.

Numeric Estimates of Risk Perception

Much of the work on numeric estimates of subjective probabilities, i.e. risk perceptions, has been conducted on health issues, and in particular the risks of smoking. Due to the focus on health issues, and the fact that the data to be used in the current research focuses on smoking risk beliefs, much of the focus of this section will focus on smoking risk perceptions. However, it is important to note several other studies that have used numeric estimation of risk perceptions. Among the earliest was the seminal work by Lichtenstein and colleagues. Importantly, the authors find evidence that estimation of subjective probabilities relies on somewhat different processes than frequency estimation (Lichtenstein et al. 1978). Also, they found that respondents tended to overestimate small risks and underestimate large risks, a finding that has been found similarly in other studies (Slovic 1987). These findings suggest the possible use of the availability heuristic for subjective risks (e.g. direct experience, news), that people avoid extremity in response, or the anchoring and adjusting of answers (e.g. on low/high anchors). These biases possibly are due to a number of sources, including disproportionate media exposure, memorability, or imaginability (Lichtenstein et al. 1978).

Much of the early research on smoking focused on people's knowledge and beliefs about smoking as harmful to health rather than subjective estimation of the risks (e.g. Salber et al. 1963, Kelson et al. 1975). National polls focused on whether people were aware of and believed smoking was harmful rather than asking the amount of risk

perceived. For example, in a 1949 Gallup poll, 60% said that they believed that smoking was harmful to health (questions used in this section can be found in Appendix 1). In 1954, Gallup asked questions in January and June national surveys about the awareness of the link between cancer and smoking. In the January and June surveys, 83% and 90%, respectively, said they had recently heard or read smoking had been linked to lung cancer. Beliefs about smoking causing cancer in those surveys were significantly lower. In January, 41% said they believed smoking causes lung cancer, with 42% agreeing to the same in June. This number grew steadily, with 83% expressing the belief that smoking causes lung cancer in a 1981 Gallup survey.

The first published work including data on numeric subjective probability estimates appears in Lee (1989).[2] Australian respondents gave the chances of an average smoker developing lung cancer (along with other questions about other diseases) on an 11-point scale, with each point numbered between 0 and 100 in intervals of 10. Respondents were also asked about their own chances using the same scale. Although the purpose of the paper is not to present subjective probability estimates, it is clear that both smokers and non-smokers gave high chances for lung cancer, with smokers giving an average response of about 50 and non-smokers a response of about 60. Chances for personal risk for both smokers and non-smokers were significantly less than population risks. However, smokers' perceived personal chances were higher than that of non-smokers indicating smokers understand they have a higher risk (Lee 1989).

---

[2] Kristiansen et al. (1983) implement magnitude estimation using line drawing, but is not considered here to be the same as asking directly for an X out of 100 (or other base) subjective probability estimate.

The shift of focus to numeric risk estimates of smoking was mainly precipitated by the work of Viscusi (1990). Data collected in 1985 through a survey funded by the tobacco industry allowed individuals' numeric estimates of subjective risk from smoking to be analyzable. The target question asked, "Among 100 smokers, how many of them do you think will get lung cancer because they smoke?" Variations of this question have been asked repeatedly over time and are now the standard measures in numeric subjective risk estimates for smoking. The mean response to this question in 1985 was 42.6, significantly greater than the expected lung cancer rate among smokers of 6 - 13 out of 100.[3] This finding has been consistently found in other research (Krosnick 2001; Romer and Jamieson 2001; Viscusi 1992, 2002; Viscusi and Hakes 2008). Table 1 presents the mean number of smokers expected to get lung cancer based on similar absolute risk questions over time, denoting the source of the data. The youth and adult 1999 data are those employed in this study, and will be discussed further.

[3] The estimate of 6 – 13 out of 100 comes from Viscusi (2002), derived from data contained in scientific reports in various Surgeon General Reports. Earlier estimates of this number were 5-10 out of 100 (Viscusi 1991).

Table 1

Mean Absolute Risk Estimates of Lung Cancer Risk for Smokers

Year	Source	Mean
1964	Industry	16.4*
1977	Industry	45.6
1980	Industry	26.3*
1985	Viscusi 1990	42.6
1991	Viscusi 1992	38.0
1995[4]	Sutton 1998	19.0*
1997	Viscusi & Hakes 2008	47.2
1998	Viscusi & Hakes 2008	47.6
1999	Annenberg 2 Youth	60.4
1999	Annenberg 2 Adult	48.5
2000	Krosnick 2001	43.4

*denotes data collected via face-to-face personal interview

Although these data first appeared in published accounts beginning with Viscusi (1990), similar data has been of interest to the tobacco industry for some time. The tobacco industry regularly hired polling companies such as Gallup and Roper to collect data on various topics of interest. These data were found in the various online tobacco document databases. A question on subjective smoking risk measured numerically was first asked in 1964, shortly after the seminal first United States' Surgeon General's report on smoking and health. Although the wording was slightly different from the standard measures used in later research, the item asked about the same construct requesting a

[4] This study was conducted in the United Kingdom.

numeric subjective probability. Unlike the later published data, the mean was significantly lower, estimated at 16.4 out of 100 smokers.[5]

Later industry-funded surveys switched to wording consistent with the above standard absolute risk question. The next instance not in the published literature came from data collected in 1977. The mean for this year was more in line with later findings at 45.6 out of 100. In 1980, the next time the question was asked, the mean was 26.3. This decrease is an unexpected shift downward in the mean, especially given that the next time the question was asked, in 1985, the mean was the 42.6 reported by Viscusi (1990). There is no reason to expect that risk perceptions should have fluctuated in such a manner; the only major change from one survey to another was the data collection methodology. Both the 1977 and 1985 surveys were collected by telephone surveys, while the 1980 survey was collected using face-to-face personal interviews. Strikingly, the lowest means in Table 1 come from data collected using face-to-face personal interview surveys. Although other differences exist among the surveys, the only difference consistent for all three (Roper 1964, 1977, Sutton 1998) was that they were face-to-face surveys. In spite of other differences, such as denominator (e.g. Krosnick (2001) uses a denominator of 1000 smokers instead of the usual 100), the means for all the telephone surveys are similarly high from 1977 through the present day.

As alternatives to the standard absolute risk measures, other numeric measures have been suggested to better capture risk perception. These arise from the possibility that numeric estimates to the standard absolute risk question are feasible measures, but

---

[5] The mean was not presented directly in the report, but was estimated using the midpoint of the ranges for the frequency estimates weighted by the proportion of respondents in that range.

that individuals perceive risk in a more comparable rather than absolute manner. Specifically, it is not the risk of a smoker alone that matters; rather it is the risk of smoking compared to not smoking. Krosnick (2001) explores this idea and suggests two additional measures: the risk difference and the relative risk. The risk difference is the difference between the absolute risk of the smoker and the absolute risk of the non-smoker. That is, the risk to smokers above and beyond the risk of lung cancer to non-smokers. Relative risk compares the magnitude of the difference between these two absolute risk measures as the ratio of the two estimates. Comparing the measure of absolute risk to smokers to these two new measures, Krosnick (2001) finds that the relative risk measure predicts smoking status best, while the absolute risk to smokers predicts the worst. Further, although the mean of absolute risk to smokers, like prior studies, indicates an overestimate of risk, the measure for relative risk indicates people underestimate the risk of smoking.

Vague Quantifier Measures of Risk Perceptions

Another possibility is that numeric responses used in any manner are suboptimal in measuring risk perceptions. An alternative way to measure risk perceptions is by a more qualitative scale, providing vaguely quantified risk options, similar to a study on smoking risk perceptions conducted by Slovic (1998). Numeric responses require that individuals understand and estimate numeric responses properly given the scale. Although some have argued that people can use numbers to make subjective probabilities (Krosnick 2001), this assumption seems to be tenuous. First, numeric literacy among the general public is quite low (Peters et al. 2007). Studies have also shown that a number of respondents do not answer with logically consistent responses to subjective probability

measures using numeric scales, particularly among those with lower cognitive abilities (Belli et al. 1999). Further, the use of numeric scales led to a number of respondents satisficing, again particularly among those with low cognitive abilities. People have also been found to incorrectly select greater magnitude probabilities when frequency and denominators differed (e.g. 24.14 out of 100 versus 2414 out of 10000) by choosing cases where the higher total frequency was greater, neglecting the denominator (Yamagishi 1997).

Other findings in the survey literature concern the impact of numeric labels placed on scales (e.g. varying levels of number of hours of television watched at the end points of the scale). Studies have found that using different values produces differing estimates of frequency or duration (Schwarz et al. 1985; Wright et al. 1994). This result has also been found in measures of smoking risk perceptions (Borland 1997). Differing scale anchors and response options led to differing percentages of the population over- and underestimating the risk of smoking, contrary to expectations. Assuming respondents are cognitively able to assess risk numerically, one should expect similar responses regardless of the scale values. Borland (1997) suggests that in the case of the standard absolute risk measure used for smoking, some respondents may perceive the response scale as a danger analogue scale. That is, given inability to process the task as constructed, the interpretation is for a scale indicating levels of danger, possibly 0 for “no danger” to 100 for “great danger”.

However, Viscusi (2002) argues that vague quantifiers are not suited for measuring smoking risk perceptions because these cannot be compared to an objective estimate, and there is not a “correct answer”. It is not possible to know what “a little bit

risky" means in objective terms and thus not ascertainable if smoking risk is over- or underestimated. Still, there are number of reasons why the survey designer should not rule out the use of vague quantifiers completely, as suggested by many practitioners, especially for measurement of subjective probabilities. The most important and overriding factor is whether the respondents naturally think about and encode numeric information as vague quantifiers. Respondents must map their answers to survey questions onto the response options given (Tourangeau et al. 2000). Survey designers must provide response options that best represent the respondent's answer; not doing so produces error-prone data, even if respondents did provide an "acceptable" answer.

Studies suggest that subjective probabilities are internally represented as verbal information (Zimmer 1984). Supporting this view, several studies have found most people prefer to convey probability information as vague quantities (Olson and Budescu 1997, Wallsten et al. 1993). In a study of students, 65% preferred to convey uncertainty information through vague quantities (Wallsten et al. 1993). This number is likely low compared to the general public as half of the sample studied were MBA students who took the survey in a statistics class. Similarly, American college students rated scales with vague quantifiers as easier to use and more representative of their feelings towards risk assessment than numeric-based scales (Diefenbach et al. 1993). These studies have also found that subjective probabilities measured by vague quantifiers predicted respondents' behaviors and preferences at least as well, and frequently better than, numeric scales (Diefenbach et al. 1993, Windschitl and Wells 1996; Weinstein and Diefenbach 1997). These findings may be stronger in the general population, as it seems

reasonable to assume that a college student may be more proficient and comfortable with numeric scales.

The increased validity vague quantifier scales (e.g. Windschitl and Wells 1996) exhibit in these studies is likely due in part to the processes involved. Numeric responses to subjective probability estimates elicit more rule-based thinking, whereas vague quantifier scales result in more associative processing (Windschitl and Wells 1996, Windschitl and Weber 1999). Many human behaviors are not dictated by rule-based processes but are related to associative processes, resulting in the better prediction of behavioral outcomes. Vague quantifiers may therefore also be better measures of risk perceptions, particularly when relationships between behaviors and other variables are of interest. An additional benefit is that since vague quantifiers are simpler to use and map responses to, the cognitive effort required is reduced. Although not possible to test here, this reduction in cognitive effort needed indicates that responses may be less affected by mode of data collection, which apparently influences numeric estimates.

Summary and Gaps in the Literature

Based on the above literature review, a number of factors are important in the measurement properties of questions about quantitative information, as well as a number of important gaps in the literature that should be filled to better understand these factors. Of most importance for this dissertation is the response format used. As the literature shows, quantitative information may naturally be stored as vague quantities, for both frequency information and subjective probabilities (Conrad et al. 1998, Reyna and Brainerd 2008, Windschitl and Wells 1996). However, some of the literature in survey

research and risk perceptions suggests that vague quantifiers are to be avoided (Tourangeau et al. 2000, Schaeffer 1991, Viscusi 2002).

Not recognizing that people may more naturally think about quantitative information in vague terms, it has been suggested that recommendations to avoid vague quantifiers may in part be due to researcher bias (Windschitl and Wells 1996). In addition, there have been relatively few studies that have been published on vague quantifiers in survey research, and fewer yet that have examined the measurement error properties of these measures. Findings do show vague quantifiers mean different things to different respondents (Bradburn and Miles 1979, Schaeffer 1991). However, none of these studies examine the accuracy of vague quantifier responses. This lack of accuracy measures is due in part to the lack of verifiable information available in most survey situations.

One survey that had such verifiable information was Lu et al. (2008), which found that vague quantifiers performed as well as numeric response scales. However, these researchers set the questions up such that the response was about the percentage of time completing an activity, and numeric translations were based on percentages, rather than the more standard way to ask about frequency in surveys, i.e. the number of times an activity was conducted (Tourangeau et al. 2000). That is, the one study that has examined the accuracy of vague quantifiers did so for a question format that is not commonly used in survey research in the collection of behavioral frequency data. Further, the numeric response format was an eleven point scale, rather than the open ended response formats often used for frequency estimation. Overall, this research leaves a

significant gap in the literature as to which, if any, response format is more accurate for standard frequency-type questions.

For similar reasons, there is a gap in the literature examining the predictive validity of vague quantifier and numeric response options for frequency questions. Predictive validity has been shown to be important in understanding measurement properties for other question formats for frequency questions (Chang and Krosnick 2003). For response options for questions about frequency information, the only study that examined this issue is Lu et al. (2008), who found no differences between response scales in terms of their predictive validity. Again, the results of Lu et al. are based on percentage-completed scales, with transformations of the vague quantifier scale based on a percentage, rather than on the frequency an event occurred or numeric open ended responses. Additionally, for both the accuracy and predictive validity of response scales, the results of Lu et al. (2008) are the only available evidence, and thus, even absent the problems with using a percentage scale, more studies are need to provide converging evidence about the comparative measurement properties of the response formats.

For subjective probabilities, accuracy is not necessarily appropriate for assessing measurement properties, given the subjective nature. However, some researchers have examined subjective probabilities compared to actual risk estimates to ascertain over or underestimation of risks (e.g. Viscusi 1990, 2002, Krosnick 2002, Benjamin et al. 1997). Instead, in order to assess the measurement properties of subjective probabilities using different response formats, predictive validity has been used between responses and theoretically related variables (Krosnick 2002, Windschitl and Wells 1996, Diefenbach et al. 1993). However, several gaps exist in these studies. First, these studies have not

examined all of the numeric responses and vague quantifier scales simultaneously. Krosnick (2002) examined absolute risk, relative risk, and risk difference measures, but not vague quantifier measures. Other studies (e.g. Windschitl and Wells 1996, Diefenbach et al. 1993) included vague quantifier and absolute risk responses, but not relative risk or risk difference measures. Second, all of these studies except Krosnick (2002) were conducted using small, non-representative samples, and Krosnick did not include the vague quantifier response option for comparison. Taken together, this research does not allow for important comparisons in respondent characteristics that possibly differentially affect the measurement properties of the different response formats.

Indeed, as the above literature review suggests, there are possible respondent characteristics that may affect the accuracy and predictive validity for different frequency and subjective probability response formats. First are the numeric capabilities (numeracy) of respondents. As noted, many people in the population have low levels of numeracy (Peters et al. 2007, Galesic and Garcia-Retamero 2010). Numeric ability has also been found to affect understanding of new quantitative information (Schwartz et al. 1997) and the ability to make decisions regarding quantitative information (Peters et al. 2006). However, there has been no examination of the effect of numeracy on the measurement error of quantitative information requested by questions like those in surveys, such as frequency and subjective probability estimation. This measurement error includes both the accuracy and predictive validity discussed above.

If respondents lack basic numeracy it follows that the ability to carry out the steps needed to answer a question about quantitative information is also likely lacking. Most

pertinent, they may be less likely to encode quantitative information as actual numbers, and/or unable to recall numeric information sufficiently, in part because the information is stored as different format, such as vague quantities. Indeed, it is suggested that many people store numeric information in a vague, “fuzzy” way rather than verbatim numeric information (Reyna and Brainerd 1991, Reyna and Brainerd 2008). Therefore, it may be that for many respondents, especially those with low numeracy, asking questions with vague quantifier response options are more in line with the natural way they think about quantitative information. As such, providing vague quantifier response options may improve the measurement properties of questions about quantitative information, especially among those with the lowest numeracy. Indeed, Lu et al. (2008) suggest, but do not test, that the similar results between numeric and vague scales, with rating scales more accurate than either of vague or numeric scales, may be due to the low numeracy of their sample. However, there is a significant gap in the literature examining the relationship between numeracy and the response formats. As of this dissertation, no such research has been found examining these two factors.

Similarly, as the literature has shown, at least in the frequency domain, actual frequency is related to accuracy, in that lower actual frequency tends to be reported more accurately than when the actual frequency is greater (Brown 1995, 1997). However, in Brown’s (1995, 1997) research all responses were collected via open-ended numeric responses. Further, it has been suggested that as actual frequency increases, it is more likely that the quantitative information is stored as vague quantities (Conrad et al. 1998). It follows from this finding, then, that as actual frequency increases, if these are more likely stored as vague quantities, the accuracy may be increased by asking for responses

using vague quantifier responses. No research that has been identified has examined the relationship between actual frequency, response formats, and accuracy.

In addition, the numeracy of the individual may also affect the relationship between actual frequency, response formats, and accuracy. The threshold of when quantitative information becomes stored naturally as vague quantifiers may differ based on numeric ability, such that those with lower numeracy are more likely to store quantitative information as vague quantifiers at lowers levels of actual frequency. Thus, for those with lower numeric ability, it may be that vague quantifier response options provide greater accuracy even at lower levels of actual frequency, while for those with higher numeracy, the improvement in accuracy with vague quantifier response formats are only at higher levels of actual frequency. At lower levels of actual frequency, those with higher numeracy may provide more accurate answers with numeric response formats. As of now, however, no study has examined these possible interactions.

An additional possible interaction with the response format provided (the main concern of the study) is the strength of the context memory for the target of the frequency estimation question. That is, how varied instances of the target event are, whether the target events are all the same or are different and unique in some way. Brown (1995) has shown the importance of different context memory in terms of accuracy and recall strategy, with more varied context memory leading to greater accuracy and more use of enumeration strategies. This increased accuracy is largely due to the increased ability to discriminate target instances with increased variation in context memory. It may then be that decreased variability in context memory leads to storage in more vague and “fuzzy” ways as suggested by Reyna and Brainerd (1991). If this vague storage is the case, then

there is an interaction between response format and context memory, where increased context memory (more variable) lends itself to greater accuracy with numeric response formats, while decreased context variability leads to greater accuracy when asked using vague quantifier response formats. As of now, there is no study found that examines this possible interaction.

As noted, numeracy also likely affects subjective probability estimation in ways similar to that of frequency estimation relating to response formats. That is, people with lower numeracy ability may be less likely to encode quantitative information as numbers, or they are unable to recall numbers correctly, in part because the information is stored as different format, such as vague quantifiers. However, as with frequency estimation, there has been no research on the effect of numeracy on the predictive validity of different response formats, either numeric open ended responses or vague quantifiers. This lack of research on numeracy, especially in regards to response formats, is considered another significant gap in the literature.

Beyond numeracy, there are also other possible factors that affect subjective probability estimation in regards to response formats, which have not yet been examined in the literature in a systematic fashion. Specifically, the potential impact of age has not been researched in terms of the measurement properties of different response formats in regards to subjective probabilities. The need for the research on age difference is important because it has been suggested that youth perceive risks in a different way than adults (Reyna and Farley 2006). Specifically, youth may perceive risks more through numeric, objective information when deciding on whether to partake in a risky behavior. Conversely, adults rely more on vague, “fuzzy” notions of risks in making risk behavior

decisions. However, these potential differences have not been studied empirically, and are a source of further research that will be examined, as discussed below.

In addition to accuracy and predictive validity, another potentially important source for the measurement properties of the different response formats is the logical consistency between numeric and vague quantifier responses. As the above literature review notes, several studies have examined the consistency of numeric translation of vague quantifiers. A number of studies have shown that there is large between-subject variation in the meaning of vague quantifiers, as defined by numeric translations (Beyth-Marom 1982, Biehl and Halpern-Felsher 2001, Budescu and Wallsten 1985, Clarke et al. 1992, Lichtenstein and Newman 1967). Age has also been shown to be an important factor in this variability (Biehl and Halpern-Felsher 2001). Within-subject variation may also exist (Clarke et al. 1992). Further, numeric meanings of vague quantifiers also seem to be partly defined by group comparisons (Schaeffer 1991).

Although logical consistency of response formats has been studied more than other sources of measurement properties of response formats (accuracy, predictive validity), there are some aspects that have not been completely examined. First, is the importance of whether vague quantifier translations are logical, for example, that the response “never” is equal to zero. Further, it should also follow that respondents, for example, should always give a translation for the quantifier “very risky” that is significantly greater than “somewhat risky”. However, this logical ordering has not been studied in an adequate manner.

Similarly, the effect of numeracy has not been examined in regards to the logical consistency between numeric and vague quantifier response options. Given the effect of

numeracy on other areas of cognition, it follows that this logical consistency may be affected as well. For example, Schwartz et al. (1997) found that more numerate individuals were able to understand changes in risk information, while Peters et al (2006) found numeracy generally improved decision making. It may be that these are due to the ability of more numerate individuals to manipulate as well as understand numbers. Given this increased ability, it follows that more numerate individuals are more able to understand the logical relation between numeric and vague quantifier response options, and are better able to have logically consistent translation between the two. Although the non-numerate may be able to think in vague ways, it may be numerate individuals are better at the translation of vague quantifiers into numeric terms. However, as of this dissertation, no research has examined the effect of numeracy on logical consistency of translation between these response formats.

This dissertation presents several studies to examine the important factors and gaps, based on the above literature review and the identified gaps in the literature provided in this summary. Most important, and of central concern, is the measurement properties of the different response formats of numeric and vague quantifier response options for both frequency and subjective probability estimation. In addition, there are several other effects, which may interact with response formats, which are of importance. First is the impact of numeracy, for both frequency and subjective probability estimation. For frequency estimation, context memory is another potentially important factor which may interact with response format as well. For subjective probability estimation, age may also have an important interactive effect. Finally, it is important to look at the logical

consistency between the two response formats, and as well as examine the effect of numeracy on logical consistency of translation.

Study 1

Data and Methods

The first data set comes from an experiment conducted at the University of Nebraska-Lincoln. The experiment is a factorial design, with two between-subjects factors and one within subjects. The purpose of the experiment is to determine which factors are related to accuracy of response, in particular the effects of those with the different response formats of vague quantifiers and numeric open ended responses. Each of the factors is related to the encoding, recalling, and reporting of frequency data. All 124 participants studied word lists, with word pairings of a target word and a context word. Subjects studied these words lists in groups, ranging from 5-20, in a classroom on a screen at the front of the room. All participants were given the instructions that word pairs would be shown, which word was to be recalled, and that a memory test for the frequency of some of the words would be given after the presentation of the list.[6] The instructions were similar to those used in Brown's (1995) Experiment 1, in which respondents were also told that a memory test would be given after a being presented a set of word pairs. However, the instruction differed from Brown (1995) in that the nature

---

[6] The exact instructions given were "This experiment is about memory, and will ask you to recall a number of words that will be presented to you on the screen at the front of the room. A pair of words will be presented. The first word is in all capital letters and is the words you will be asked to recall at the end of the experiment. The second word is to provide an example of the word you are to recall, in order to help your memory. You will be asked to recall how many times the capital words occurred in the list. If there are no questions, I will begin the word list presentation, which will take about 10 minutes. Afterward, I will give you a set of questions to answer."

of the memory test was not specified, whereas in this study the nature of the test was specified. After the presentation of the word list, there was a task to test numeracy, which completes the purpose of a filler task as well as collect numeracy information, prior to the asking about word frequencies. The numeracy test and word frequency test were conducted on paper forms. The main outcome variable is the accuracy of the response to the question how many times a target word was presented in the list.

The experiment employed a 2 (context: same; different) x 6 (frequency: 0, 2, 4, 8, 12, 16) x 2 (response form: open-ended numeric; vague scale) factorial design, with the context and form factors manipulated between subjects, and the frequency factor manipulated within subjects. The within subjects factor is the number of times the target word is presented in a list, i.e. the actual frequency. Target words were presented 2, 4, 8, 12, or 16 times, for six seconds each as was done in Brown (1995). Further, respondents were asked about target words that have not been presented, i.e. have a zero actual frequency. Studies have shown that different levels of the actual frequency lead to different recall strategies (Brown 1995, 1997, Conrad et al. 1998). At times, some strategies appear to require numeric translation to answer the given numeric open ended response option. As such, these strategies may not only directly affect the accuracy of the answer, but also may interact with other question format options, such as the response option given.

One between-subject factor manipulated the type of context word used along with the target word. There are two conditions for this factor: same-context condition and different-context condition. This format follows the conditions used in Brown (1995) which has been shown to affect recall strategy selection. In the same-context condition,

the target word is presented with the same context word at every presentation. This context word is an exemplar for the target word. For example, for the target word CITY, it would be presented with the only one context word, such as the exemplar Miami at every presentation of the word CITY. Conversely, for the different-context condition, the target word is presented in combination with a different context word at every presentation, again with each context word an exemplar of the target word. For example, for the target word CITY, for one presentation the exemplar Miami would be presented, for the next, New York, for the third, Chicago, etc. (i.e. CITY-Miami, CITY-New York, CITY-Chicago, etc.).

Target words and exemplars come from Van Overschelde et al. (2004) and McEvoy and Nelson (1982), studies of categories (targets) and category instances/norms (exemplars). A total of fifteen target words are used, and are selected based on similar criteria as in Brown (1995): first, the target word had to be clearly identified by a single noun, and second, each target word had to have a substantial number of category instances (norms). Specifically, each word had to have at least sixteen category instances so that each could be presented to have any actual frequency in the different context condition. Target words and exemplars selected for use are located in Appendix 2. In each list, all fifteen target words will be presented according to one of the five actual presentation frequencies, i.e. 2, 4, 8, 12, or 16 times.

Given the five levels of actual frequency and the fifteen words, three words are presented at each of the five actual frequencies. This presentation strategy leads to 126 presentations of target words in the study list. In order to create lists where each word is presented at each level of frequency, groups of three target words (from the fifteen) were

created, and these groups were varied at the five levels of frequency. This grouping leads to five groups of three words, requiring five lists to ensure that each target word would be presented at each level of actual frequency (i.e. CITY presented twice on one list, four times on another, and so on through CITY being presented 16 times on one list). There were five lists for the same context condition and five lists for the different context condition.

Target words were randomized in presentation throughout the lists. A random number was generated and assigned to each target word in each of the ten lists (five for each of the context conditions). The lists were then sorted by the random number, with the word with the smallest random number assigned appearing first, and the largest random number appearing last. Thus, the position of any given word was completely random. For the same-context condition, each target word in this list was paired with the selected context word. The same exemplar word was randomly selected and used for every presentation of the target word across all presentations of the target word in the same-context condition. For the different-context condition, the same lists were used with the difference being in that for each target word, a different exemplar is used. The exemplars for each target were selected at random, without replacement. Again, this randomization technique is the same as used in Brown (1995).

The second between subject factor, and of focal interest, is the different response options offered, focusing on numeric open response and vague quantifier response options, for reasons noted above. In one condition, respondents responded to the query for the frequency of a given target word, which has been presented 0, 2, 4, 8, 12, or 16 times, using an numeric open ended response. The question asked, “How often did

WORD appear in the presented list? _____times", where the blank was filled in by the respondent using any number. In the vague quantifier condition, respondents were asked the same question, but instead of a blank to fill in a number, they were presented vague quantifier response options. The vague quantifier options chosen come from Pohl (1981), with a desire to find words thought to be naturally ordered and creating a balanced scale. These options are "Never, Not Often, Somewhat Often, Fairly Often, Quite Often, Very Often". Six vague quantifier scale points were chosen in order to match the number of actual values used. The test for both conditions included questions about the fifteen presented words as well three additional words that were not presented at all, for a test of zero presented frequencies. For this test phase of the experiment, target words questioned about were randomized, with the targets asked about in a random order.

Following answering all of the questions about frequency of the target words, respondents in the vague quantifier condition were asked for a numeric translation of the response options. Specifically, participants were asked to give a numeric translation to each of the six vague quantifiers used in the above question. They were asked to translate each of the six vague quantifiers answering the question, "In the past test, how many times did you think the word WORD meant?", where WORD replaced each of the six vague quantifiers.

All respondents were also asked to complete a numeracy questionnaire in between presentation of the word lists and frequency questions. This questionnaire served both as a filler task between the presentation and testing of word lists and as an instrument to collect participants' level of numeracy information. The numeracy questionnaire consists

of nine test questions that had been earlier administered by Galesic and Garcia-Retamero (2010), and is presented in Appendix 3.

Subjects were selected from the University of Nebraska – Lincoln experimental subject pool, through the university's psychology department. An advertisement was placed to obtain participation in exchange for completion of course requirements. The total number of respondents selected is based in part on findings in a study of somewhat similar design, Brown (1995), which also used the variation of actual frequency and the same-different context factor (see below). In a test of the difference of absolute error between respondents in the same and different context (discussed above in the literature review), the statistical tests based on and an analysis of variance (ANOVA) included F(1,38) = 14.0, indicating greater absolute error for the same context condition (Brown 1995, p. 1543). When the degrees of the numerator for the F-score is one, as it is in this test, this F-score can be translated to an effect size (i.e. correlation, r) using the formula, $r = \sqrt{F/(F + df_D)}$ where F is the value of the statistical test (in this case, 14.0) and df_D is the degrees of the freedom from the denominator of the test, in this case, 38 (Friedman 1982). The F-score from Brown (1995) (F(1,38) = 14.0) translates to an effect size of r = 0.5188. Another test of mean difference between rank-order correlations between estimated and actual frequency for the same and different context conditions led to the test results of t(38) = 3.8, indicating a larger mean correlation for different conditions, suggesting greater relative accuracy in the different context condition (Brown 1995, p. 1543). A t-score, like an F-score, can be translated into an effect size, in a similar manner. In this case, the r is calculated using the equation $r = \sqrt{t^2/(\frac{t^2}{df})}$, where t equals

the test score and df is the degrees of freedom of the test. The equation for the t-score translation to effect size is essentially the same equation as the translation of the F-score into effect size, since $F = t^2$. Using the test t(38) = 3.8 translates to an effect size of r = 0.5247.

Using these effect sizes as a basis for the expected effect sizes to be found in this study, based on the table in Friedman (1982), the sample size needed for each combination of the between subjects factors is between 20 and 26, for r = .55 and r = .50, respectively. Since the effect sizes found in Brown (1995) fall between these values, the sample size selected for each combination of the between-subject factors is the midpoint is sample size values, 23. For a 2 x 2 between subjects design, 23 subjects per each factor combination leads to a total sample size required of 92. Using a simpler design, with only one between subjects factor, Brown (1995) used 40 respondents, so this increase of more than double the subjects appears reasonable with the addition of the additional two-level between-subjects factor.

Using this power analysis as a guideline, 124 subjects participated. These subjects were randomly assigned to one of the four between-subjects combinations. More subjects were tested to ensure greater power than was required as indicated by the power analysis. Of these 124, 63 completed the numeric open ended response form and 61 completed the vague quantifier response form. For the presentation context manipulation, 67 were presented the different context, with different context words presented with each target word. The remaining 57 were presented the same presentation context condition.

Hypotheses

H1: People, in general, will be more accurate when responding to frequency questions when using vague quantifiers rather than numeric open ended responses

H2: When asked for a direct numeric translation of vague quantifiers, there should be a logical consistency between numeric and vague quantifier measures.

H3: As actual frequency increases, accuracy will decrease

H4: Greater context memory will improve accuracy

H5: In general, more numerate individuals will provide more accurate responses

H6: More numerate individuals will show greater logical consistency in numeric open ended and vague quantifier responses than less numerate individuals

H7: More numerate individuals will be more accurate using numeric open ended responses than less numerate individuals

H8: More numerate individuals will be more accurate using numeric open ended responses than when using vague quantifiers

H9: As the actual frequency increases, vague quantifier responses will become more accurate.

H10: As the actual frequency decreases, the accuracy of numeric open ended responses will increase.

H11: At lower levels of actual frequency, numeric open ended responses will be more accurate than vague quantifier responses.

H12: There will be no discernible effect of context memory on the accuracy of vague quantifier responses.

H13: There will be an interaction effect between context memory and actual frequency in regards to accuracy, whereby greater context memory will lead to better accuracy at higher levels of actual frequency, but not at lower levels of actual frequency.

H14: There will be an interaction between numeracy and actual frequency in regards to accuracy, whereby greater numeracy increases accuracy at higher levels of actual frequency but not at lower levels.

H15: A three way interaction will arise between response format, actual frequency and numeracy, whereby the most numerate will show higher accuracy at higher levels of actual frequency using numeric open ended responses, but no effect at lower levels of actual frequency.

Results

Numeracy

Overall, the 124 respondents displayed a high level of numeracy. Out of the nine numeracy questions used from Galesic and Garcia-Retamero (2010), the mean number of correct responses is 7.27 (S.E. = 0.14). The median number of correct responses was 8 out of 9. Thirty respondents (24.2% of the total) answered all nine correctly, another 37 (29.84%) answered eight correctly, and 26 (20.97%) got seven correct. The minimum number correct was 3, accomplished by 3 respondents (2.42%). Since there were none that scored zero, giving no true zero point for use in interpretation in analysis, all further analyses using numeracy uses a grand-mean centered measure of numeracy. That is, the mean (7.27) was subtracted from each respondent's numeracy score in order to have a numeracy measure with mean zero.

Logical Consistency

Of the 61 respondents that answered the frequency questions using the vague quantifier scales, and thus provided numeric translations for each of the vague terms used in the scales, 58 provided whole number translations, as intended. The remaining three respondents gave translations in terms of percentiles. Since percentiles are not the correct metric, these responses were not used in the logical consistency assessment. These three respondents, however, all gave the translation of “never” as meaning zero, and so this value was used for their numeric translation for this term. For the other five vague terms, for use in assessing accuracy, the mean translated value of the total sample was imputed for these three respondents for the value their numeric translations of these terms.

The means for all six vague terms, as well as the standard error, and the minimum and maximum for each are presented in Table 1.1. All respondents except one translated “never” as meaning zero, with this one respondent translating “never” as meaning three times. In all cases, there was complete consistency within respondents. That is, each respondent gave larger numeric translations along each point of the vague quantifier scale such that all translations followed the pattern: never < not often < somewhat often < fairly often < quite often < very often. This is mirrored by the means of the total sample in the below table. This same ordering of the vague quantifier scale translation occurs at the aggregate level as well, with never being the smallest and very often the largest. This may be in part due to the way that the questions were asked. All respondents in the vague quantifier condition were asked at the end of the frequency estimation task to translate each of the scale points, in the same order they appeared in the frequency estimation task. As such, the practice of responding to each of the frequency estimations questions and

the ordering of the translation questions may have indicated to the respondent that there was a correct ordering, with each ascending point meaning a larger value.

Table 1.1

Numeric Translations for Vague Quantifier Scale

	Mean	(S.E.)	Minimum	Maximum
Never	0.05	(0.05)	0	3
Not Often	2.67	(0.16)	1	5
Somewhat Often	5.57	(0.30)	3	13
Fairly Often	8.74	(0.44)	4	15
Quite Often	12.10	(0.64)	5	25
Very Often	15.48	(0.87)	7	37

All of the differences between the mean translations seen in the above table are significant at the $p<0.01$ level, further suggesting the logical consistency within the sample for the translation of vague quantifiers. Given the strong level of logical consistency, there does not appear to be any discordance between numeric and vague quantifier responses. As such, Hypothesis 2 appears to be supported. Since all respondents were completely logically consistent, with everyone giving the same ordering of vague quantifier numeric translations, it is clear that numeracy did not have an impact on logical consistency. Therefore, there is no support for Hypothesis 6.[7]

[7] As an alternative test, actual differences between estimated values of each vague quantifier and the intended value were examined for each individual. For example, the difference between the assigned value for “very often” and 16, the intended value for that quantifier, was estimated for each individual. These differences were then summed and correlated with numeracy score, which was found to not be significantly

Finally, it is worth noting that the mean values in Table 1.1 fall remarkably close to the six possible values for the actual values of the words presented in the word lists. There is still a significant amount of variation, in part indicated by the minimum and maximum values given for each translation, with greater dispersion at the higher end of the vague quantifier scale. Even with this variation, though, the mean values of the translations for the sample are similar to the actual values of 0, 2, 4, 8, 12, and 16 used. The reason that six vague quantifier scale points were chosen was in order to match the number of actual values used, and this shows the similarity between the two, possibly indicating an overall high level of accuracy (see below).

Accuracy

Accuracy may be measured in a number of ways, but it is first necessary to place numeric values onto the vague quantifiers, that in essence conform to the distribution of the actual frequency. It is necessary to assume some distribution of values to be assigned to the vague quantifiers, or else accuracy cannot be tested with these values. Standard assignment of values to such scales where there is only a unit difference between scale points, such as 1, 2, 3, 4, for a four point scale, will obviously be unsatisfactory for assessing many measures of accuracy, which is usually defined as the difference between the estimated value (the response) and the actual value (the actual frequency). Assigning values to vague quantifier scales has been used in past research to assess accuracy based on percentage of time a behavior was completed (Lu et al. 2008).

---

different from zero ($r = -0.17$, $p = 0.19$). This finding further suggests there was no effect on numeracy on logical consistency.

In this analysis, following the suggestion of Bradburn and Miles (1979) values are assigned to vague quantifier responses based on the individual respondents' translation of these quantifiers to numeric values. As part of the experimental procedure used, subjects gave one translation of what the vague quantifiers meant at the end of the frequency recall test, and the values are used for each instance that quantifier was used. For example, a respondent who said that "very often" translated to 17 times, has the value 17 used for each time they selected that a word had occurred very often. For three respondents, as noted above, the translations given were in the form of percentiles rather than in terms of whole numbers, except for the translation of "never", which was given for all three as zero. For these respondents, the mean translations of the sample are used as their values for the remaining five vague quantifiers (zero is used for never translation).

Accuracy can be measured several different ways. Four are used in this analysis. First, the absolute difference between the reported frequency and the actual frequency can be used (i.e. |actual frequency – estimated frequency|). This difference estimates the total accuracy (error) in the measure, but fails to capture whether a measure is more or less error-prone in a particular direction, such as over- or underestimation of actual frequency. As such, another measure of accuracy is the signed difference (i.e. actual frequency – estimated frequency).

Two additional previously used measures that are to be used here include the rank-order correlations between the actual frequency and the estimated frequency, and the regression slope fitting estimated frequency to actual frequency (Brown 1995). The correlation tests the relative accuracy of a participant, while the regression slope tests the

degree to which there is a bias overestimating or underestimating actual frequency. These two measures have to be estimated over all responses that a respondent gives, whereas the absolute and signed differences are calculated at the response level, with each response having a particular level of error. Therefore, the effect of actual frequency is only examinable when looking at absolute and signed differences.

Examining the regression slopes, overall, respondents underestimated actual frequencies, as indicated by the overall mean of 0.72 (S.E. = 0.03). The mean did not differ significantly for those responding using vague quantifier (M = 0.71) or numeric open ended responses (M = 0.73) as tested by a pairwise t-test $t(122) = 0.32$, $p = 0.75$. Similarly, contrary to Brown's (1995) findings, there was no difference between those in the same context (M = 0.73) and different context (M = 0.71), $t(122) = 0.23$, $p = 0.81$ conditions. Finally, the correlation between the slopes and numeracy (mean centered) was -0.05, and not significantly different from zero, $p = 0.55$.

To control for all of these effects simultaneously, and to test for possible interactions between these potentially important variables, a linear regression model was estimated, with slope of the participants as the dependent variable. The independent variables are the response form used (vague quantifier form = 1, numeric open ended form = 0), the context words used with the target words (same context = 1, different context = 0), numeracy (mean centered), and all of the possible interactions between these three variables. Results are presented in Table 1.2. None of the independent variables are estimated to be significantly different from zero. Mirroring this fact is that the omnibus F-test fails to reject the null hypothesis that the addition of any of the independent variables has an effect over the intercept. In other words, the expected value

of the dependent variables, the slopes, is constant over all independent variables, including all of the interactions.

Table 1.2

Regression on Respondents' Slopes

Variable	Coefficient	(s.e.)
Same Context	0.002	(0.10)
Vague Quantifier Form	-0.03	(0.09)
Numeracy	-0.00005	(0.04)
Context*Numeracy	0.05	(0.06)
Context*Form	0.04	(0.14)
Numeracy*Form	-0.05	(0.06)
Context*Form*Numeracy	-0.05	(0.09)
Intercept	0.73*	(0.07)

n = 124, F = 0.46, p = 0.86, R^2 = 0.03, *p < 0.05

The results examining the rank-order correlations are the nearly the same as that for the slopes. Overall, there tends to be a high correlation between estimated frequency and actual frequency, r = 0.75. This suggests that overall, there was a high level of relative accuracy. For analysis of correlations, such as differences in means and regressions, correlations were transformed to Fisher z-scores. Using these, it is found that there is no difference between vague quantifier (M = 1.06) or numeric open ended responses (M = 1.08) as tested by a pairwise t-test t(122) = 0.46, p = 0.64. There is also no difference between same and different context in terms of the correlations, t(122) = 0.42, p = 0.68 (same context M = 1.06, different context M = 1.09). There was also not a

significant correlation between numeracy and the Fisher z-score transformations, $r = 0.09$, $p = 0.35$.

Again, to control for all of the effects and all of the possible interactions simultaneously, linear regression was employed. The dependent variable is the transformed correlations to Fisher's z-scores, with the independent variables again being the response form, context words used, numeracy, and all of the interactions of these three. Results are presented in Table 1.3. As with the regression on the slopes, none of the coefficients for any of the independent variables reached the standard level of significance of $p<0.05$. The main effect of numeracy, however, did reach significance at $p<0.10$. If the higher level of alpha is accepted, this suggests that those with higher levels of numeracy also display somewhat higher correlations, and thus higher relative accuracy, in the cases when the response form and context used both equal zero. That is, numeracy increases correlations when numeric open ended responses are used in the different context condition. However, it is again noted that this coefficient is not significant at the standard $p<0.05$ level. Further, the omnibus F-test fails to reject the null hypothesis that the addition of any of the independent variables has an effect over the intercept. In other words, the expected value of the dependent variables, the slopes, is constant over all independent variables, including the main effect for numeracy.

Table 1.3

Regression on Respondents' Correlations

Variable	Coefficient	(s.e.)
Same Context	-0.004	(0.09)
Vague Quantifier Form	-0.009	(0.09)
Numeracy	0.06	(0.04)
Context*Numeracy	-0.06	(0.06)
Context*Form	-0.06	(0.13)
Numeracy*Form	-0.08	(0.05)
Context*Form*Numeracy	0.09	(0.09)
Intercept	1.10*	(0.06)

n = 124, F = 0.55, p = 0.80, R^2 = 0.03, *p < 0.05

Given the lack of identified effects from the previous analyses on the two forms of accuracy, regression slopes and rank-order correlations, there appears to be little to no support for many of the hypotheses that can be tested using these measures. The only possible exception is that there is a marginal significance for the main effect of numeracy in the regression on respondents' correlation coefficients. This marginal significance suggests, possibly, that numeracy leads to higher level of (relative) accuracy, giving some support to Hypothesis 5, that more numerate individuals will provide more accurate responses.

The only remaining hypothesis that is supported is Hypothesis 13, which predicted a null result in terms of no interaction between context memory and vague quantifier response in regards to accuracy. Indeed, this null result was found. Otherwise, none of the potentially testable hypotheses were supported. Hypotheses 1, 4, and 8 all

were not supported. The remaining hypotheses include some discussion of actual frequency, which neither of these measures of accuracy can address.

Therefore, to examine the effect of actual frequency and the effect of context, response form, and numeracy on different forms of accuracy, the absolute and signed differences are analyzed. These can be analyzed at the true frequency count levels; that is, it is possible to examine the effect of increasing the actual count on these differences. Examining signed differences first, Table 1.4 presents the mean signed difference at each level of actual frequency, i.e. 0, 2, 4, 8, 12, and 16, for each version of the response form and context condition combination that respondents fell under. For example, Vague-Same means respondents who were responded using vague quantifier responses and were presented the same context condition.

Table 1.4

Signed Differences at Levels of Actual Frequency, by Form and Context

Actual Frequency	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0	0.34*	0.39*	0.92*	0.73*
2	1.25*	0.76*	2.67*	1.55*
4	0.85	0.53	2.88*	2.73*
8	0.80*	-0.94	1.36*	1.50*
12	-1.52*	-3.26*	0.05	-1.43*
16	-3.80*	-3.49*	-3.14*	-2.69*
Overall	-0.34	-1.08*	0.79*	0.34

*$p < 0.05$

As can be seen, generally across all conditions, it appears that at lower levels of actual frequency there are more overestimation errors, and at higher levels of actual frequency, there is greater underestimation. Across all levels of actual frequency, it appears that there is slight underestimation for vague quantifiers and slight overestimation for numeric open ended responses, however, the overall signed differences are not significantly different from zero for the vague quantifier-same context combination and numeric open ended-different context response combination. Both response forms show a similar pattern in the direction of errors at each level of actual frequency, except in the middle, where actual frequency equals eight. Here, numeric open ended responses still are overestimated, while for vague quantifier responses in the different context condition the shift to underestimation has already happened. For both sets of numeric open ended responses, this shift to underestimation occurs at the next highest level of actual frequency. Also, at lower levels of actual frequency, the size of the error is smaller for vague quantifier responses, but at the two highest levels of actual frequency, numeric open ended responses show somewhat smaller error sizes.

To further examine the effect of actual frequency, as well the effect of response form, word context, and numeracy, a multivariate model should be used. Given the structure of the data, hierarchical linear modeling (HLM), is most appropriate (Luke 2004, Raudenbush and Bryk 2002). The respondents each gave a total of eighteen frequency estimation responses, three for each level of actual frequency. Thus for each response, there is a different actual frequency value associated with the response. These responses are nested within each respondent, each of whom has respondent level

characteristics of interest. Specifically, numeracy, response form, and presentation context vary at the respondent level, not at the response level.

The model is therefore a two-level model, modeling response accuracy for each response, with the level-1 independent variable being actual frequency. The second level of the model includes random effects at the respondent level (j). The level two, i.e. respondent level, covariates are the aforementioned numeracy (mean centered), response form, and presentation context. To ensure all interactions of the effects are accounted for, as well as main effects, both intercepts and slopes are modeled as random and used as outcomes (Luke 2004). For the signed differences form of the model, the model can be written using the following equations. First, using the HLM notation, the model is presented in the following three equations:

Level-1 Model:

$$SIGNED_{ij} = \beta_{0j} + \beta_{1j}(Actual\ Frequency) + r_{ij} \quad (1)$$

Level-2 Model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Context) + \gamma_{02}(Form) + \gamma_{03}(Numeracy) + \gamma_{04}(Context * Numeracy) + \gamma_{05}(Context * Form) + \gamma_{06}(Form * Numeracy) + \gamma_{07}(Context * Numeracy * Form) + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Context) + \gamma_{12}(Form) + \gamma_{13}(Numeracy) + \gamma_{14}(Context * Numeracy) + \gamma_{15}(Context * Form) + \gamma_{16}(Form * Numeracy) + \gamma_{17}(Context * Numeracy * Form) + u_{1j} \quad (3)$$

Where *i* represent the individual level and *j* represents the respondent level, β and γ are parameter coefficients, and *r* and *u* are error terms for the various portions of the equation. Rewriting these equations in the mixed model format makes clear the full

model and how all of the interactions are modeled, as well as the fixed and random portions of the model. This is done in the following equation:

$$SIGNED_{ij} = \gamma_{00} + \gamma_{01}(Context) + \gamma_{02}(Form) + \gamma_{03}(Numeracy) + \gamma_{04}(Context * Numeracy) + \gamma_{05}(Context * Form) + \gamma_{06}(Form * Numeracy) + \gamma_{07}(Context * Numeracy * Form) + \gamma_{10}(Actual\ Frequency) + \gamma_{11}(Context * Actual\ Frequency) + \gamma_{12}(Form * Actual\ Frequency) + \gamma_{13}(Numeracy * Actual\ Frequency) + \gamma_{14}(Context * Numeracy * Actual\ Frequency) + \gamma_{15}(Context * Form * Actual\ Frequency) + \gamma_{16}(Form * Numeracy * Actual\ Frequency) + \gamma_{17}(Context * Form * Numeracy * Actual\ Frequency) + u_{1j}(Actual\ Frequency) + u_{0j} + r_{ij} \quad (4)$$

The $u_{1j}(Actual\ Frequency) + u_{0j} + r_{ij}$ portion of equation 4 is the random portion of the model; the remainder is the fixed part of the model. Before running this model, a random-intercept only model is run with signed differences as the dependent variable, in order to calculate the intraclass correlation coefficient (ICC). The ICC measures the proportion of the variance in the dependent variable (signed differences in this case) accounted for by the level 2 units, i.e. respondents (Luke 2004). Formally, this model is written, in HLM terms, by the following equations:

$$SIGNED_{ij} = \beta_{0j} + r_{ij} \quad (5)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (6)$$

In mixed model terms the equation is as follows, which can be seen to be the same as a one-way random effects ANOVA model (Luke 2004):

$$SIGNED_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (7)$$

All models are calculated using HLM 7 (Raudenbush et al. 2011).The calculated ICC is 0.33, suggesting that respondents account for about 33% of the variability in the observed signed differences. This ICC is moderately high, suggesting the potential

usefulness of multilevel models (Luke 2004). Given that, the full model is presented in Table 1.5. Robust standard errors are used, that is, standard errors that are consistent even when HLM assumptions are not (Raudenbush et al. 2011). Only the full model is presented, as it alone allows for an examination of all hypotheses simultaneously. Thus, no reduced models are examined.

Table 1.5

Hierarchical Linear Model of Signed Differences on Response and Respondent Characteristics

Variable	Coefficient	(s.e.)
For Intercept1, β_0		
Intercept2, γ_{00}	2.22*	(0.37)
Same Context, γ_{01}	0.59	(0.63)
Form, γ_{02}	-1.15*	(0.44)
Numeracy, γ_{03}	-0.60*	(0.25)
Context*Numeracy, γ_{04}	0.88*	(0.43)
Context*Form, γ_{05}	-0.07	(0.72)
Form*Numeracy, γ_{06}	0.61*	(0.29)
Context*Form*Numeracy,γ_{07}	-0.99*	(0.47)
For Actual Frequency slope, β_1		
Actual Frequency, γ_{10}	-0.28*	(0.06)
Context*Frequency, γ_{11}	-0.01	(0.09)
Form*Frequency, γ_{12}	-0.02	(0.08)
Numeracy*Frequency, γ_{13}	0.003	(0.04)
Context*Numeracy*Frequency, γ_{14}	0.05	(0.05)
Context*Form*Frequency, γ_{15}	0.05	(0.14)
Form*Numeracy*Frequency, γ_{16}	-0.05	(0.05)
Context*Form*Numeracy*Frequency, γ_{17}	-0.05	(0.09)

n = 124, *p < 0.05

The results indicate a number of interesting findings. First, in regards to the overall mean (i.e. intercepts as the outcome) of the signed differences Table 1.6 gives predicted values of the mean. Given the more difficult nature of interpreting signed differences from a regression model, predicted values of the mean (i.e. intercept) allow for increased ability to interpret results. The first noticeable result from the model is that presentation context on its own does not appear to have an effect. However, the main effect of response form suggests that, when mean centered numeracy is at its mean (i.e. numeracy = 0) and when different presentation context is used, vague quantifiers lower the signed mean difference. Since the overall mean is positive, suggesting overestimation (intercept = 2.22), and the reduction caused by vague quantifier response is not greater than the mean (thus causing the mean to become negative), this reduction can be viewed as an overall reduction in error, with lower overestimation. When numeracy is equal to zero, and same context is used, the reduction in the mean error is somewhat greater, but is not statistically different from that of the different context condition. These findings suggest that, at mean levels of numeracy, vague quantifier responses reduce error by reducing overestimation error.

Table 1.6

Predicted Mean Signed Error by Context, Form, and Numeracy

Numeracy	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0 (Mean)	1.59	1.08	2.81	2.22
1.73 (High)	1.41	1.09	3.29	1.18
-4.27 (Low)	2.05	1.05	1.65	4.80

When numeracy deviates from the mean of zero, the picture becomes more complicated. Since the maximum value of mean centered numeracy is 1.73 (minimum -4.27), several conclusions can be drawn. First, for the numeric open ended response form and given the different context presentation, greater numeracy leads error to be reduced, while lower than average numeracy increases overestimation error (see Table 1.6). As also seen in Tables 1.5 and 1.6, at higher levels of numeracy, vague quantifier response forms decrease overestimation error overall for both presentation contexts, although there is a bit more of a reduction in the different context condition. Similarly, for the different context condition, greater numeracy leads to less overestimation when using a numeric open ended response form. However, for the same context condition and numeric open ended response form, overestimation is increased at the highest levels of numeracy. Compared to the mean numeracy, overestimation error is decreased or within rounding error of three of the four instances, the lone outlier being the same context, numeric open ended form where higher numeracy led to greater overestimation error.

Subjects with lower numeracy, interestingly, show higher overestimation errors than respondents with higher numeracy in two cases, but lower overestimation errors in the other two cases (see Table 1.6) Specifically, those with the lowest numeracy given the same context presentation and vague quantifier form had somewhat higher levels of overestimation error (though still smaller than the overall mean) than those with higher levels of numeracy. Similarly, those with the lowest numeracy given different context and numeric open ended response had the highest overestimation error of all categorizations. However, consistent with other levels of numeracy, different context and vague quantifier response led to the lowest levels of overestimation error. What is

somewhat different is that for the same context condition and use of the numeric open ended form, there was reduced overestimation error compared to the overall mean of numeracy and higher levels of numeracy. Therefore, although it appears that in most cases numeracy decreases overestimation error, there are a few cases where unexpectedly this does not occur.

Greater clarity is at hand when examining the effect of actual frequency and the interactions between the respondent characteristics and actual frequency (slopes as the outcome). First, it is evident increases in actual frequency lead to increasing likelihood that there will be underestimation. This underestimation with increasing actual frequency is evident in Table 1.4 as well. However, none of the interactions with actual frequency are statistically significant. This lack of significance suggests that there is no relationship between respondent level and response level variables. Thus, although the effect of actual frequency is evident, actual frequency does not interact with any other variables to affect the signed differences.

The results of this hierarchical model provide support to some of the hypotheses proposed and lack of support to others. First, there is support for Hypothesis 1, in that generally, people were more accurate (i.e. overestimation error was reduced) when using vague quantifier responses rather than numeric open ended responses. In examination of Table 1.6, comparing response forms across similar context conditions, vague quantifier responses produced less overestimation error in all but one instance. Further, vague quantifier responses reduced error from the overall mean in every case.

Hypothesis 3 is also somewhat supported, in that increases in actual frequency altered the level of accuracy such that at the highest levels of actual frequency, there was

greater underestimation. However, it is not as clear that at all increasing levels of actual frequency there are decreases in accuracy, which are due to the nature of signed differences (see Table 1.4). At mid-levels of actual frequency (i.e. actual frequency = 8), it appears that lower levels of error may be occurring than at points lower and higher than 8 on the actual frequency scale. This effect may be simply due to the shifting of some respondents to begin underestimating, as is clearer at higher levels of actual frequency, while others respondents are still overestimating, as with lower levels of actual frequency. If some are overestimating and others underestimating, the average effect may make it appear that there are lower levels of error. Given this shifting from over- to underestimating, it is not always clear that increasing the actual frequency will lead to reduced errors overall when examining signed differences. This issue of signed differences is resolved when examining absolute differences below.

Still, other hypotheses are more clearly supported or not supported using the signed differences model. Hypothesis 4, that greater context memory (i.e. different context presentation) on its own reduces error is not supported. The main effect for context memory is not significant, and given that all interactions equal zero when context presentation is set to different context (i.e. different context memory = 0), the main effect is the effect of interest. Hypothesis 5 is also not clearly supported, since although greater numeracy does at times appear to decrease errors, at times it appears it can increase error, depending on the context memory and response form. Similarly, Hypothesis 7 is not clearly supported, since it depends on context memory whether more numerate individuals are more accurate than less numerate respondents using numeric open ended responses. When more numerate individuals responded on the numeric open ended form,

this decreased error generally for different context presentation, but increased overestimation for same context presentation. The reverse pattern was true for lower numeracy respondents when using numeric open ended response forms.

Hypothesis 8 is somewhat supported, as more numerate individuals are more accurate when using numeric open ended responses, but only when given the different context condition. In the same context condition, higher numeracy leads to a reduction in overestimation error when responding to vague quantifier response form. Hypothesis 12 is not supported, however, as there is some evidence that context memory does have an effect on vague quantifier response accuracy, as seen by the three way interaction between context, response form, and numeracy. More numerate individuals responding in the same condition shows less overestimation error when using vague quantifier response forms.

None of the hypotheses regarding interaction with actual frequency are supported. This lack of support for all of these hypotheses is evidenced by the lack of statistically significant interaction between actual frequency and any other effects. Therefore, Hypotheses 9, 10, 11, 13, 14, and 15 are not supported.

Signed differences detect the direction of error, but are less easily interpreted in terms of overall error. Absolute error detects total error more clearly, and is analyzed in a similar manner to that of the signed differences. First, Table 1.7 presents the mean signed difference at each level of actual frequency, i.e. 0, 2, 4, 8, 12, and 16, for each version of the response form and context condition combination that respondents fell under.

Table 1.7

Absolute Differences at Levels of Actual Frequency, by Form and Context

Actual Frequency	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0	0.34*	0.39*	0.92*	0.73*
2	1.99*	1.96*	3.05*	2.36*
4	3.15*	2.61*	4.05*	3.97*
8	4.55*	4.46*	4.23*	4.34*
12	5.52*	5.77*	5.02*	5.75*
16	7.02*	6.03*	6.52*	5.73*
Overall	3.76*	3.61*	3.96*	3.88*

*$p < 0.05$

Across all conditions, as actual frequency increases error also tends to increase, with all absolute differences significantly greater than zero. Across all levels of actual frequency, it appears that there is slightly less error for vague quantifiers than numeric open ended responses, and slightly less error for the different context condition. Vague quantifiers appear to have less error at lower levels of actual frequency than numeric open ended responses, but a slight reversal occurs at higher levels of actual frequency, with numeric responses showing somewhat smaller errors.

In order to examine absolute differences more completely, as was done with signed differences, a hierarchical linear model is used. The model used is identical to that described in equations 1-4, with the exception of using absolute differences as the dependent variable instead of signed differences. Similarly, the intraclass correlation is calculated using the same equations as 5-7, again substituting only absolute differences

for signed differences. The ICC is estimated as 0.15, suggesting that respondents account for about 15% of the variability in the observed absolute differences. The full model is presented in Table 1.8, with robust standard errors, and again, only the full model is used in order to test all of the hypotheses simultaneously.

As can be seen in Table 1.8, first examining the mean (intercepts as outcome), the main effects for different context condition, response form, and the interaction between these two are not found to be statistically significant. This indicates a lack of effect for these variables. However, the interaction of these variables with numeracy and the numeracy main effect are all statistically significant, indicating the importance of numeracy on absolute error. The coefficients suggest that for the different context condition when answering via numeric open ended responses, greater numeracy reduces error.

It is important to note, that even though the error is larger, it does not increase the overall amount of error from the grand mean. Further, greater numeracy actually increases error in the same context condition compared to those with lower numeracy presented in the same context condition, when responding using numeric open ended responses. However, when responding on the vague quantifier form in the same context condition, greater numeracy reduces error and less numeracy increases error. Even with the increase by responding to vague quantifier forms with the same context memory, the less numerate still overall have lower than the mean error. That is, overall, vague quantifier responses tend to either reduce the level of error or maintain the level of error at all levels of numeracy and context memory.

Table 1.8

Hierarchical Linear Model of Absolute Differences on Response and Respondent Characteristics

Variable	Coefficient	(s.e.)
For Intercept1, β_0		
Intercept2, γ_{00}	1.69*	(0.41)
Same Context, γ_{01}	0.27	(0.67)
Form, γ_{02}	-0.64	(0.47)
Numeracy, γ_{03}	-0.58*	(0.27)
Context*Numeracy, γ_{04}	1.03*	(0.44)
Context*Form, γ_{05}	-0.28	(0.75)
Form*Numeracy, γ_{06}	0.58*	(0.30)
Context*Form*Numeracy,γ_{07}	-0.97*	(0.48)
For Actual Frequency slope, β_1		
Actual Frequency, γ_{10}	0.29*	(0.04)
Context*Frequency, γ_{11}	-0.01	(0.06)
Form*Frequency, γ_{12}	-0.06	(0.05)
Numeracy*Frequency, γ_{13}	0.02	(0.02)
Context*Numeracy*Frequency, γ_{14}	-0.02	(0.03)
Context*Form*Frequency, γ_{15}	0.05	(0.08)
Form*Numeracy*Frequency, γ_{16}	-0.03	(0.03)
Context*Form*Numeracy*Frequency, γ_{17}	-0.01	(0.06)

n = 124, *p < 0.05

Table 1.9 displays these effects, presenting predicted means (intercepts) for various combinations of the three way interactions between different conditions and

levels of numeracy, as in Table 1.6 for signed differences. As indicated by the Context*Numeracy coefficient in Table 1.8, those with lower numeracy have smaller error in the same context condition. Indeed, the lowest error predicted is for those with the lowest numeracy in the same condition, using numeric open ended responses. Similarly, those with the highest levels of numeracy have higher levels of error in the same context condition compared to those with similarly high numeracy using the different context condition (within the same response form). The highest level of predicted error is for those with the lowest numeracy, when given the different presentation context and responding to the numeric open ended response form. Although generally it appears that vague quantifier response form leads to lower error, which is indeed the direction the model coefficients indicate, it is important to note that the some of the model coefficients are not statistically significant. Rather, taking that into consideration, it shows that error is reduced or maintained at about the mean level of estimated error when using vague quantifier responses.

Table 1.9

Predicted Mean Absolute Error by Context, Form, and Numeracy

Numeracy	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0 (Mean)	1.04	1.05	1.96	1.69
1.73 (High)	1.14	1.04	2.73	0.68
-4.27 (Low)	0.79	1.06	0.05	4.17

Examining the slopes coefficients in Table 1.8, it is evident that actual frequency increases error. This increase in error is also reflected in Table 1.7. However, as with the

model for signed differences, none of the interactions with actual frequency is statistically significant. This lack of significance suggests that there is no relationship between respondent level (context, response form, and numeracy) and response level (actual frequency) variables. Thus, although the effect of actual frequency is evident, actual frequency does not interact with any other variables to affect absolute error.

These results inform whether the proposed hypotheses are supported or not. In regards to absolute differences error, Hypothesis 1 is only partially supported, as in general, people were not more accurate using vague quantifier responses than when using numeric open ended responses. Rather, it appears that after controlling other factors, there is no difference between the two response forms. This similarity is evident by the lack of significance of the main effect of the response form. Further, the vague quantifier response form only significantly affects error at different levels of numeracy, which at times also depends on the context condition. However, when vague quantifiers did have an impact on responses through these interactions, it was in the direction of reducing error, thus the partial support for the hypothesis.

Hypothesis 3 is clearly supported, as increases in actual frequency increases error, i.e. decreases accuracy. However, like Hypothesis 1, Hypothesis 4 is not supported. The main effect for context is not significant, and the effect of context is only apparent at different levels of numeracy, and depending on the person's numeracy, the effect of context differs. That is, sometimes greater context memory (i.e. different context condition) improves accuracy and sometimes the opposite occurs. For similar reasons, Hypothesis 5 is not strongly supported. Although at times greater numeracy does lead to lower error (e.g. under different context and numeric open ended response format

conditions), at other times, higher numeracy leads to greater error (e.g. under same context and numeric open ended response format conditions).

Hypothesis 7 is only partially supported, as evidenced by the coefficients in Table 1.8. When the numeric open ended form is used, more numerate individuals display less error, but only under the different context condition. Under the same context condition, vague quantifier responses reduce error for those with higher numeracy. Hypothesis 8 is also partially supported, as seen by the Form*Numeracy interaction, with more numerate individuals having more error using the vague quantifier form, but only under the different context condition. For the same context condition, more numerate individuals display less error using the vague quantifier form. Hypothesis 12 is not supported, as some effect is found, as seen in the three way interaction Context*Form*Numeracy. This interaction suggests that context does affect error on the vague forms, at either higher or lower numeracy (in opposite directions).

Like the signed differences model, for the absolute error model, none of the hypotheses regarding the interactions with actual frequency are supported. The absence of support for all of these hypotheses is seen by the lack of statistically significant interactions between actual frequency and the other effects. Thus, Hypotheses 9, 10, 11, 13, 14, and 15 are not supported.

Discussions and Conclusions

This study is the first that has compared the accuracy between vague quantifier and numeric open ended responses, as well as examining logical consistency between the two response forms. The examination of logical consistency and accuracy was accomplished by using an experimental design in which actual frequencies were known.

Included in this experiment was the important factor of context memory, which has been shown to affect accuracy (Brown 1995). Two conditions were used, the same and different context conditions, as was also done in Brown (1995). Further, the effect of numeracy was examined on validity, as has been done in Studies 2 and 3 in this dissertation. However, unlike those other studies, which relied on proxy measures of numeracy available in the data sets obtained, this study generated numeracy scores based on a previously used scale (Galesic and Garcia-Retamero 2010). Thus, it was possible to examine the interaction between context memory and response form, along with numeracy and the actual frequency of the event.

The results of this experiment provide several findings. First, individuals in this study are highly numerate. This high numeracy is likely due to the population selected from: students enrolled at a university in a psychology course. The numeracy of the student sample here is significantly higher than that of the national samples used by Galesic and Garcia-Retamero (2010). However, although there is a high level of numeracy overall in the sample used here, there is variation. Some respondents display significantly lower levels of numeracy, with as low as 3 out of 9 questions correct. Further, this variation allowed for detection of differences in numeracy in regards to accuracy (discussed below).

The second finding of note is that of the strong logical consistency between vague quantifiers and numeric translations. All respondents ordered the translations of the vague quantifiers in a consistent manner, with “never” consistently being the smallest value and “very often” consistently being the largest, with intermediate values for the other scale points. This ordering is also observed at the aggregate level, with each increasing scale

point statistically significantly larger from never to very often. Further, all but one participant gave “never” a translation of zero, showing strong sematic logical consistency as well. Given the logical consistency shown by all respondents, numeracy does not have an impact on this consistency, contrary to expectations.

Numeracy does have an impact on accuracy, however. In fact, of all of the possible variables affecting accuracy, numeracy consistently has the most impact. Numeracy, in some form, affects the relative accuracy (correlation coefficients), signed error, and absolute error. It only does not appear to have an effect on the size of individual regression slopes. How this effect occurs, however, is not consistently in one direction or the other. In some situations, greater numeracy is associated with decreased error, in others increased error. For example, for numeric open ended responses in the different context condition, greater numeracy is associated with a reduction in error. However, in the same context condition and with numeric open ended responses, greater numeracy is associated with greater error.

Regardless, results support the view that numeracy is a potentially important variable in survey research that to this point has not been fully explored. There have been no studies, except those in this dissertation, that have attempted to relate the effect of numeracy to validity in a survey. This study found that there is indeed an impact, albeit one that is not consistently in one direction. What this suggests is that survey researchers should account for the potential impact of numeracy in survey response, taking a measure of numeracy and including it as a covariate, at least when the response at issue is quantitative in nature.

What appears to have less of an impact than numeracy is that of the level of context memory, as measured by the presentation context, either the same context or different context. Unlike Brown (1995), who found a clear effect for context in the direction that greater context memory (i.e. different context presentation) led to greater accuracy, this is not case in this experiment. Indeed, for regression slopes (bias) and correlation coefficients (relative accuracy) context has no apparent effect. For signed and absolute differences, context memory interacts mainly with numeracy (and sometimes response form) to have an effect. Like numeracy, however, the effect was not always in a consistent direction. For example, for both signed and absolute error greater numeracy is associated with greater error under the same context condition (using numeric open ended responses). However, like Brown (1995), there is a clear effect for the actual frequency of presentation, with greater frequency leading to increased error.

Although context memory is found to be important at times, depending on other respondent conditions, the differences between these findings and Brown (1995) may need to be reconciled. One potential important difference related to power is the number of words used in each study. Brown (1995) used twice the words used in this study, (126 vs. 252); as such, it may be that the importance of context memory occurs at a certain threshold, not reached in this study. Similarly, it may also be that the lack of effect is due to when given fewer words to study, participants were able to estimate the frequency better than when more items are presented, and context memory loses its importance, as the participants can rely on other strategies for recall.

Participants in Brown (1995) also were required to “think aloud” during the response process, explaining their thought process, in order to ascertain what recall

strategies were used. The current study did not require “think aloud” responses. It may also be that “thinking aloud” may affect recall in some way that affected the results. Finally, it should be noted that the instructions regarding the memory test seems to differ slightly between the two studies. Although both studies informed participants a memory test would follow the presentation of a set of words, Brown (1995) states that the type of memory test was not specified, whereas in this study participants were informed about which words they would be tested on. The difference in instructions could affect differences in encoding and recall. Studies have indeed found that differences in instruction affect these aspects of memory (Brown 1997). Still, there is lack of evidence of an impact of context memory in this study suggests that perhaps context memory should be studied further, to see if the results of Study 1 are anomalous, or to determine if context memory is important only in certain conditions, such as those administered by Brown (1995).

Finally, of particular importance for this research is the effect of response form, either vague quantifier or numeric open ended response form. For bias (regression slopes) and relative accuracy (correlation coefficients), the response forms do not appear to have any effect on error. That is, one response form did not prove to be more effective than the other on accuracy. For signed error and absolute error, response form appears to have a more immediate impact. Vague quantifier response form generally reduces the error in signed differences, but has a less clear impact on absolute differences. Although the coefficient for the main effect of vague quantifiers was in the expected direction of reducing absolute error, this coefficient was not significant. Still, based on the effect of the interactions, vague quantifier responses either maintained or reduced absolute error

from the overall mean error. Thus, this study suggests that vague quantifiers are overall at least equally as accurate as numeric open ended responses, and at times provide more accurate data, with some exceptions.

In situations in which exceptions are observed, vague quantifiers still reduce error from the overall mean, but perhaps not to the same extent as numeric open ended responses. These exceptions occur due to interactions with numeracy and context memory. However, the reverse is not true; numeric open ended responses do not always tend to reduce or maintain error. Rather, at times, numeric open ended responses clearly increase overall error, as seen above.

Based on these findings, few of the hypotheses posited are consistently supported. The only hypothesis clearly supported is that of Hypothesis 3, which states that as actual frequency increases, accuracy will decrease. Many of the other hypotheses receive partial support, due to the interactive nature of the model. At times error is increased in one condition or level of numeracy, but is decreased in a different combination that renders support for the hypothesis incomplete at best. Other hypotheses receive no support at all. Of particular importance, Hypothesis 1, that in general, vague quantifiers would be more accurate than numeric open ended responses received some support in two of the measures of accuracy, and is not contradicted.

Although this is the first study to examine accuracy differences between response formats, and how this is impacted by context memory and numeracy, there are some limitations to this study that warrant discussion. First, is the way that logical consistency is measured. Logical consistency was measured by the numeric translations at the end of the questionnaire by all respondents using the vague quantifier scale. For each translation

section, the scale points were ordered from top to bottom in order from never to very often. Since the order was given to the respondents in this way, as well as having experience with the scale from answering all the frequency questions, this could influence responses to ensure that there was greater logical consistency than might have otherwise been.

Second, although this study used a more complete measure of numeracy than have other studies, the responses to the numeracy scale led to a high skew, toward the upper end of numeracy. The overall mean was 7.27 out of 9, suggesting a highly numerate sample, compared to other samples (Galesic and Garcia-Retamero 2010). In part due to this, numeracy scores were mean centered. Even so, there is a ceiling effect in that numeracy scores can only reach so high (1.73) but can reach significantly lower (-4.27). Further, given the skew in the data to the upper end, there is not a large amount of variation in the data. Still, even though this may seem like a limitation, numeracy effects are still estimable.

Related to this skew in the numeracy are the population studied and the generalizability of results. Specifically, participants used in this study were selected from a subject pool drawn from university students taking a psychology course. As indicated by the numeracy scores, this population and sample is likely to be more cognitively able than an average person in the wider population. Given the nature of the population, generalizability of the results is in question. That is, it may not be possible to infer that the results found here would also be found in the general population.

However, it is worth noting that vague quantifiers are thought to be cognitively less demanding than numeric open ended responses (Bradburn and Miles 1979). If this is

indeed the case, as it is also argued here that it is, then the findings in this study suggest that it may be vague quantifiers will perform even better in the general public. In this study, vague quantifiers tended to perform at least as well as numeric open ended responses, and at times, better. If this is true among more a cognitively able sample, then for a less cognitively able sample, likely to come from a general population, then the cognitively easier vague quantifiers could perform even better.

This proposition has not been tested, however, and suggests an avenue for future research. That is, a similar kind of study should be conducted on a sample from the general population, where cognitive ability and numeracy would theoretically be lower and more similar to those samples used in other surveys. A further study could also restrict the population to those with the lowest cognitive ability and/or lowest numeracy to see what effect of the different response forms arise.

Additionally, different areas of frequency estimation could be studied for differences between vague quantifier and numeric open ended responses. This study examined only recall of word lists. However, it may be that other kinds of recall, for example, number of times a doctor was visited, may lead to different results. Beyond that, asking about word lists is not a frequent survey question; rather survey questions are often about behaviors or other numeric information. This study suggests recall of numeric information is sometimes affected by response form, as well as context memory and numeracy, but whether this holds for actual behaviors is an empirical question that can be answered.

Another aspect that could be studied further is the effect of context memory. According to Brown (1995), greater context memory increases accuracy. This study did

not find that the effect of context memory was so clear or evident. Additional studies could examine the role of context memory to further the knowledge of its effects, and to examine under what conditions context memory is important and how. For example, it may be that context memory may only be important as the number of items to be remembered increases, or is differentially affected by numeracy, as it appeared to be in this study.

Study 2

Data and Methods

The second data set comes from the National Survey of Student Engagement (NSSE). The NSSE is an annual survey that has collected data from college students from hundreds of participating institutions, with several thousand surveys collected annually (NSSE 2011).The survey collects data from randomly selected college freshman and seniors at the participating institutions. Although freshman and seniors were the target population, some respondents came from other classes. The primary purpose of the survey is twofold: 1) assessing the time and effort undergraduate degree-seeking students spend on educational activities, and 2) assessing what schools are doing to focus student efforts to these activities (NSSE 2011). Data from these surveys are collected via one of two survey modes, either by a paper or Web survey.

The particular NSSE data for use in this research comes from 2006, when additional data were collected on questions using vague quantifier response options. Twelve questions, regarding active and collaborative learning and student-faculty interaction, were first asked about using vague quantifier response options. Then, at the end of the Web survey (these questions were repeated at the end of the Web survey only)

students were reminded of their earlier answers. The respondents were asked to quantify their response for each question by filling in a number into an open ended response space to indicate the number of times intended by the vague term, and to select the time frame thought of and that the number of times occurred in. For example, a person saying an event occurred “very often” was asked how many times this event occurred and if this number occurred either per day, week, month, academic term, or academic year. Thus, students entered a numeric response and selected an appropriate time frame. Figure 1 displays an example from the Web survey showing the way this question was asked to respondents (from Nelson-Laird et al. 2008).

Given the differences in the rate of occurrence selected by the respondent in the numeric translation of the vague quantifier response, that is, whether the number of times occurred per day, week, month, academic term or academic year (see Figure 1), it is necessary to transform numeric open ended responses to a constant time frame, such as is done in Nelson-Laird et al. (2008). In this case, all numeric answers were transformed to be on a per week time frame, through either multiplying answers by five (for those that said number per day) or dividing by the appropriate number for those saying per month, academic term or year. Specifically, time frames were adjusted by using the following multipliers: day = 5, week = 1, month = 0.231, academic term = 0.067, and academic year = 0.033. The twelve questions asked about in this manner are presented in Appendix 4.

The twelve questions used have been used grouped together in two scales, one for active and collaborative learning and one for student-faculty interaction. Of the twelve items, seven are for the active-collaborative learning scale, with the remaining five

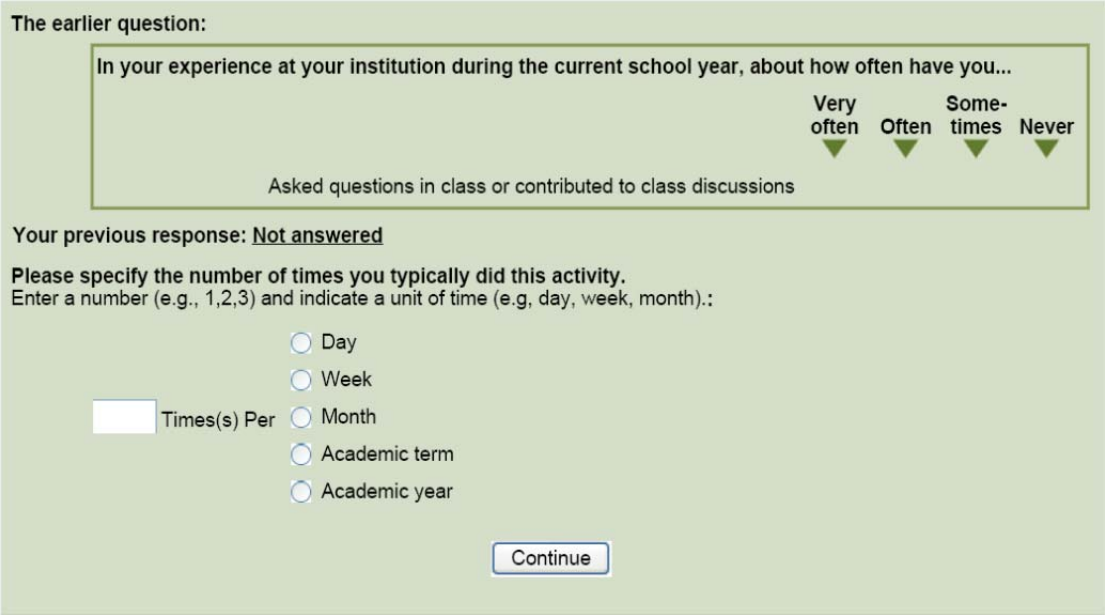

Figure 1. Example question for absolute frequency intended by vague quantifier.

belonging to the student-faculty interaction scale (see Appendix 4). The responses for both the vague quantifier and numeric response translation response options have both been used for the scales (Nelson-Laird et al. 2008). For both response formats, the responses to each question are summated to create the scale value for each respondent. For the numeric translation scale, responses to each question were first transformed to a z-score (by taking the difference from the overall mean and dividing by the standard deviation). The z-scores for each question were then summed up to form the numeric scale.[8] Unlike Study 1, which used translated numeric equivalents, for the vague quantifier scale, scales are summated for the vague quantifier questions by placing values on each of the response options, from 0 through 3 for the options ascending from 0 for

[8] For the numeric translation scale, a simple addition of the numeric responses given was also examined in forming the scales. The results in terms of directionality and significance were near identical to the z-score transformed scales.

“never” to 3 for “very often”. This assignment was used for two reasons; first because others have summed the scale in this way (i.e. Carini et al (2006)) and second, because numeric translation were needed for accuracy assessments but is not needed for predictive validity assessments. Responses are then summed up based on the questions for each of the two scales for each respondent.

Carini et al (2006) examined both the active and collaborative learning and student-faculty interaction for outcomes using the 2002 NSSE. Unlike the data used here, the 2002 NSSE data only had the vague quantifier questions available for scale construction and not the numeric responses as well. Using the vague quantifier version of these scales, Carini et al. (2006) examine the predictive validity of the scales for important theoretically related outcomes. Specifically the authors examine education outcomes, including self-reported grades in college, as “student engagement is generally considered among the better predictors of learning and personal achievement” (Carini et al 2006, p. 2). The results show small but significant positive correlations for both active and collaborative learning and for student-faculty interaction with grades, suggesting the important theoretically related relationship. However, there is no comparison, as it was not possible at the time, to examine the numeric scale format and which response format (numeric or vague) has higher levels of predictive validity.

The authors also note the potential importance of satisfaction with the educational experience as an important educational outcome and construct (Carini et al. 2006). The authors measure satisfaction using two questions from the NSSE. The first asks, “How would you evaluate your entire educational experience at this institution? Poor, Fair, Good, Excellent”. The second satisfaction question asks, “If you could start over again,

would you go to the *same institution* you are now attending? Definitely no, Probably no, Probably yes, Definitely yes" (also available in Appendix 4). Given the important nature of satisfaction as an educational factor, as noted by Carini et al (2006), satisfaction measures were considered outcomes to relate the vague quantifier and numeric response options to the various questions regarding active and collaborative learning and for student-faculty interaction.

The survey also asked about a number of details regarding student engagement as well as student characteristics. Importantly, data about the student's Scholastic Aptitude Test (SAT) is included. Numeracy may potentially be indicated by the student's SAT math score. Initially, this score was divided into quartiles of the distribution, i.e. four levels of numeracy, low to high. However, additional collapsing showed that no differences in results, first by dividing the respondents into three divisions of numeracy, and then dividing students into simple low-high numeracy dichotomies. Given the lack of differences and large number of missing cases on the SAT math score variable (discussed below), the dichotomy of high-low numeracy was chosen to examine differences on this variable, using the split at the median level. .

The data available for this survey is a twenty percent random subsample of the respondents, due to licensing restrictions. The data comes from the available student respondents to the Web survey, which represent 26,204 first-year and 36,263 senior students who were randomly selected from 149 institutions (Nelson-Laird et al. 2008). A twenty percent random subsample leads to a sample size of 10,767. Analyses of these respondents' answers focus on the logical consistency between vague quantifiers and the associated numeric open ended responses and on the predictive validity of the different

responses, in regards to variables relating to the academic outcomes and perception of education.

It is important to note that the data from the NSSE may be similar to a cluster sample for surveys (Kish 1965). This clustering would arise from the fact that respondents are selected within participating institutions (the clusters). Responses are therefore not completely independent. Lack of independence of response could occur because the similar environment or the underlying causes in selecting the same university leads to more similar responses than would be expected if the response were independent (Kish 1965). This potentially leads to more homogenous responses than if a simple random sample (SRS) was taken in conducting the same survey. Hence, each observation provides less information than an observation in a simple random sample.

As such, estimating the variance of point estimates using SRS assumptions and techniques may be incorrect (Kish 1965, Lohr 2010). Special techniques should be used that incorporate the clustering effects. In this case, that would entail controlling for the potential clustering of responses within universities. Failure to do so would lead to incorrect variance estimates; however, point estimates remain unbiased. Using the incorrect variance would lead to incorrect hypothesis testing and estimation of confidence intervals, which are based on the standard error resulting from the biased variance estimates using SRS assumptions.

In order to do so, it is first necessary to examine whether there is indeed a clustering effect or not. To test for clustering effects, it is necessary to calculate the design effect (Kish 1965). The design effect (*deff*) is an inflationary factor of the variance for the cluster sample, where the cluster variance is inflated over that the variance

assuming SRS by the formula *variance(cluster) = variance(SRS)*deff.* If the design effects are large, then appropriate methods must be used to compensate for the clustered design of the data. Specifically, it is necessary to use an alternative variance estimation process, such as Taylor series approximation (Kish 1965), which is employed here.

Hypotheses

Given that Studies 2 and 3 both are similar, with Study 2 examining frequency estimation and Study 3 examining subjective probabilities, hypotheses 1 through 4 are the same for both, with the remaining hypotheses being study specific.

H1: In general, vague quantifiers will show higher levels of predictive validity with theoretically related variables compared to numeric responses.

H1a: There will be higher correlations and better model fit for variables measured by vague quantifiers compared to numeric open ended responses.

H2: When asked for a direct numeric translation of vague quantifiers, there should be a logical consistency between numeric and vague quantifier measures.

H3: In general, more numerate individuals will provide responses showing greater predictive validity.

H4: More numerate individuals will show greater logical consistency in numeric open ended and vague quantifier responses than less numerate individuals.

H5: More numerate individuals will show greater predictive validity using numeric open ended responses than less numerate individuals.

H6: However, more numerate individuals will not show greater predictive validity using numeric open ended responses than when using vague quantifiers.

Results

Data Management

After transforming the numeric data to weekly rates, the distribution of these numeric translations were examined. Visual inspection showed that some responses were extreme and not plausible (e.g. an event occurring 50000 times a week). Overall, these extreme responses were few. The data were cut at the 99th percentile of the distribution, which would in all cases lead to more reasonable responses with a minimum of data cut. For nine of the numeric translations, the use of the 99th percentile led to cuts of translations greater than 10. For frequency of questions asked in class, the 99th percentile was 50. It was 15 for working in class with other students. For discussed ideas outside of class with others, the 99th percentile was 20. These cut data are used in all following analyses. Further, in order that the vague quantifier scales be anchored at zero, all of the vague quantifier responses were scaled from zero (“never”) to three (“very often”).

A potential problem of data missingness arises when attempting to divide the data along numeracy measures, i.e. SAT math scores. In particular, more than half the data has missing values on this variable, reducing the overall n to 4761 when using numeracy as a measure. Still, this provides a reasonably large sample, but results based on numeracy must be made with this caveat. Levels of numeracy were divided into high- and low-based, divided along the median of 570 (mean = 565.55). Respondents with SAT math scores higher than 570 are considered to have high numeracy and those with lower than

or at 570 are considered to have low numeracy. This leads to 2211 respondents in the high numeracy division and 2250 respondents in the low numeracy division.[9]

The active-collaborative learning scale and student-faculty interaction scale were calculated for both vague quantifier and z-scored numeric responses. These calculations were done by summing the responses for the vague quantifier scale (with values of 0 to 3 for each response) and separately summing the numeric responses (with values ranging from 0 to 10 for nine questions, 0 to15 for one question, 0 to 20 for one question, and 0 to 50 for one question, as discussed above). Means for each scale and the standard error (accounting for clustering, see below) are presented in Table 2.1. Means are given for the total sample, as well as for high and low numeracy divisions.

Table 2.1

Active-Collaborative Learning Scale (ACLS) and Student –Faculty Interaction Scale (SFIS) Means

	ACLS – Vauge (s.e.)	ACLS – Numeric (s.e.)	SFIS – Vauge (s.e.)	SFIS Numeric (s.e.)
Low Numeracy	9.82 (0.10)	16.76 (0.44)	6.71 (0.10)	3.51 (0.09)
High Numeracy	9.91 (0.15)	17.96 (0.50)	6.56 (0.14)	3.54 (0.13)
Total Sample	9.96 (0.08)	16.75 (0.28)	6.58 (0.07)	3.46 (0.06)

[9] The sample was also divided into four and then three levels of numeracy to test for sensitivity to the divisions. The divisions were made such that first a quarter of the sample and then near 33% of the sample belonged to each division. Dividing the sample in these ways changed none of the results and thus a simpler division of two groups was used, for parsimony and for increased sample sizes, given the large number of missing cases.

Effects of Clustering

Given that there may be a clustering effect in the data, since respondents are selected and grouped within the 127 universities included in the survey, it is important to check for possible clustering effects in the data. This check is done by estimating the design effect (*deff*) for each of the main variables used in this study. The design effect is an inflationary factor that shows the ratio between the estimated variance controlling for the complex survey design (i.e. clustering) and the variance calculated assuming simple random sampling (SRS), which is the default variance estimation in most analyses. Larger design effects suggest a greater increase in the estimated variance due to the complex survey design (Kish 1965). However, clustering does not affect calculation of point estimates, such as means, correlation coefficients and regression coefficients, only the variance for these estimates.

The design effect is statistic specific, and may be different for different variables and statistics (Lohr 2010). To examine whether clustering is a possible issue that needs to be accounted for in the NSSE data, design effects were calculated for the means of the variables of interests. This includes the 12 vague quantifier responses to the two scales, the 12 numeric translations for each of these responses, and the three related variables of interest, grades and the two satisfaction questions, for a total of 27 design effects estimated. Design effects were estimated based on the Taylor series approximation variance estimates for the clustered survey design (Kish 1965).

The results show that, overall, clustering is an issue that must be accounted for in the NSSE. The calculated design effects ranged from 1.104 to 9.267. Only three of the 27 estimated *deff* were less than 2, with the mean *deff* being 3.888. This mean *deff* suggests,

on average, a near four times increase in the estimated variance due to the clustered design compared to the simple random sampling assumptions frequently used in analyses. Given the evident clustering effects on the variance estimation, it would be inappropriate to use simple random sampling assumptions in variance estimation, and hence hypothesis testing, for the remaining analyses. Therefore, appropriate estimation procedures are employed using the SAS system (SAS 2010). Specifically, for means, PROC SURVEYMEANS will be used, and for regressions, PROC SURVEYLOGISTIC will be used. Variances will all be estimated using the Taylor series approximation.

Logical Consistency

To examine the relationship between vague quantifier responses and the numeric translations, means translation at each level of vague quantifier response was examined for each of the twelve question pairs. Assuming that the vague quantifier and the numeric translations estimates are measuring the same construct as intended, there should be a level of concordance between them. This analysis is in essence similar to prior research examining the numeric translations to responses to vague quantifier scales (Bradburn and Miles 1979, Nelson-Laird et al. 2006). Tables 2.2–2.13 present the means for the numeric translations at each level of the responses to the vague quantifier scale. These means were calculated for the full sample and both divisions of numeracy, with the noted caveat that this leads to a large reduction in the number of cases due to the missingness in SAT sccores, although more than 2000 cases remain for each division.

Table 2.2

Number of Times Asked Question in Class by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	1.47[a] (n = 58)	1.96[a] (n = 39)	1.52[a] (n = 240)
Sometimes	3.99[b] (n = 763)	3.75[b] (n = 631)	3.66[b] (n = 3113)
Often	9.77[c] (n = 865)	10.47[c] (n = 722)	9.56[c] (n = 3620)
Very Often	17.40[d] (n = 756)	18.47[d] (n = 754)	16.87[d] (n = 3405)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.3

Number of Times Made a Class Presentation by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.16[a] (n = 160)	0.12[a] (n = 170)	0.16[a] (n = 755)
Sometimes	0.39[b] (n = 1015)	0.33[b] (n = 1027)	0.38[b] (n = 4287)
Often	0.72[c] (n = 817)	0.57[c] (n = 711)	0.65[c] (n = 3448)
Very Often	0.99[d] (n = 407)	0.74[d] (n = 233)	0.93[d] (n = 1748)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.4

Number of Times Worked With Other Students During Class by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.23^{a} (n = 245)	0.13^{a} (n = 267)	0.21^{a} (n = 1011)
Sometimes	0.94^{b} (n = 1129)	0.80^{b} (n = 1109)	0.88^{b} (n = 4783)
Often	1.99^{c} (n = 722)	1.87^{c} (n = 562)	1.90^{c} (n = 3103)
Very Often	2.52^{d} (n = 280)	3.15^{d} (n = 179)	2.94^{d} (n = 1251)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.5

Number of Times Worked With Other Students Outside Class by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.14^{a} (n = 177)	0.17^{a} (n = 118)	0.13^{a} (n = 789)
Sometimes	0.73^{b} (n = 963)	0.65^{b} (n = 836)	0.68^{b} (n = 3959)
Often	1.54^{c} (n = 800)	1.66^{c} (n = 716)	1.56^{c} (n = 3103)
Very Often	2.29^{d} (n = 407)	3.01^{d} (n = 420)	2.65^{d} (n = 1912)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.6

Number of Times Tutored or Taught Other Students by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.04^{a} (n = 1144)	0.02^{a} (n = 791)	0.04^{a} (n = 4516)
Sometimes	0.72^{b} (n = 823)	0.66^{b} (n = 781)	0.68^{b} (n = 3580)
Often	1.64^{c} (n = 234)	1.91^{c} (n = 316)	1.85^{c} (n = 1172)
Very Often	2.92^{d} (n = 136)	3.12^{d} (n = 189)	3.25^{d} (n = 721)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.7

Number of Times Participated in a Community-Based Project by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.06^{a} (n = 1231)	0.04^{a} (n = 1274)	0.05^{a} (n = 5633)
Sometimes	0.39^{b} (n = 713)	0.34^{b} (n = 542)	0.38^{b} (n = 2852)
Often	1.05^{c} (n = 252)	1.11^{c} (n = 180)	0.99^{c} (n = 1014)
Very Often	1.46^{c} (n = 132)	1.64^{c} (n = 93)	1.60^{d} (n = 489)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.8

Number of Times Discussed Ideas from Classes with Others Outside of Class by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.20[a] (n = 127)	0.28[a] (n = 87)	0.20[a] (n = 471)
Sometimes	1.37[b] (n = 800)	1.08[b] (n = 635)	1.23[b] (n = 3276)
Often	3.13[c] (n = 796)	3.22[c] (n = 794)	3.08[c] (n = 3543)
Very Often	5.05[d] (n = 488)	5.64[d] (n = 475)	5.29[d] (n = 2206)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.9

Number of Times Discussed Grades or Assignments with an Instructor by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.06[a] (n = 127)	0.12[a] (n = 139)	0.10[a] (n = 547)
Sometimes	0.40[b] (n = 870)	0.37[b] (n = 905)	0.37[b] (n = 3912)
Often	0.89[c] (n = 758)	0.96[c] (n = 644)	0.90[c] (n = 3264)
Very Often	1.25[d] (n = 552)	1.34[d] (n = 389)	1.36[d] (n = 2180)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.10

Number of Times Talked About Career Plans with a Faculty Member or Advisor by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.03[a] (n = 339)	0.04[a] (n = 389)	0.04[a] (n = 1775)
Sometimes	0.25[b] (n = 1008)	0.23[b] (n = 955)	0.22[b] (n = 4382)
Often	0.54[c] (n = 548)	0.56[c] (n = 447)	0.56[c] (n = 2335)
Very Often	1.01[d] (n = 351)	1.13[d] (n = 281)	0.99[d] (n = 1392)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.11

Number of Times Received Prompt Feedback from Faculty on Academic Performance by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.08[a] (n = 89)	0.10[a] (n = 78)	0.12[a] (n = 432)
Sometimes	0.66[b] (n = 740)	0.64[b] (n = 671)	0.64[b] (n = 3223)
Often	1.45[c] (n = 1011)	1.51[c] (n = 947)	1.47[c] (n = 4351)
Very Often	1.99[d] (n = 414)	2.19[d] (n = 338)	2.07[d] (n = 1667)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.12

Number of Times Worked with Faculty Members on Activities Other Than Coursework by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.04[a] (n = 1089)	0.02[a] (n = 865)	0.04[a] (n = 4733)
Sometimes	0.44[b] (n = 683)	0.42[b] (n = 715)	0.39[b] (n = 3029)
Often	1.08[c] (n = 303)	1.21[c] (n = 283)	1.13[c] (n = 1230)
Very Often	1.72[d] (n = 176)	2.26[d] (n = 165)	1.98[d] (n = 668)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Table 2.13

Number of Times Discussed Ideas from Classes with Faculty Members Outside of Class by Vague Quantifier Response

Vague Quantifier Response	Low Numeracy	High Numeracy	Total Sample
Never	0.13[a] (n = 761)	0.10[a] (n = 634)	0.13[a] (n = 3209)
Sometimes	0.73[b] (n = 4345)	0.63[b] (n = 984)	0.67[b] (n = 4345)
Often	1.44[c] (n = 378)	1.59[c] (n = 290)	1.53[c] (n = 1473)
Very Often	2.22[d] (n = 197)	2.23[d] (n = 139)	2.21[d] (n = 738)

Means are numeric translations of vague quantifier response. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

Several findings are evident based on the data presented in these tables. First, vague quantifiers evidently mean different things in response to different questions and topics. For example, using the total sample, the translated meaning for the response of "very often" ranges from a mean of 0.93 times a week for number of times a presentation was made in class to a mean of 16.87 times a week for asked a question in class. Similarly, the translation of "never" ranged from a mean 0.04 times a week talking with faculty members about career plans to a mean 1.52 times a week for asked a question in class. These differences in meaning are found across all divisions of the sample. This finding is consistent with other studies, which have also found different numeric meanings for vague quantifiers across different behaviors (Bradburn and Miles 1979).

The second important finding is that, overall, there appears to be a great deal of logical consistency, in the total sample and in both the high and low numeracy divisions. For all questions and across all divisions of the sample, the mean numeric translations for each level of the vague quantifier are ordered in a consistent and expected manner. That is the translation for "never" is the smallest number of times per week, followed by "sometimes", "often" and "very often", in that order. For nearly all questions and divisions of the sample, the mean translations are all significantly different from one another, further strengthening the finding of logical consistency.

Only two instances of means were not significantly different from one another. For both the low numeracy and high numeracy divisions, the mean translations for "often" and "very often" are not significantly different from one another for the number of times participated in a community-based project. This lack of significance is likely partly due to small numbers of respondents giving each of these responses, but other

questions have similar distribution of respondents with significant differences still found. Further, the difference just barely fails to reach statistical significance at the $p<0.05$ level; at $p<.10$ the means are significantly different. In both cases, the mean differences are in the expected direction, with “very often” larger than “often”. Overall, these two instances do not contradict the pattern of logical consistency between vague quantifier and numeric responses.

The only possible exception to the argument of logical consistency between the two measures is that for all questions, the mean translation of never is significantly different from zero at the $p<0.05$ level. Although in some cases, the mean approaches zero, being as low as 0.04 in several instances for the total sample and those in the low numeracy division and as low as 0.02 for the high numeracy division in a couple of instances, these are still significantly different from zero. Further, for all three divisions, never means more than one for number of times asked a question in class (total sample = 1.52; low numeracy = 1.47; high numeracy = 1.96).

However, this result is countered by the fact that for most of the questions, the large majority of respondents who said “never” to the question gave a numeric translation of zero. The only exception is for asked a question in class, where still the plurality (33.75% in the total sample, 36.21% in the low numeracy division, 38.46% in the high numeracy division) gave zero as the numeric translation for “never”. For the remainder of the questions, the percentage giving zero as the translation of “never” ranged from 68.45-93.49% for the total sample, 68.93-91.87% for the low numeracy division, and 72.28-95.20% for the high numeracy division. This finding of large majorities for most questions translating “never” as zero is additional support that overall, responses to vague

quantifiers and numeric responses are logically consistent with one another, at least for this population and set of behaviors.

These findings relate directly to two of the proposed hypotheses, one of which receives support and one that does not. First, with the direct translation between measures, there appears to be strong logical consistency, which lends support to Hypothesis 2. However, there is no apparent difference in this high level of logical consistency between measures when examining different numeracy levels, which is the expectation based on Hypothesis 4.

Predictive Validity

Predictive validity has been shown to be an important aspect of the measurement properties of frequency questions in past research (Chang and Krosnick 2003, Lu et al. 2008). This type of validity can be measured through the relationship between the measures of interest and theoretically related variables. In this case, the measures of interest are the active-collaborative learning and student-faculty interaction scales, which are summations of the twelve questions presented above (Carini et al. 2006, Nelson-Laird 2008). Seven are used in the active-collaborative scale and five in the student-faculty interaction scale (see Appendix 4). Since the twelve questions are asked using both vague quantifier scales and z-scored numeric translations, four total scales are available for use; two for vague quantifier scales and two for z-scored numeric translations.

The theoretically related variables, as noted in Carini et al. (2006), are grades, measured on an eight-point scale, and two satisfaction questions using four point scales (see Appendix 4). Given the ranked ordering of the data, Spearman's rho is the appropriate correlation coefficient to employ. The correlations of the two vague

quantifier and two numeric scales with these three measures are presented in Tables 2.14-2.19, with the correlations of the theoretically related variables and the active-collaborative learning scales presented first, followed by those with the student-faculty interaction scales. Along with the correlations for the total sample, the correlations found in the high and low numeracy divisions are also presented. In addition, correlations were transformed to Fisher z-scores in order to examine significant differences. Due to the clustering effects, this may not be the most appropriate manner for testing for significant differences; however, no other method is currently known, and this test gives a possible indication of significant differences. In any case, the point estimates for the correlation coefficients are not affected by the clustering, and are interpretable in terms of size and direction.

The results presented in these tables point in a single direction, being that vague quantifier responses have higher levels of predictive validity than when using numeric open ended responses. In every case except one, for every theoretically related variable for each of the divisions of the sample (high and low numeracy, total sample) the point estimates for the correlations are larger for vague quantifier responses than for numeric responses. The one exception occurred in the high numeracy division for the ACLS correlation with same college preference, where the numeric scale was slightly larger at the third place after the decimal, 0.164 to 0.163. Further, for the total sample, every one of these differences are statistically significant at the $p<.05$ level using the Fisher z-score transformation. Although the correlations as whole are not large, the importance follows from the comparison between the two different measures. In addition, similarly small correlations were found in Carini et al (2006), so these small correlations are expected.

Table 2.14

Correlation of Active-Collaborative Learning Scales with Grades

	Low Numeracy	High Numeracy	Total Sample
ACLS – Vague	0.18[a]	0.14[a]	0.18[a]
ACLS – Numeric	0.12[a]	0.06[b]	0.10[b]

ACLS = Active-Collaborative Learning Scale
Different superscript letters indicate significant differences of correlations within columns at $p<0.05$ level.

Table 2.15

Correlation of Active-Collaborative Learning Scales with College Experience Rating

	Low Numeracy	High Numeracy	Total Sample
ACLS – Vague	0.26[a]	0.23[a]	0.23[a]
ACLS – Numeric	0.15[b]	0.17[b]	0.17[b]

ACLS = Active-Collaborative Learning Scale
Different superscript letters indicate significant differences of correlations within columns at $p<0.05$ level.

Table 2.16

Correlation of Active-Collaborative Learning Scales with Same College Preference

	Low Numeracy	High Numeracy	Total Sample
ACLS – Vague	0.19[a]	0.16[a]	0.16[a]
ACLS – Numeric	0.15[a]	0.16[a]	0.13[b]

ACLS = Active-Collaborative Learning Scale
Different superscript letters indicate significant differences of correlations within columns at $p<0.05$ level.

Table 2.17

Correlation of Student-Faculty Interaction Scales with Grades

	Low Numeracy	High Numeracy	Total Sample
SFIS – Vague	0.16[a]	0.15[a]	0.14[a]
SFIS – Numeric	0.11[a]	0.11[a]	0.10[b]

SFIS = Student Faculty Interaction Scale
Different superscript letters indicate significant differences of correlations within columns at $p < 0.05$ level.

Table 2.18

Correlation of Student-Faculty Interaction Scales with College Experience Rating

	Low Numeracy	High Numeracy	Total Sample
SFIS – Vague	0.28[a]	0.29[a]	0.27[a]
SFIS – Numeric	0.18[b]	0.19[b]	0.19[b]

SFIS = Student Faculty Interaction Scale
Different superscript letters indicate significant differences of correlations within columns at $p < 0.05$ level.

Table 2.19

Correlation of Student-Faculty Interaction with Same College Preference

	Low Numeracy	High Numeracy	Total Sample
SFIS – Vague	0.19[a]	0.21[a]	0.19[a]
SFIS – Numeric	0.17[a]	0.16[a]	0.15[b]

SFIS = Student Faculty Interaction Scale
Different superscript letters indicate significant differences of correlations within columns at $p < 0.05$ level.

In regards to the divisions of high and low numeracy, the pattern is similar to that of the total sample, with the correlation estimates, while overall being relatively small, all being higher in both divisions for vague quantifier responses in every comparison. However, not every difference is significant at the $p<0.05$ level using the Fisher z-score transformation. This lack of significance occurred between the active-collaborative learning scales correlations with grades and student-faculty interaction scales correlations with grades and same college preference for the low numeracy division. For the high numeracy division, there were no significant differences for active-collaborative learning scales correlations with same college preference and student-faculty interaction scales correlations with grades and same college preference for the low numeracy division.

Beyond these findings, there are no discernible differences between high and low numeracy divisions. In some cases, the correlations are higher among those with high numeracy; in others the correlations are higher among those with lower numeracy. The lack of difference between the two divisions holds when looking at either scale difference. In only one instance was there a significant difference across divisions, with the correlation for the active collaborative learning scale using numeric response being higher in the low numeracy division than the higher numeracy division. It appears that numeracy, like for logical consistency, did not generally have an impact on predictive validity, at least using correlation coefficients as a measure.

More support is evident for the proposed hypotheses on predictive validity than is evident for the tests for logical consistency. First, Hypothesis 1 (and 1a) was clearly supported. Scales using vague quantifiers showed higher levels of predictive validity, with higher correlations in every case and every division between all theoretically related

variables and vague quantifier scales compared to scales using numeric responses. Similarly, Hypothesis 6 is supported by the correlations in the above tables. Like the total sample and low numeracy division, high numeracy individuals also showed higher levels of predictive validity when using vague quantifier responses than when using numeric responses.

However, some hypotheses were not supported. Those with high numeracy did not show higher levels of predictive validity compared to those with lower numeracy, counter to Hypothesis 3. In fact, the one case where there was a significant difference between the two divisions, there was a higher correlation among the lower numeracy division. Hypothesis 5 is also not supported, as the higher numeracy division did not have higher correlations when using numeric responses than those in the lower numeracy division. Again, the one instance of a significant difference was in the opposite direction of the prediction, with the lower numeracy division having a higher correlation using numeric responses. The overall picture shows little difference between high and low numeracy divisions, contrary to these two hypotheses.

Another way to test the predictive validity of the vague quantifier and numeric responses scales is to use regressions predicting the theoretically related variables, while including important control variables. Since the purpose is to compare models in regards to which scale best predicts the theoretically related variables in order to further assess the predictive validity, not all possible control variables predicting these outcomes are required. The control variables selected, based on those used in Nelson-Laird et al. (2008), are class standing (i.e. freshman, senior, or other), gender, full-time attendance status, and age (categorized as 19 and younger, 20-23, 24-29 and 30 and older).

The purpose of these regression models is to identify which scales increase model fit and predictive capability. Since all of the outcome variables (grades and two satisfaction questions) are ordinal-level variables, ordered logistic regressions were used to take into account the clustering effect of respondents within universities. Therefore, separate models for each of the scales predicting each of the three related variables are compared using the criterion of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Agresti 2002). Since the AIC and BIC are related to the sample size, all models are restricted to include only cases where data is available for all risk measures to ensure comparability. Another useful indicator for model fit, the area under the Receiver Operator Characteristic (ROC) curve, a measure of diagnostic accuracy, could not be estimated since this requires binary outcomes. A base model was also estimated for each of the three outcome variables, including only the demographic variables to show improvement in the models based on the inclusion of the various scales. The AIC and BIC for each of the three outcome variables, for each of the three divisions of the sample (low numeracy, high numeracy, and total sample) are presented in Tables 2.20-2.25.

The results mirror those of the correlation analyses, with the models employing the vague quantifier versions of the scales performing better than either the base model or the model using the numeric versions of the scales, as indicated by the lower AIC and BIC scores. These findings hold true in all cases for both the active-collaborative learning scale and for the student-faculty learning scale and when examining the AIC and BIC. The models are restricted to the same sample and include all of the same control variables, only differing in which version of the scale (numeric or vague quantifier

version) are used. Therefore, the differences in the AIC and BIC can be solely attributed to the differential predictive validity for each version of the scales. As such, it appears that vague quantifiers have higher levels of predictive validity than numeric scales. This finding holds across all levels of numeracy, as those with both higher and lower levels of numeracy show higher levels of predictive validity using vague quantifier scales than when using numeric scales. Again, this is true for all outcome variables, for both the active-collaborative learning scale and for the student-faculty learning scale, and when using either the AIC or BIC.

Table 2.20

Logistic Regression Indicators of ACLS on Grades

	Low Numeracy (n = 2133)		High Numeracy (n = 1915)		Total Sample (n = 9119)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	7840.73	7920.04	6532.96	6610.76	32665.89	32765.54
ACLS – Vague	7788.05	7873.03	6510.69	6594.05	32432.30	32539.07
ACLS – Numeric	7827.24	7912.22	6531.06	6614.42	32623.00	32729.78

Table 2.21

Logistic Regression Indicators of ACLS on College Experience Rating

	Low Numeracy (n = 2136)		High Numeracy (n = 1915)		Total Sample (n = 9124)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	4318.85	4375.52	3795.91	3851.48	18402.35	18473.53
ACLS – Vague	4191.79	4254.13	3662.75	3723.88	17836.73	17915.04
ACLS – Numeric	4286.22	4348.56	3772.62	3833.75	18244.01	18322.32

Table 2.22

Logistic Regression Indicators of ACLS on Same College Preference

	Low Numeracy (n = 2136)		High Numeracy (n = 1914)		Total Sample (n = 9123)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	5026.31	5082.98	4319.69	4375.26	20689.32	20760.51
ACLS – Vague	4953.59	5015.92	4243.79	4304.91	20404.25	20483.25
ACLS – Numeric	4992.60	5054.93	4294.22	4355.35	20590.64	20668.95

Table 2.23

Logistic Regression Indicators of SFIS on Grades

	Low Numeracy (n = 2199)		High Numeracy (n = 1986)		Total Sample (n = 9439)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	8092.80	8172.54	6751.94	6830.26	33711.35	33811.48
SFIS – Vague	8052.60	8136.04	6720.71	6804.62	33548.39	33655.68
SFIS – Numeric	8068.59	8154.03	6744.54	6828.45	33664.79	33722.08

Table 2.24

Logistic Regression Indicators of SFIS on College Experience Rating

	Low Numeracy (n = 2203)		High Numeracy (n = 1987)		Total Sample (n = 9474)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	4453.07	4510.05	3956.48	4012.43	19056.29	19127.82
SFIS – Vague	4295.67	4415.15	3761.60	3823.14	18321.07	18399.76
SFIS – Numeric	4417.89	4480.56	3929.97	3991.50	18888.83	18967.51

Table 2.25

Logistic Regression Indicators of SFIS on Same College Preference

	Low Numeracy (n = 2203)		High Numeracy (n = 1986)		Total Sample (n = 9474)	
	AIC	BIC	AIC	BIC	AIC	BIC
Base	5166.31	5223.29	4491.82	4547.76	21380.83	21452.37
SFIS – Vague	5092.11	5154.79	4377.42	4438.96	20988.55	21067.23
SFIS – Numeric	5132.47	5195.14	4461.42	4522.96	21260.15	21338.72

Like the correlation analyses, these results also lend support to some of the hypotheses, while not supporting others. First, these results support Hypothesis 1 (and 1a) in that vague quantifier scales show higher level of predictive validity than numeric scales. Given that this higher level of predictive validity is found across all divisions of the sample, for all models using the different scales, for all measures of model fit, this hypothesis is strongly supported. In conjunction with the correlations above, it is suggested that this hypothesis may be confirmed.

In addition, Hypothesis 6 is also supported, as it is in the correlation analyses. Specifically, even among those with higher levels of numeracy, vague quantifier scales outperformed its numeric counterparts in terms of predictive validity. Like with the total sample and those with lower numeracy, those with higher numeracy show higher levels of predictive validity using vague quantifier scales than with numeric scales for all models, scales, and criteria.

Hypotheses 3 and 5 are not directly testable using the AIC and BIC. The AIC and BIC are sample specific, and therefore only comparable within, not across, samples. Still,

the results are suggestive. Similar patterns emerge for both numeracy divisions across all models and scales, for both the AIC and BIC. This similarity is suggestive of similar predictive validity for each numeracy division. This is bolstered by the findings of the correlation analysis which more directly show no differences between the two divisions in terms of predictive validity. In combination, this further strengthens the claim that Hypotheses 3 and 5 are not supported, i.e. that there are no differences between high and low numeracy divisions in terms of predictive validity.

Finally, the estimated coefficients and odds ratios for the different scales in predicting the three dependent variables controlling for demographics are presented in Tables 2.26-2.31. The standard errors reported reflect the clustered sample design using the Taylor series approximation. As in other tables, the divisions for low and high numeracy are presented along with that of the total sample.

All of the coefficients are significant, and all coefficients are in the expected direction, with increases in the scale being associated with increases in the three dependent variables. That is, increases in both the active-collaborative learning scale and student-faculty interaction scale, regardless of scale type (vague or numeric), leads to increases in the predicted grades, satisfaction with college experience and the stated preference for the same college.

Table 2.26

Scale Coefficients and Odds Ratios of ACLS on Grades

	Low Numeracy (n = 2133)		High Numeracy (n = 1915)		Total Sample (n = 9119)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
ACLS - Vague	0.085* (0.011)	1.089	0.064* (0.017)	1.066	0.087* (0.007)	1.091
ACLS – Numeric	0.045* (0.011)	1.045	0.026* (0.016)	1.026	0.037* (0.006)	1.038

*p < 0.05

Table 2.27

Scale Coefficients and Odds Ratios of ACLS on College Experience Rating

	Low Numeracy (n = 2136)		High Numeracy (n = 1915)		Total Sample (n = 9124)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
ACLS - Vague	0.148* (0.013)	1.159	0.167* (0.014)	1.181	0.151* (0.006)	1.163
ACLS – Numeric	0.073* (0.010)	1.075	0.072* (0.013)	1.074	0.078* (0.007)	1.081

*p < 0.05

Table 2.28

Scale Coefficients and Odds Ratios of ACLS on Same College Preference

	Low Numeracy (n = 2136)		High Numeracy (n = 1914)		Total Sample (n = 9123)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
ACLS - Vague	0.107* (0.011)	1.113	0.121* (0.014)	1.129	0.103* (0.007)	1.109
ACLS – Numeric	0.071* (0.012)	1.073	0.072* (0.014)	1.017	0.060* (0.007)	1.061

*p < 0.05

Table 2.29

Scale Coefficients and Odds Ratios of SFIS on Grades

	Low Numeracy (n = 2199)		High Numeracy (n = 1986)		Total Sample (n = 9439)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
SFIS - Vague	0.079* (0.011)	1.082	0.077* (0.014)	1.080	0.077* (0.006)	1.080
SFIS – Numeric	0.064* (0.014)	1.066	0.039* (0.013)	1.040	0.041* (0.006)	1.042

*p < 0.05

Table 2.30

Scale Coefficients and Odds Ratios of SFIS on College Experience Rating

	Low Numeracy (n = 2203)		High Numeracy (n = 1987)		Total Sample (n = 9474)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
SFIS - Vague	0.173* (0.014)	1.189	0.216* (0.014)	1.241	0.186* (0.008)	1.224
SFIS – Numeric	0.081* (0.017)	1.085	0.078* (0.018)	1.081	0.075* (0.007)	1.094

*p < 0.05

Table 2.31

Scale Coefficients and Odds Ratios of SFIS on Same College Preference

	Low Numeracy (n = 2203)		High Numeracy (n = 1986)		Total Sample (n = 9474)	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio	Coefficient	Odds Ratio
SFIS - Vague	0.114* (0.014)	1.121	0.158* (0.013)	1.171	0.131* (0.008)	1.158
SFIS – Numeric	0.079* (0.015)	1.082	0.082* (0.018)	1.086	0.087* (0.009)	1.091

*p < 0.05

In addition, all of the coefficients and odds ratios are larger for the vague quantifier versions of the scales than the numeric versions in every comparison. The larger estimates are also reflected in larger standardized coefficients (not shown) for the vague quantifier version of the scales in every comparison. Taken together, this difference in effect size suggests that changes along the vague quantifier scales have more influence in the predicted outcomes of grades, satisfaction with college experience,

and the stated preference for the same college than do numeric versions of the same scales. This finding further supports the position that vague quantifiers display higher levels of predictive validity than numeric responses. However, there are no discernible differences across division in the sample suggestive of differential effects of numeracy.

Discussions and conclusions

This study examined vague quantifier and numeric open-ended responses in terms of the logical consistency between the two responses and in terms of the comparative predictive validity of the two measures. Overall, the results show that respondents were logically consistent between vague quantifier and numeric open ended responses. Values for numeric responses increased along increases in the vague quantifier scale, with each scale point producing statistically significant larger numeric translations.

Further, although none of the means for “never” on the vague quantifier scale equaled zero, the plurality (and frequently majority) of respondents translated “never” to mean zero in every case as well, further indicating strong logical consistency, being consistent semantically as well as in ordering. There were no important differences between numeracy divisions, suggesting that in this case, numeracy did not have an impact on logical consistency.

This study also examined the comparative predictive validity, which is important in determining which measure has better measurement properties (Chang and Krosnick 2003). No study to date has examined the comparative predictive validity of these two types of measures for frequency estimation. The closest to examining this was Lu et al (2008) which compared vague quantifier and numeric scales, finding little differences between the two. Unlike that study, this research finds significant differences between

response formats. Specifically, scales using vague quantifier responses displayed higher levels of predictive validity than those using numeric open ended responses.

This higher level of predictive validity held regardless of which scale was being inspected (i.e. active-collaborative learning scale or student-faculty interaction scale). It also did not matter which outcome variable (i.e. grades, satisfaction with college experience, stated preference to for the same college) these scales were correlated with or predicting. Numeracy level also had no impact; for both low and high numeracy divisions, vague quantifier versions of the scales also performed better in every instance. Again, this suggests, that at least in this sample, numeracy does not impact whether vague quantifier or numeric responses perform better in terms of predictive validity. For all divisions of the sample, it appears that vague quantifiers have better measurement properties.

As indicated by the findings of no differences between levels of numeracy, several of the hypotheses were not confirmed. Specifically those hypotheses regarding differential effects for logical consistency and predictive validity across numeracy levels were not supported. However, the most important hypothesis overall was supported. Using different measures, such as correlations and regressions, the results suggest that the vague quantifier measure outperforms numeric open ended measures in terms of predictive validity for frequency estimation (Hypotheses 1 and 1a). The higher predictive validity is also indicative of potentially lower measurement error for the vague quantifier scale. Also, there was the expected logical consistency between vague quantifiers and the direct numeric translations (Hypothesis 2). Based on these results, it is suggested that

greater use of vague quantifiers are used in frequency estimation surveys, especially if the goal is to increase the predictive validity of responses.

Another important finding of this study, although not necessarily related to the use of vague quantifier of numeric responses, is the clustering effects found. There were generally large clustering effects (design effects) found for all of the twenty-seven questions of main interest used in this study. Overall, the mean design effect approached 4 (mean = 3.888), indicating a near four-fold increase in variance due to clustering compared to the usual simple random sampling assumptions. To find clustering was somewhat unexpected, for several reasons. First, the data received is a 20 percent random subsample of the entire data set, which was expected to reduce any observed clustering. Second, a university is not necessarily a homogenous place. Although it may be more homogenous than an entire city, it is generally expected that there is diversity in many colleges, which should limit the possible clustering.

Finally, and most unexpectedly, is that the survey is conducted via the web. Although clustering effects can occur in web surveys, clustering effects are most frequently thought of occurring in face to face surveys (Lohr 2010). The expectation for greater clustering in face to face surveys is in part due to the effect of the interviewer, which is absent in web surveys. The finding of this study of large design effects in a web survey indicates the importance of examining for possible clustering in surveys regardless of the mode used or the presence or absence of interviewers. This finding also suggests that other analyses that have used the NSSE data that have not used analyses to account for clustering may have found results that are not actually supported by the data, if correct analyses were employed.

Although this study took appropriate steps to estimate variances correctly, there are some possible limitations to this research. First is the measure used for numeracy, SAT Math scores. This measure had a large number of missing cases, dropping the overall number of respondents by more than half when examining numeracy. Further, it may be that SAT Math scores are not reflective of numeracy in a direct way. Different divisions of SAT Math scores into various levels of numeracy did not alter results, which eventually led to the simple two divisions used in the current study. In all division of the sample (two, three, or four ways) the results were the same: there were no substantive differences between levels of numeracy in regards to logical consistency or predictive validity. Although this may be due to a real lack of differential effect of numeracy, an alternative explanation could be that the lack of differences is because SAT Math scores are not directly reflective of numeracy.

Another limitation is the use of Fisher's z-score transformation to test for statistical differences for correlations. As noted, special techniques are needed to correctly estimate variances for point estimates for use in hypothesis testing when using clustered data. However, at the moment, it is not clear how to transform correlations to z-scores to test for differences taking into account clustered data. Therefore, the choices were to not use any hypothesis tests or use one with some limitation. The latter was selected as an attempt to examine differences statistically with the acknowledgement of the limitation.

A final limitation worth noting is the limited population examined in this study. All respondents were college students, meaning that, for the most part, the respondents were of a certain age range. The use of this population limits the potential generalizability

somewhat. It may be that college respondents are different than the remainder of the population in key regards in use of numeric and vague quantifier responses. For example, the higher level of education compared to other portions of the population may lead to increased logical consistency. It may also be that college students, for whatever reason, are more likely to think in terms of vague quantifiers than others in the population. Even if this is unlikely, it is impossible to tell given this dataset.

Further research on this topic is warranted. First, steps should be taken to overcome the limitations discussed above. Better measures of numeracy could be used to examine differences (see Study 1). A method to transform correlations to z-scores taking into account the complex survey design should be studied. Finally, this type study could be conducted on a larger cross-section of the population to increase generalizability and examine potentially important differences in the population, such as age differences (see Study 3).

In addition to these three, further research could focus on other areas. Besides predictive validity, other aspects of the measurement properties could be examined, such as accuracy (e.g. Study 1). Although different studies have examined these various aspects of measurement properties and measurement error, it is possible to examine them all in one study in a unified manner. Current limitations in data do not allow for examination of all measurement properties at once, and further data collection may create a study so this is possible. Additionally, this study only examined the predictive validity in the domain of frequency estimation. It is possible that there are different response option effects when examining different topic areas, such as subjective probabilities (see Study 3).

Finally it may be worthwhile to examine mode differences in regards to logical consistency and predictive validity. This survey was conducted on the web, but different responses may occur in different modes. Although it is not clear that this would be the case, it may be worth examining. Related to the possible difference in response across modes, more studies should examine the possible clustering that may occur in web (or mail) surveys, and when this clustering may occur. For example, it may only occur with specialized samples conducted within institutions, such as was the case with the current survey. However, as of now, this appears to be an open question.

Study 3

Data and Methods

Data for the third study, regarding measurement of subjective probabilities, comes from the Annenberg Perceptions of Tobacco Risk 2 surveys. The Annenberg Perceptions of Tobacco Risk 2 surveys were conducted in the fall of 1999 and winter of 2000 using two national samples in the United States selected through random digit dialing (RDD). The first sample consisted of youths ages 14 – 22, with the second consisting of adults ages 23 and older. After households were selected at random, a screener at initial contact ascertained if at least one household member fell in the age range for the active sample. Parental notification and acceptance were required for minor respondents under the age of 16. The survey instruments were nearly identical for each sample; the differences between the surveys consisted of a few additional questions for minors not relevant for the adult sample (e.g. “Do you live with your parents?”). The final sample size for the youth sample is 2002, while the sample size for adults is 1504 (for a combined sample size of 3506). The response rate is reported to be 51% (Jamieson and Romer 2001).

Several questions were asked regarding the perceived risk of smoking. These questions were asked near the beginning of the survey, following short sections on demographic information, a think aloud item about smoking and health, and questions about smoking history. Respondents were then asked a quantitative risk assessment in the standard format: “Now I would like you to imagine 100 cigarette smokers, both men and women, who smoked cigarettes for their entire adult lives. How many of these 100 people do you think will die from lung cancer?” (A complete list of questions used in for this study can be found in Appendix 5). Although this question asks about how many will die from lung cancer as opposed to how many will develop lung cancer, findings suggest respondents perceive that developing lung cancer is nearly synonymous to dying (Viscusi 2002). That is, there is not a fine discrimination by respondents between asking about “die from lung cancer” and “get lung cancer”. This is supported by the fact that the five-year survival rate for lung cancer is only 15% (Jemal et al. 2009). This measure is considered the absolute risk measure for smokers.

This question was followed by a similar one except instead of asking about smokers, it queried about non-smokers. Responses were for how many non-smokers out of 100 would die from lung cancer. This item is the absolute measure for non-smokers. This question is structured in the same manner as the standard subjective probability question on smoking risk, and therefore likely activates similar cognitive processes. Further, this question allows for computation of the risk difference and the relative risk estimates. Asking the respondent for the risk difference or relative risk between smokers and non-smokers directly would likely be too cognitively difficult (Krosnick 2001). Decomposing the question allows for more cognitively simple tasks in each question

which can be combined later to obtain the desired measures. This method has been found effective in previous research (Krosnick 2001).

The risk difference measure used is calculated as the difference between the absolute risk measure for smokers and the absolute risk measure for non-smokers. The relative risk measure is estimated by dividing the absolute risk measure of smokers by the absolute risk measure for non-smokers. A problem that arises is for those respondents that respond with a zero to the absolute risk measure for non-smokers. To overcome the potential for a zero in the denominator of the estimate a small amount is added (0.5) to these responses to allow for estimable relative risks, consistent with practices in similar research (Krosnick 2001).

The question asking for perceived risk of smoking using a vague quantifier response scale followed shortly after the numeric scale questions. This question was phrased as follows: “In your opinion, would smoking everyday be very risky for your health, somewhat risky, a little risky or not at all risky for your health?” Unlike Studies 1 and 2, there was not a direct numeric translation of the vague quantifier question. As such, it may be that there will be less logical consistency between the two measures than was observed in those studies. This expectation is due to the fact that since respondents were not directly asked to translate the vague quantity, the link between the two is not explicit, and therefore consistency may not materialize in a manner similar to prior studies. Since the respondent was not asked about and did not have to think about what the vague quantity meant in numeric terms, consistency may be lessened.

However, the central research question is which of the above four measures capture perceived risk best: absolute risk to smokers, risk difference between smokers

and non-smokers, relative risk between smokers and non-smokers, or risk on a vague quantifier scale. Therefore, techniques for examining the measurement properties and potential validity of these questions are employed. Since the concern is relative performance in measuring risk perceptions and the ability of these measures to predict outcome behaviors (i.e., smoking), the goal is not to necessarily ascertain whether smoking risks are over- or underestimated or the causes of risk perception. Previous research has indicated that youth and adults may perceive risk differently, with youth seeing risk more objectively and adults in a vaguer, intuitive sense (Reyna and Farley 2006). However, adults can be expected to have greater numeric ability compared to youths. As such, analyses examine not only the combined data but also the data for the adult and youth samples separately. In addition, as a proxy for numeracy, level of education is used, which is a variable collected in the survey. Although education may not be a direct measure of numeracy such as numeracy tests or standardized tests scores in mathematics, education is related to numeracy, with lower levels of education related to lower levels of numeracy (Galesic and Garcia-Retamero 2010). Given the lack of other measures of numeracy, education appears to be the most appropriate proxy measure. As in Galesic and Garcia-Retamero (2010), high education is defined as having at least some college education, with low education being defined as having a high school degree or less education.

The theoretically related variable of interest for use in assessing predictive validity is smoking status, that is, whether the respondent is a smoker or non-smoker. Being a smoker is defined as those responding that they smoked one or more cigarettes per day, otherwise respondents are defined as non-smokers. Generally, risk beliefs should

be related to risk behaviors (Slovic 1987, Diefenbach et al. 1993). Those with higher levels of risk perceived are generally less likely to partake in that behavior. Regarding smoking risk perceptions specifically, other research has related numeric measures of perceived risk to smoking status (Viscusi 1990, 2002, Krosnick 2002). Like other risk research, the findings suggest that higher levels of perceived risks of smoking are related to lower smoking incidence.

Hypotheses

Given that Studies 2 and 3 both are similar, with Study 2 examining frequency estimation and Study 3 examining subjective probabilities, hypotheses 1 through 4 are the same for both, with the remaining hypotheses being study specific.

H1: In general, vague quantifiers will show higher levels of predictive validity with theoretically related variables compared to numeric open ended responses.

H1a: There will be higher correlations and better model fit for when using variables measured by vague quantifiers compared to numeric open ended responses.

H2: When there is an indirect numeric translation of vague quantifiers, there should not be a logical consistency between numeric and vague quantifier measures.

H3: In general, more numerate individuals will provide responses showing greater predictive validity.

H4: More numerate individuals will show greater logical consistency in numeric open ended and vague quantifier responses than less numerate individuals.

H5: Younger respondents will show lower levels of predictive validity than adults.

H6: Younger respondents will show lower levels of logical consistency than adults.

H7: More educated (numerate) individuals will show greater predictive validity using numeric open ended responses than less numerate individuals.

H8: However, more educated (numerate) individuals will not show greater predictive validity using numeric open ended responses than when using vague quantifiers.

H9: Younger respondents will show higher levels of predictive validity when responding with numeric open ended responses than adults using numeric open ended responses.

H10: Younger respondents will show higher levels of predictive validity when responding with numeric open ended responses than when responding with vague quantifier responses.

Results

Measures of Risk

The estimates of central tendency for the risk measures for the adult, youth, and combined samples are included in Table 3.1. The means and standard errors are presented for the numeric measures and the proportion of those saying that smoking everyday would be "very risky" on the vague quantifier scale. The means for the three of the numeric risk estimates of absolute risk to smokers, risk difference, and relative risk, if taken at face value, indicate overestimation of risk. Like previous findings, for all samples, the perceived absolute risk of smoking is significantly larger than the actual risk of 6 - 13 in 100 (Viscusi 2002; Viscusi and Hakes 2008). Using this same criterion, the risk difference also suggests overestimation of the risks of smoking, although for youths the mean difference is significantly greater than for adults, and significantly greater for lower educated than higher educated. According to data from the 1989 United States'

Surgeon General's report on smoking and health, the actual relative risk of smokers to non-smokers for lung cancer is about 13:1 (Krosnick 2001). The relative risk mean for adults is nearly exactly this estimate, with higher educated having close to this estimate. For the youths and those with lower education, however, the relative risk is overestimated.[10] This overestimation inflates the mean of the combined sample. The means of the absolute risk of lung cancer for non-smokers is actually the most overestimated (relatively), as the lung cancer death rate in non-smokers is less than 0.02 per 100 (Thun et al. 2006). The proportion saying that smoking is "very risky" is quite high for all samples, with over 80% of the samples responding this way. This response, too, suggests that perceived risk in smoking is great.

Table 3.1

Risk Measure Tendencies

	Absolute Risk (Smokers)	Absolute Risk (Non-smokers)	Risk Difference	Relative Risk	Saying "Very Risky"
Youth (14-22)	60.39	13.45	46.84	30.43	0.81
(s.e.)	(0.56)	(0.40)	(0.64)	(1.21)	
Adult (23+)	48.47	15.18	33.21	13.10	0.83
(s.e.)	(0.73)	(0.43)	(0.72)	(0.86)	
Low Educ.	58.19	14.59	43.50	28.89	0.82
(s.e.)	(0.59)	(0.41)	(0.66)	(1.19)	
High Educ.	51.09	13.51	37.56	14.32	0.82
(s.e.)	(0.72)	(0.41)	(0.71)	(0.89)	
Combined	55.42	14.17	41.19	23.25	0.82
(s.e.)	(0.46)	(0.29)	(0.49)	(0.81)	

[10] Relative risk estimates are larger than would seem by the means for absolute risk to smokers and nonsmokers due to the large skew in the data, with a large number of respondents who gave zero to absolute risk to nonsmokers also giving large estimates to absolute risks to smokers. For example, for the total sample, 446 respondents gave zero responses to the absolute risk to nonsmokers question, but had a mean of 63.41 for absolute risk to smokers.

Logical Consistency

To ascertain the relationship between the vague quantifiers and the numeric values, means were calculated for each of the numeric risk values at each level of the vague quantifier scale. Assuming that the vague quantifier and the numeric risk estimates are measuring the same construct, there should be a level of concordance between them. Table 3.2 presents the means for the standard absolute risk to smokers measure at each level of the responses to the vague quantifier scale for each of the samples.[11] Absolute risk for smokers is the numeric measure most often used in risk perception studies, so this is the comparison of central concern. This analysis is similar to research attempting to identify the numeric values of vague quantifiers (Budescu and Wallsten 1985, Clarke et al. 1992, Lichtenstein and Newman 1967). However, unlike those studies, there was not an explicit request for a numerical translation of the vague quantifier selected by the respondent. As expected, the highest means for all samples are among those saying smoking is “very risky”. All means are significantly greater than those at each of the next two levels down the scale at the $p<.05$ level. However, among adults and across both education levels, the mean for those saying smoking is “very risky” is not significantly different from those saying smoking is “not at all risky”. Further, for adults, the cell mean for “not at all risky” is significantly larger than those giving either of the next two higher risk responses on the quantifier scale. For those with higher education, the cell mean for

[11] Since means are based on cases where both measures are not missing, the total means are based on somewhat smaller numbers than those in Table 1, resulting in minor differences in means across tables.

"not at all risky" is significantly larger than the mean for those saying "a little risky" and not significantly different from those "somewhat risky". Finally, for both the youth and lower educated, as well as the combined sample, the cell means of the lower three responses on the scale are not significantly different from one another in any instance. As an additional test, in order to ensure that small cell size did not affect results, the final two categories of the vague quantifier scale were combined and all analyses conducted on the full scale were also conducted on the collapsed scale. No substantive findings changed by using the collapsed scale.

Table 3.2

Absolute Risk Means by Vague Quantifier Response

Vague Quantifier Response	Youth (14-22)	Adult (23+)	Low Educ.	High Educ.	Combined
Very Risky	61.86[a] (n = 1637)	51.52[a] (n = 1153)	60.38[a] (n = 1662)	53.29[a] (n = 1093)	57.58[a] (n = 2790)
Somewhat Risky	54.42[b] (n = 252)	35.00[b] (n = 194)	48.46[b] (n = 260)	42.19[b] (n = 182)	45.97[b] (n = 446)
A Little Risky	49.81[b] (n = 59)	28.24[b] (n = 45)	45.22[b] (n = 76)	26.78[c] (n = 27)	40.48[b] (n = 104)
Not At All Risky	52.08[b] (n = 26)	42.79[a] (n = 19)	53.00[a,b] (n = 28)	42.69[a, b] (n = 16)	48.16[b] (n = 45)
Total	60.42 (n = 1974)	48.39 (n = 1411)	58.19 (n = 2042)	51.09 (n = 1318)	55.41 (n = 3385)

Means are estimated lung cancer fatalities out of 100 smokers. Different superscript letters indicate significant differences of means within columns at $p<0.05$ level.

In all samples, it appears that a large inconsistency occurs among those saying that smoking is "not at all risky". For the first three points on the scale, there appears to

be a decreasing estimate of absolute risk, even if this decrease is not significant. At the last response level, however, the mean increases for every division of the sample. This is at least partly due to the small number of people selecting this response on the vague quantifier scale. Smaller number of responses for the middle two categories ("somewhat risky" and "a little risky") may also help explain the lack of significant differences between them, as the absolute risk measure displays high levels of variance. Still, there is a significant decline between the middle two categories for those with higher education. Similar patterns of linear decline in the means of the risk difference and relative risk was found across the first three response categories of the vague quantifier scale, with the last response option showing higher means (not shown).

These patterns of the means of the absolute risk measure across the responses to the vague quantifier scale lack the consistency desired to suggest that both are measuring risk perceptions in an equivalent manner. First, as indicated, the means do not always fluctuate in the expected manner. Further, and of potentially more importance, is that if one accepted the numeric values as accurate, then at all response levels, the risk of smoking is greatly overestimated. Those saying that smoking is only "a little risky" or "not at all risky" still vastly overestimate the risks, as do people saying smoking is "very risky". Semantically, this does not seem consistent. To say smoking is only "a little risky" or "not at all risky", and that the risk is numerically similar to those saying "very risky", with all having vast numeric overestimates of risk suggests problems in understanding of the scales. This inconsistency occurs across all sample divisions, suggesting that all respondents have difficulty with one or both of the response options. Given the lack of numeracy among the general population (e.g. Peters et al. 2007) and the

wide use and expressed preference for responding with vague quantifiers, it seems possible the absolute risk measure is problematic.

To that end, it is important to examine the pattern for those with lower and higher levels of education, which are proxies for numeracy. Of all the sample divisions, only the division of higher educated respondents show statistically significant declines over the first three response options, with other divisions showing no difference between the middle response options. This finding suggests that, at least in some small way, increased numeracy does increase the logical consistency between vague quantifier and numeric absolute risk measures. However, it should be noted that as with other divisions in the sample, the cell mean for final response option for “not at all risky” is not significantly different or greater than other response options indicating greater risk. Additionally, as with all other divisions of the sample, higher educated respondents also give absolute risk estimates suggesting vast overestimation of the risk at all levels of response on the vague quantifier scale, further indicating discordance in responses.

Based on Table 3.1, the above results indicate that overall there appears to be a logical discordance between vague quantifier and absolute risk responses. The fact that there is discordance between the two suggest support for Hypothesis 2, stating that there should be discordance when the translation from vague quantifiers to numeric responses is not done directly. However, this discordance may also be due to the way the questions were also asked. The numeric question asked about anyone’s risk for lung cancer, whereas the vague quantifier scale asked about health risk to oneself. Hence, not only were the apparent objectives of the questions different with regard to the persons who are

susceptible to risks, but also with regard to whether risk is targeted to a specific health condition or to health more broadly.

Not all of the hypotheses regarding logical discordance between vague quantifier and numeric absolute risk measures are supported, however. Given the pattern of cell means, lower and higher educated respondents show similar problems. For both, the cell mean for final response option for “not at all risky” is significantly not different or greater than other response options indicating greater risk. Further, both sets of respondents also give absolute risk estimates suggesting vast overestimation of the risk at all levels of response on the vague quantifier scale, further indicating discordance in responses. However, higher educated respondents did show significant declines over the first three response options, with lower educated showing no difference between the middle response options. This suggests partial support for Hypothesis 4, as more educated respondents may be showing some level of greater concordance in responses than lower educated, but overall, there is still a high level of discordance even for higher educated respondents.

Finally, there is little support for Hypothesis 6. Both youth and adult samples show a similar pattern. The only difference is that, among adults, those saying “not at all risky” are significantly greater than the middle response but not different from the “very risky” response. Conversely, for youth, the last response is not statistically different from the middle responses and is significantly smaller than the “very risky” response. Further, adults give smaller absolute risk estimates at each level of the vague quantifier response than youth, but both still show levels of overestimation at each level of response. Given

the similar pattern and overall level of discordance in both youth and adults, it is not possible to say that either show less logical discordance than the other.

Predictive Validity

Past research suggests that the efficacy of these measures is indicated by the relationship to an outcome behavior of interest (Diefenbach, Weinstein, and O'Reilly 1993; Windschitl and Wells 1996). In this case the behavioral outcome of interest is smoking, with higher risk perceptions observed with lower incidence of smoking. Smoking is defined in the current context as those respondents saying they smoke at least one cigarette a day. Using this definition, the proportion of smokers in the youth sample is 0.191 and 0.184 in the adult sample. For the low education division, the proportion of smokers is 0.199 and for the high education division it is 0.171. Overall, for the full sample the proportion of daily smokers is 0.188. In order to further assess the utility of these varying numeric (measured on the 0-100 open format) and vague quantifier (measured on a 1-4 ordinal scale) measures, each measure was correlated with smoking status, dichotomized as smoker/non-smoker (smoker = 1). Point bi-serial correlations were estimated using the numeric risk estimates and a rank bi-serial correlation was estimated between the vague quantifier scale and smoking status. Correlations were transformed to their Fisher Z-scores in order to test for significant differences. Results are presented in Table 3.3.

Table 3.3

Correlation of Risk Measures with Smoking Status

	Youth (14-22)	Adult (23+)	Low Educ.	High Educ.	Combined
Absolute Risk	-0.08[a]	-0.17[a]	-0.15[a]	-0.08[a]	-0.12[a]
Risk Difference	-0.09[a]	-0.22[a]	-0.15[a]	-0.11[b]	-0.14[a]
Relative Risk	-0.08[a]	-0.04[#,b]	-0.10[a]	0.004[#,a]	-0.07[b]
Vague Quantifier	-0.19[b]	-0.35[c]	-0.21[b]	-0.23[c]	-0.27[c]

Different superscripts indicate significant differences of correlations within columns at $p<0.05$ level.
#indicates correlation not significantly different from zero at $p<0.05$ level

Nearly all correlations are in the expected direction, with higher scores on a measure related to lower smoking incidence. The one exception is for the correlation of smoking status and relative risk in the higher educated sample, and this is not significantly different from zero. Overall, the adult sample's correlations between the risk measures and smoking status are higher than for the youths. Only the relative risk measure in the adult sample has a correlation with smoking status not significantly different from zero. Similarly, the correlations tend to be somewhat higher for low education than for high education, excepting the correlation with the vague quantifier measure. Importantly, although all the measures generally show a relationship to the behavior of interest, the relationship is significantly stronger for the vague quantifier scale in every division of the sample. This finding is consistent with prior research (Windschitl and Wells 1996). This consistency with previous studies further indicates

that the vague quantifier scale may have better predictive validity and preferred in measuring smoking risk perceptions.

The findings in Table 3.3 provide support for some of the proposed hypotheses and against others. First, the data provide strong support for Hypotheses 1 and 1a. The correlations for all divisions of the sample are stronger for vague quantifier measures than for any of the numeric measures. These stronger correlations suggests greater predictive validity for vague quantifier measures compared to numeric open ended responses. However, Hypothesis 3 is not supported. Contrary to the expectation, the more numerate sample did not show higher levels of predictive validity. Rather, those in the higher educated sample tended to have smaller correlations with smoking status than did those in the lower education division. Still, only one of the differences is statistically significant, with the correlation between smoking and the relative risk measure significantly greater in the lower educated sample ($p<0.05$). Regardless, the pattern is contrary to Hypothesis 3.

Similarly, Hypothesis 7 is not supported. More educated respondents did not show higher levels of predictive validity when using numeric measures than less numerate individuals. In fact, the trend tends to be somewhat opposite this expectation, as noted above. Hypothesis 8 is supported based on the data in Table 3.3, however. More numerate individuals did not show higher correlations (predictive validity) using numeric measures than vague quantifiers. Like all other divisions, more educated respondents showed higher correlations using vague quantifier measures compared to numeric open ended measures.

Hypothesis 5 is also generally supported by the correlations in Table 3.3. Younger respondents did provide lower correlations in all instances, except between smoking and the relative risk measure. For this correlation with relative risk, the difference between youth and adult samples was not significant ($p = 0.20$). For the remaining correlations, all of the correlations were larger for the adult sample than the youth sample at the $p<.05$ level. These findings all suggest greater predictive validity generally among the adult sample compared to the youth sample.

The remaining hypotheses concerning the predictive validity in the youth and adults samples are not supported. The youth sample did not display higher correlations than adults when responding using numeric open ended measures (Hypothesis 9), as might be expected by the argument by Reyna and Farley (2006) that youths perceive risk in terms of the objective measures as opposed to adults, who are argued to see risks in a vague, fuzzy gist manner. Rather, higher correlations were generally found for the numeric measures in the adult sample, with two of the three larger at statistically significant levels. The lone remaining measure, for relative risk, was not significantly different in the two samples. Finally, Hypothesis 10 is not supported. Contrary to this hypothesis, the youth sample, like all other divisions, had higher correlations for the vague quantifier measure than for the other numeric measures.

As another test to further gauge the predictive validity of the various risk perception measures, logistic regressions were conducted predicting smoking status, including several important control variables. Since the purpose is to compare models to ascertain which risk perception measure is best in modelling smoking behavior in order to further assess the predictive validities, not all possible control variables predicting

smoking are required. The variables used are race, age, and education. Specifically, white respondents were compared to all others, as whites have been found to hold distinctive perceptions of risks compared to other races (Finucane et al. 2000). Education was measured as six increasing categories of educational achievement, beginning with eighth-grade education and ending with completion of a college degree or higher. For comparisons across low and high educated samples, only race and age are used as control variables, as it does not make sense to include an education variable.

Since the purpose is to identify which measure of risk perception increases model fit and predictive capability, separate models for each of the risk measures are compared using the criterion of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Agresti 2002). Since the AIC and BIC are influenced by the sample size, all models are restricted to include only cases where data is available for all risk measures to increase comparability. In addition, the AIC and BIC both penalize for model complexity, as more included variables increase AIC and BIC values. Given the skewed nature of the vague quantifier scale, indicators were created for each of the response options, leaving the indicator for “not at all risky” out of the model as the baseline category. By including dichotomous indicators for three of the response categories on the vague quantifier scale, there are two more parameters in the vague quantifier scale models than those models with numeric measures, giving a slight initial edge to the numeric risk perception models in terms of calculation of the AIC and BIC. Lower AIC and BIC are considered to be better model fits. As an additional indicator for model fit, the area under the Receiver Operator Characteristic (ROC) curve, a measure of diagnostic accuracy, was estimated. Higher ROC is indicative of better model fit. A base model was

also estimated, which includes only the demographic variables to show improvement based on the measures of risk perception. The AIC, BIC, and area under the ROC curve are presented in Tables 3.4-3.6.

Table 3.4

Logistic Regression Models Fit Indicators by Age

	Youth (n = 1923)			Adult (n = 1366)		
	AIC	BIC	Area Under ROC Curve (s.e.)	AIC	BIC	Area Under ROC Curve (s.e.)
Base	1666.81	1689.05	0.726 (0.015)	1274.93	1220.24	0.683 (0.019)
Absolute Risk	1662.76	1690.57	0.728 (0.015)	1149.26	1175.36	0.725 (0.018)
Risk Difference	1660.88	1688.69	0.729 (0.015)	1135.63	1161.73	0.739 (0.018)
Relative Risk	1666.48	1694.29	0.728 (0.015)	1197.46	1223.56	0.686 (0.019)
Vague Quantifier	1615.92	1654.86	0.756 (0.014)	1102.02	1138.56	0.758 (0.017)

Table 3.5

Logistic Regression Models Fit Indicators by Education (Numeracy)

	Low Educ. (n = 1994)			High Educ. (n = 1295)		
	AIC	BIC	Area Under ROC Curve (s.e.)	AIC	BIC	Area Under ROC Curve (s.e.)
Base	1938.14	1954.93	0.644 (0.016)	1130.10	1145.60	0.614 (0.020)
Absolute Risk	1902.53	1924.92	0.630 (0.016)	1118.30	1138.96	0.633 (0.020)
Risk Difference	1892.53	1914.92	0.665 (0.016)	1106.60	1127.26	0.650 (0.020)
Relative Risk	1923.12	1945.51	0.623 (0.015)	1132.01	1152.67	0.616 (0.020)
Vague Quantifier	1831.53	1865.11	0.704 (0.015)	1073.31	1104.30	0.694 (0.020)

Table 3.6

Logistic Regression Models Fit Indicators by Total Sample

	Combined Total (n = 3289)		
	AIC	BIC	Area Under ROC Curve (s.e.)
Base	3092.64	3117.03	0.548 (0.012)
Absolute Risk	3042.95	3073.44	0.623 (0.012)
Risk Difference	3017.10	3047.60	0.644 (0.012)
Relative Risk	3078.75	3109.24	0.586 (0.012)
Vague Quantifier	2922.50	2965.19	0.657 (0.013)

The results are consistent and clear across all samples. By all criteria, the models using the vague quantifier scale fit better than those using numeric risk estimates. Even with the penalty for the additional parameters, the AIC and BIC are smallest for models using the vague quantifier scale. Further, although the area under the ROC curves for the various models are similar, it is always greatest for the model using the vague quantifier scale. This confirms the findings from the correlation analysis while including controls related to risk perception. For almost all samples, relative risk performed worst on all criteria, with higher BIC than the base model in adult and youth samples, as well for the higher educated sample. Risk difference performed slightly better than did the absolute risk measure, especially in the youth sample, where the BIC for the absolute risk model is larger than the base model. It appears that while the standard measure of absolute risk is related to smoking status, the vague quantifier and risk difference measures model the smoking decision better.

These results also provide additional evidence regarding the hypotheses about predictive validity in combination with the correlation results. Again, most clearly, Hypotheses 1 and 1a are supported by the results of the logistic regression models, with all model diagnostics suggesting greater predictive validity for the vague quantifier measure compared to all of the numeric open ended response formats, regardless of the sample division. Based on the AIC, BIC, and ROC, it is also clear that Hypothesis 8 is supported, as numeric measures did not display more predictive validity than the vague quantifier measure for more educated respondents. Further, Hypothesis 10 is not supported based on all of the model diagnostics. Like all other sample divisions, youth

responses using the vague quantifier scale shows higher predictive validity compared to numeric measures, contrary to expectations. These findings all conform to the findings of the correlation analyses presented above.

For hypotheses comparing differences across samples, it is not possible to use the AIC and BIC as these assume that the models all come from the exact same sample for comparative usage. However, the ROC does provide some evidence in regards to these hypotheses. There is little support for Hypothesis 3, as the area under the ROC is not consistently greater for low or high educated respondent models, and no difference is significant at the $p<.05$ standard. Similarly, Hypothesis 7 is not supported, as the area under the ROC curve is not always higher for numeric measures among more educated respondents compared to less educated, and there are no significant differences. In combination with the results from the correlation analyses, the results suggest little support for Hypotheses 3 or 7, and more educated (numerate) respondents did not display higher predictive validity in any of their responses.

There is also little support for Hypothesis 5 based on the ROC from these models. The ROC is similar between the adult and youth samples, sometimes higher for one, sometimes for the other, but never significantly different. Still, Hypothesis 5 is supported based on the findings in Table 3.3, and given the nonsignificant differences in the ROC analyses for the logistic models, it suggested that there is general support for this hypothesis. Based on these models, there is some support for Hypothesis 9. For two of the three numeric measures, youth responses led to models with higher areas under the ROC curve. However, these differences were not significant, and for the third numeric measure, risk difference, it was higher for adults. In combination with the lack of support

in the correlation analyses, it suggests that youth do not give responses with higher predictive validity using numeric measures than do adults.

The estimated coefficients and odds ratios for the risk measures controlling for demographics are presented in Tables 3.7-3.9. Among the numeric risk measures, the risk difference measure produced a greater effect than either than the absolute risk or relative risk measures (which was also reflected in the standardized coefficients, not shown). Looking across samples at each measure shows that the youth sample has smaller estimated coefficients and odds ratios closer to one than the adult sample. This attenuates the effects found in full sample data, leading to effects in between the youth and adult samples. For the vague quantifier measure, none of the included category coefficients included in the youth model are significant. However, chi-square tests for the joint effect of all three indicators are highly significant in all five samples (for youth, $\chi^2 = 54.65$, for adults, $\chi^2 = 104.65$, for lower educated sample, $\chi^2 = 118.07$, for higher educated sample, $\chi^2 = 66.78$ for combined sample, $\chi^2 = 185.90$, all df = 3, p<.0001). This significance, in addition to the greater fit statistics for this model compared to others, suggests the efficacy of the vague quantifier scale.

Table 3.7

Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking by Age

	Youth (n = 1923)			Adult (n = 1366)		
	Coefficient	(s.e.)	Odds Ratio	Coefficient	(s.e.)	Odds Ratio
Absolute Risk	-0.006*	(0.002)	0.994	-0.021*	(0.003)	0.980
Risk Difference	-0.006	(0.002)	0.994	-0.025*	(0.003)	0.976
Relative Risk	-0.002	(0.001)	0.998	-0.005	(0.003)	0.995
Vague Quantifier						
Very Risky	-0.113	(0.593)	0.893	-1.379*	(0.537)	0.252
Somewhat Risky	1.042	(0.606)	2.835	0.133	(0.551)	1.142
A Little Risky	0.965	(0.661)	2.625	0.955	(0.620)	2.599

*p < 0.05

Table 3.8

Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking By Education (Numeracy)

	Low Educ. (n = 1994)			High Educ. (n = 1295)		
	Coefficient	(s.e.)	Odds Ratio	Coefficient	(s.e.)	Odds Ratio
Absolute Risk	-0.012*	(0.002)	0.988	-0.013*	(0.003)	0.987
Risk Difference	-0.013*	(0.002)	0.987	-0.016*	(0.003)	0.984
Relative Risk	-0.005	(0.001)	0.995	-0.001	(0.002)	0.999
Vague Quantifier						
Very Risky	-0.216	(0.562)	0.806	-1.493*	(0.543)	0.225
Somewhat Risky	1.108	(0.573)	3.028	-0.209	(0.560)	0.811
A Little Risky	1.386*	(0.609)	3.998	0.329	(0.668)	1.389

*p < 0.05

Table 3.9

Estimated Coefficients of Risk Measures in Logistic Regression Models Predicting Smoking By Education (Numeracy)

	Combined Total (n = 3289)		
	Coefficient	(s.e.)	Odds Ratio
Absolute Risk	-0.013*	(0.002)	0.987
Risk Difference	-0.015*	(0.002	0.985
Relative Risk	-0.005*	(0.001)	0.998
Vague Quantifier			
Very Risky	-0.749*	(0.371)	0.473
Somewhat Risky	0.614	(0.380)	1.848
A Little Risky	0.962*	(0.420)	2.618

*p < 0.05

Finally, models were tested adding each of the numeric measures of risk perception to the vague quantifier model (not shown). Although vague quantifiers better predict smoking status, it may be that numeric risk may still be important in decision making in addition to that of vaguely quantified risk. Using the AIC and BIC as criteria for model improvement, adding the numeric measures consistently improved model fit for absolute and risk difference measures in the adult sample, for risk difference in the higher educated division, and for all numeric risk measures for the lower educated sample (not shown). By adding these to the vague quantifier model, both the AIC and BIC decreased. In these models, the numeric risk estimate is significant and in the expected direction (greater perceived risk decreasing smoking likelihood), while not changing the

direction or significance of the other indicators in Table 3.5. These findings suggests the possible efficacy of asking both numeric and vague quantifier measures for many respondents, including, contrary to expectation, lower educated samples. Also somewhat contrary to expectation, for the youth samples, adding any numeric risk measure to the vague quantifier model led to increases in the AIC and BIC. The lone exception was when the risk difference measure was added, which decreased the AIC from 1615.92 to 1614.63. The BIC for this model, however, increased from 1654.86 to 1659.12.

Discussion and conclusions

How perceptions are measured is important in risk evaluation and for survey methodology generally. The differences in potential risk perception measures have been largely overlooked. One measure is often selected for a survey without tests of comparative validity. In the case of smoking risk perceptions, research has long been divided on how well people understand the hazards from smoking. These divisions have frequently focused on the results of the standard question asking for a numeric estimate of absolute risk of smoking. However, using national surveys of youths and adults, the current research shows that the standard question of absolute risk does not perform as well as other risk measures. Given the potential issues with this question, such as the numeracy of respondents and individuals' potentially representing risk in a "fuzzy" manner (e.g. Reyna 2004), this may not be surprising. Similarly, other numeric measures did not relate to smoking behavior as well as more qualitative estimates. The means of these numeric measures of risk did not conform to a pattern expected given the verbal representations of risk respondents gave in other parts of the survey. To say smoking is

“not at all risky” on one measure, yet apparently vastly overestimate the risk on the standard absolute risk measure does not seem coherent.

Since people naturally speak in vague quantifiers and often prefer to express themselves in this way, it may not be surprising that the vague quantifier measure performed better than numeric measures. Indeed, higher predictive validity was found for a measure using a vague quantifier scale, consistent with other findings (Windschitl and Wells 1996). The vague quantifier scale was more highly correlated with smoking behavior and led to better fitting predictive models; this was true for all divisions of the sample. This finding suggests that all age groups report risks in similar ways, possibly contrary to prior theory (e.g. Reyna and Farley 2006). Like adults, youth smoking behavior shows a greater relationship with vague, rather than numeric, risk estimates. Understanding how youth perceive smoking risk is of particular importance given that many smokers begin at young ages (Escobedo et al. 1990). Additionally, higher correlations and better model fit were similarly found for the vague quantifier scale in both higher and lower education groups. This suggests, contrary to expectations, that numeracy did not have a substantive impact on measurement of subjective probabilities, at least in terms of impact on numeric and vague quantifier measures.

The results of the current study also mirror those found by Kahneman et al. (1993) in a study of contingent valuation. Like that study, this study found that qualitative measures of attitude were psychometrically superior to numeric estimates, at least in predicting decisions. Further, both studies suggest that numeric estimates may be expressions of similar attitudes as those on the qualitative scale. Rather than being considered true values of numeric risk (or willingness to pay), these numeric estimates

may be simply expressions of attitude on an arbitrary scale (similar to the suggestion of Borland (1997)).

Although several of the hypotheses were not confirmed regarding the effects of age or numeracy, the most clearly important hypothesis in terms of its importance for survey methodology was supported. That is, by all measures and tests, the results suggest that the vague quantifier measure outperforms numeric open ended measures for subjective probabilities (Hypotheses 1 and 1a). These results therefore are suggestive of lower measurement error, an important component of Total Survey Error (Groves 1989), for the vague quantifier scale. As such, the current results point to greater use of vague quantifier scales compared to numeric open ended response options in designing surveys, at least when asking questions about subjective probabilities and decision making studies.

There are several limitations to the current research. The first is the use of education as a measure of numeracy. Although correlated with numeracy, as indicated in Galesic and Garcia-Retamero (2010), education is not a direct measure of numeracy. The fact that this is not a direct measure may contribute to the lack of findings of differences across these divisions. Although other studies have also found no difference between levels of numeracy in regards to logical consistency and predictive validity using different measures of numeracy, it is possible that different results may have occurred with a different numeracy measure.

A second limitation is different wording for the numeric and vague quantifier questions. Unlike other studies, which use direct translation of vague quantifier to numeric meaning, this study used two related questions using the two scale formats. However, the question wordings were different between the two, and different question

wording may have an impact on how people respond, even if the target opinion is the same (Schuman and Presser 1981). Importantly, in this case the targets asked about in the question are also different, which may increase the chances of non-comparability. In the numeric questions, the question is asking about a hypothetical 100 smokers, whereas the vague quantifier question is asking about the respondent. The effect of the difference between asking about self or others has been found previously (Schwarz and Bienias 1990). Further, one question asked about lung cancer risk and the other asked about health generally. The differences in questions used in this study may explain the difference in logical consistency and predictive validity. It must be stated that these differences make the potential for comparison between the two measures somewhat limited. However, for predictive validity at least, similar results were found when using the same question, with respondents simply translating vague quantifiers into numeric responses (Study 2). This similarity in results suggests that the differences in questions may not have a large effect in the current findings, however, the difference must be noted.

These findings indicate areas for further research. First, research should examine what impact data collection mode has on responses to perceived risk measures. The current data were collected using telephone surveys, and as indicated by the trend of absolute risk measures, differences may be observed using other modes, such as face-to-face or self-administered questionnaires. Second, although vague quantifiers are more related to smoking behavior than numeric measures, the best formulation of the vague quantifier scale should be identified in terms of wording and number of scale options. Not only may the question change, but also the number of categories. The current

research uses a scale with four response options. More options, either even or odd numbered, may be preferable. Third, it is important to see if these findings hold for other areas of risk perception. This study focused on smoking, but this may not hold for other types of risks. Finally, it should be examined whether it is possible to analyze responses to vague quantifier scales to better discriminate individuals' level of perceived of risk, possibly using latent models. Regardless, numeric estimates of risk from surveys, at least in the case of smoking, should be used with the understanding that these do not predict behaviors of interest as well as qualitative scales. The fact that this holds among youth samples is additionally important as most smoking decisions are made in younger years.

Summary

When asking about quantitative information in surveys, there are a number of ways to provide response options to respondents. Three main response options have been developed and used in requesting quantitative information from respondents: numeric open ended, numeric scales, or vague quantifier scales (Tourangeau et al. 2000). This dissertation has focused on two of these, numeric open ended and vague quantifier scale responses. Numeric scales were not examined because these have been shown to have potentially biasing properties (Schwarz et al. 1985).

Vague quantifier scales are often argued against by many survey researchers, suggesting instead that numeric open ended responses be used (Beyth-Marom 1982, Schaeffer 1991, Tourangeau et al. 2000). This suggestion for avoidance is for several reasons. First, these scales provide response options that are, as the name suggests, inherently vague. Due to this, there is often a large variation in the numeric translation assigned to vague quantifiers (e.g. Wallsten et al. 1985). Further, the scales also have

relative meaning, such as where on the scale a respondent believes they are in comparison to similar others (Schaeffer 1991). Additionally, vague quantifiers have different meanings for different targets (Windschitl and Wells 1996). For example, “a lot” of risk from smoking may be different from “a lot” of risk from arsenic.

However, although many have argued against the use of vague quantifiers, there are also reasons that these scales may be preferable to numeric open ended responses. First, it is not clear that respondents are able to think about quantitative information in an appropriate manner. Studies have shown that overall there is a lack of numeracy (numeric literacy) in the population (e.g. Galesic and Garcia-Retamero 2010). Second, theories such as “fuzzy-trace” and other dual-process theories suggest that people frequently rely on vague, intuitive representations of numeric information rather than on verbatim representation of the numbers (Reyna and Brainerd 2008). Third, other research has shown that the relationship between subjective beliefs and behaviors are stronger when using vague quantifier scales than numeric responses (Windschitl and Wells 1996). Finally, and of particular importance, is that it has been argued that it is more cognitively burdensome to ask about numeric information than vague quantifiers (Bradburn and Miles 1979).

Even though there are arguments that have been put forth for and (mainly) against the use of vague quantifiers, there is a lack of studies that have compared the validity of numeric open ended and vague quantifier responses. This dissertation aimed to fill these gaps with the main objective being examination of which response format has better measurement properties for questions requesting numeric data, specifically, which has the least measurement error. This main objective was achieved through a set of four refined

objectives. First, was to assess the accuracy of the different response formats and compare this accuracy to see which, if either, performed better. Second, and similarly, was to assess and compare the predictive validity of the different response formats. Third, was to examine the logical consistency between the measures, numeric open ended and vague quantifiers, which is related to the face validity of the measures. Fourth, and finally, was to study under what circumstances these measures differed on these aspects of accuracy, predictive validity, and logical consistency.

These objectives were achieved through three studies, each of which examined different aspects of these objectives. The first of these studies (Study 1) used a new and unique data set, created using an experimental method. This experiment produced data that allowed for the examination of accuracy of the different response forms, numeric open ended and vague quantifiers. This data was generated as the actual frequency of events was controlled and known by the experimenter, and respondents were then asked about how frequently the event (word appearance) occurred by either numeric open ended or vague quantifier responses. This data addresses the first objective on differential accuracy.

Respondents who answered the frequency questions using vague quantifiers also had to provide translations of the scale at the end of the experiment. That is, for each of the six scale points, a numeric translation of meaning was given. Not only did this allow for examination of accuracy of the responses by placing a value for each of the vague quantifier scale points, this translation also allowed for the examination of logical consistency of numeric and vague quantifier meanings, i.e. the third objective.

In addition, the experiment presented respondents word lists along with a context word. This context word allowed for the examination of differences in context memory and its effect on response accuracy, in conjunction with response form. In one of the experimental conditions, each target word was presented with the same context word at each presentation; for the other condition, a different context word was presented with each target word. The manipulating of context memory replicated the experiment conducted by Brown (1995). Further, this experiment collected data on numeracy from all respondents, using a scale previously used (Galesic and Garcia-Retamero 2010). By including the manipulation of context memory and the measure of numeracy, the fourth objective was also satisfied, whereby it was possible to examine the circumstances that the two response forms of interest differed in regards to accuracy. The numeracy measure also allowed for examination of any effect on logical consistency, again part of the fourth objective.

Studies 2 and 3 both examined the predictive validity of the different response measures as the main outcome, that is, the second objective. Examination of predictive validity requires relating the measure(s) of interest to a theoretically related variable. Both used preexisting data sets that contained both numeric open ended and vague quantifier responses on the same or related questions. This also allowed for examination of logical consistency (objective three) between the two measures. Both data sets also contained theoretically related variables that had been studied previously, although not in a comparative manner as done here. The difference between these two studies is the domain of the question asked; Study 2 examined the area of behavioral frequency, while Study 3 examined the area of subjective probabilities.

Additionally, both Studies 2 and 3 estimated the effect of numeracy on predictive validity, although the measures used were not as direct as that used in Study 1. Study 2 used SAT Math scores, while Study 3 used education as a proxy (Galesic and Garcia-Retamero 2010). Study 3 also examined the effect of age on predictive validity of the different response forms, numeric open ended and vague quantifiers. Theory such as fuzzy-trace theory suggests that youth and adults perceive risks in different ways, which may affect the validity of the different response forms (Reyna and Farley 2006) across age groups. Specifically, youth perceive risk in a more objective, numerical manner, with adults perceiving risk in a fuzzy, vague way. By examining the differences in numeracy in Studies 2 and 3 and age in Study 3, the fourth objective was also met.

Overall, the findings from these studies provide a generally consistent view that vague quantifiers should not be avoided, contrary to many practitioners' warnings. Study 1 showed that in general, vague quantifiers perform as well as, and in some cases, better than numeric open ended responses in regards to accuracy. Studies 2 and 3 are more definitive; by all analyses vague quantifiers outperform numeric open ended responses, with higher levels of predictive validity in all instances. That is, all three studies find that vague quantifiers are not worse than numeric open ended responses, and in many cases, clearly provide higher levels of validity, which is a goal of any survey question.

Further, Studies 1 and 2 show that there is a high level of logical consistency between numeric open ended and vague quantifier responses. This logical consistency suggests that respondents use both response forms in a coherent way. Studies 1 and 2 therefore show potential face validity for both measures. This concordance was expected given that there was a direct translation between numeric and vague quantifier measures.

However, Study 3 found discordance between the two response forms. This discordance may be due to the fact that unlike in other studies, a direct translation of vague quantifiers to numeric responses was not requested. Rather, the means of numeric open ended responses were examined at the corresponding level of a separate vague quantifier response question.

Finally these studies examined the circumstances where these findings may differ. In regards to accuracy, there tended to be an effect for numeracy, but the effect was not always in a consistent direction. Numeracy sometimes increased accuracy, depending on response form and context memory, and others, decreased accuracy in a relative manner. For example, for numeric open ended responses and different context, greater numeracy tends to reduce error. However, for the same context and numeric open ended responses, greater numeracy appears to be related to greater error. Although generally, vague quantifier responses were as accurate as or more accurate than numeric open ended responses, numeracy at times affected this level of relative accuracy. Similarly, context memory sometimes affected vague quantifier responses in a similarly inconsistent way, with the end result being the same or greater accuracy generally, with some exceptions, but there was no consistent effect observed. Indeed, the overall effect of context memory found in Brown (1995) was not replicated here, suggesting potential revisiting of the effect of context memory.

In regards to predictive validity and logical consistency, however, there was no such variation across different conditions. Regardless of level of numeracy, in both Studies 2 and 3, predictive validity was higher for vague quantifier responses in every analysis. Similarly, in Study 3, there was no difference across ages, with predictive

validity higher for vague quantifier responses again. Results were also the same for logical consistency, with the findings of all studies replicated across all levels of numeracy, and in Study 3, age.

Given that in many of the analyses results were consistent across levels of numeracy and other factors, such as age, suggest that vague quantifiers are more suitable for all respondents than numeric open ended responses. At least this statement is clearly true if the desire is to increase predictive validity. Although numeracy had an impact on accuracy, vague quantifier responses generally performed at least as well, if not better, than numeric open ended responses. Context memory appeared to have less of an impact on accuracy, although context did impact accuracy through interactions with response form and numeracy. Again, these findings suggests that while conditions do exists that may differentially effect the relative accuracy of the different response forms, in almost all conditions vague quantifiers perform as well or better than numeric open ended responses. In combination with the findings on lack of effect of conditions on predictive validity, findings suggest the use of vague quantifiers for all respondents.

These findings and the relative cognitive ease of response using vague quantifiers suggest that there should be more usage of these in surveys. One way to accomplish this, while achieving more “precise” data many researchers desire, is to adopt the methods used in Study 1 and advocated by Bradburn and Miles (1979). A series of questions can be asked using vague quantifier responses, and at the end of the survey, respondents can be asked for their numeric translations of each of the scale points. Therefore, numeric information is provided, shown here to be done in a logically consistent manner, but only

had to be done once. The cognitive burden is reduced, and the benefits of using vague quantifiers found in this dissertation are maintained.

Finally, consideration is merited as to why the common view for avoidance of vague quantifiers has arisen and numeric estimates preferred. First are all of the issues that were pointed out previously that have been found in studies. These issues include that vague quantifiers are inherently vague, have large variation in the numeric translation assigned to vague quantifiers, and that there is relative meaning, both relative to similar other people and relative to other target events.

Another possible explanation is that of researcher influence (Windschitl and Wells 1996). First, researchers tend to be relatively numerate individuals, able to think about and understand numbers better than the general public. Second, researchers may prefer numeric data. Numeric data is thought to be more accurate than vague quantities, is more amenable to the types of analyses that researchers prefer, and can be presented with more certainty of meaning than vague quantifiers, e.g. whether people over- or underestimate risks. Finally, specifically in the case of smoking risk perceptions, one explanation for this preference is the establishment of the absolute risk measure over time. This question was first used by industry research as far back as 1964 and asked several times thereafter (see Table 1). Industry data were then used in published risk perception literature, further establishing it as the standard measure (e.g. Viscusi 1990).

However, none of these reasons actually mean that people have accurate representations of numeric information or that this is the best way to measure beliefs in surveys. Instead, this dissertation shows that more accurate and valid representations can

be conveyed through more qualitative response options. That is, vague quantifier response options can successfully be used in surveys asking about a variety of topics.

References

Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. New York: Wiley

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research*, *25,* 318-340.

Begg, I., Maxwell, D., Mitterer, J. O., & Harris, G. (1986). Estimates of Frequency: Attribute or Attribution?*Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 496-508

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, *6,* 383-406.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly, 65,* 45-74.

Belli, R. F., Smith, L., Andreski, P., & Agrawal, S. (2007). Methodological comparisons between CATI event history calendar and conventional questionnaire instruments: Quality of Life Course Retrospective Reports. *Public Opinion Quarterly*, *71,* 603-622

Benjamin, D., Dougan, W, & Buschena, D. (1997). Individuals' Estimates of the Risks of Death: Part II—New Evidence *Journal of Risk and Uncertainty, 22,* 35-57

Beyth-Marom, R. (1982). How Probable is Probable? A Numerical Translation of Verbal Probability Expressions *Journal of Forecasting 1,* 257-269

Bickart, B. A., Blair, J., Menon, G., & Sudman, S. (1990). Cognitive Aspects of Proxy Reporting of Behavior *Advances in Consumer Research, 17,* 198-206

Biehl, M., & Halpern-Felsher, B. L. (2001). Adolescents' and adults' understanding of probability expressions *Journal of Adolescent Health*, *281,* 30-35.

Biemer, P. P., &Lyberg, L. E. (2003). *Introduction to Survey Quality.* New York: John Wiley and Sons

Blair, E., & Burton, S. (1987). Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Question *Journal of Consumer Research, 14,* 280-288

Blair, E., Sudman, S., Bradburn, N. M., & Stocking, C. (1977). How to ask questions about drinking and sex: Response effects in measuring consumer behavior. *Journal of Marketing Research, 14,* 316-321.

Bless, H., Bohner, G., Traudel, H., & Schwarz, N. (1992). Asking Difficult Questions: Task Complexity Increases the Impact of Response Alternatives *European Journal of Social Psychology, 22,* 309-312

Bogart, L. M., Walt, L. C., Pavlovic, J. D., Ober, A. J., Brown, N. R., & Kalichman, S. C. (2007). Cognitive strategies affecting recall of sexual behavior among high-risk men and women.*Health Psychology. 26*, 787-793 .

Borland, R. (1997). What do people's estimates of smoking related risk mean? *Psychology and Health*, *12*, 513-521.

Bradburn, N. M., & Miles, C. (1979). Vague Quantifiers *Public Opinion Quarterly 43,* 92-101.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys *Science, 236,* 157-161

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1539-1553.

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 898-914.

Brown, N. R (2008). How metastrategic considerations influence the selection of frequency estimation strategies. *Journal of Memory and Language, 58*, 3-18.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation, *Psychological Review, 100,* 511-534

Brown, N. R., & Sinclair, R. C. (1999). Estimating number of lifetime sexual partners: Men and women do it differently. *Journal of Sex Research, 36,* 292-297.

Brown, N. R., Williams, R. L., Barker, E. T., & Galambos, N. L. (2007). Estimating frequencies of emotions and actions: A web-based diary study. *Applied Cognitive Psychology, 21*, 259-276.

Bruce, D., Hockley, W. E., & Craik, F. I. (1991). Availability and category-frequency estimation. *Memory & Cognition, 19,* 301-312.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic statements *Organizational Behavior and Human Decision Processes36,* 391-405.

Burton, S., & Blair, E. (1991). Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys *Public Opinion Quarterly, 55,* 50-79

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student Engagement and Student Learning: Testing the Linkages *Research in Higher Education, 47*, 1-32

Chang, L., & Krosnick, J. (2003). Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week". *Sociological Methodology, 33,* 55-80.

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of Orally Presented Verbal Expressions of Probability by a Heterogeneous Sample *Journal of Applied Social Psychology 22*, 638-656.

Cohn, L. D., Schydlower, M., Foley, J., & Copeland, R. L. (1995) Adolescent misinterpretation of health risk probability expressions *Pediatrics*, *95,* 713-716.

Conrad, F. G., Brown, N. R., & Cashman, E. (1993). How the memorability of events affects frequency judgments. *American Statistical Association, Proceedings of the Section on Survey Methods Research, Volume 2* (pp. 1058-1063). Alexandria, VA: American Statistical Association.

Conrad, F., Brown, N. R., & Cashman, E. (1998). Strategies for estimating behavioral frequency in survey interviews. *Memory, 6,* 339-366.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Thousand Oaks, CA, US: Sage

De Bruin, W. D., Fischoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: “It's a Fifty–Fifty Chance” *Organizational Behavior and Human Decision Processes*, *81,* 115-131.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284, 970-974,*

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility *Health Education Research,82,* 181-192.

Dominitz, J., & Manski, C. F. (1997). Perceptions of economic insecurity: Evidence from the survey of economic expectations *Public Opinion Quarterly, 61,* 261-287

Elmo Roper and Associates. (1964). *A study of the reactions to the Surgeon General's Report on cigarette smoking.* Report prepared for Philip Morris. February.

Escobedo, L. G., Anda, R. F., Smith, P. F., & Remington, P. L. (1990). Sociodemographic characteristics of cigarette smoking initiation in the United States *Journal of American Medical Association, 264,* 1550-1555.

Fischoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A Psychometric study of attitudes towards technological risks and benefits.*Policy Sciences, 9,* 127-152.

Fiser, J., & Aslin, R. N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes *Psychological Science, 12, 499-504.*

Friedman, H. (1982). Simplified Determinations of Statistical Power, Magnitude of Effect and Research Sample Sizes *Educational and Psychological Measurement, 42,* 521-526

Galesic, M., & Garcia-Retamero, R. (2010). Statistical Numeracy for Health: A Cross-cultural Comparison With Probabilistic National Samples *Archives Internal Medicine, 170, 462-468.*

Galesic, M., Garcia-Retamero, R., & Gigerenzer G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology, 28, 210-216.*

Gaskell, G. D., O'Muircheartaigh, C. A., & Wright, D. B. (1994). Survey Questions about the Frequency of Vaguely Defined Events: The Effects of Response Alternatives *Public Opinion Quarterly, 58,* 241-254.

Greene, R. L. (1984) Incidental learning of event frequency *Memory & Cognition, 12*, 90-95.

Groves, R. (1989). *Survey Errors and Survey Costs.* New York: John Wiley.

Groves, R. M., & Magilavy, L. J. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys *Public Opinion Quarterly, 50,* 251-266.

Hasher, L., & Zacks, R. T. (1984). Automatic and Effortful Processes in Memory, *Journal of Experimental Psychology: General, 108,* 356-388.

Hasher, L., Zacks, R. T., Rose, K. C., & Sanft, H. (1987). Incidental Encoding of Frequency Information *The American Journal of Psychology, 100,* 69-91.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78, B53-B64.*

Hintzman, D. L., & Curran, T. (1994). Retrieval Dynamics of Recognition and Frequency Judgments: Evidence for Separate Processes of Familiarity and Recall *Journal of Memory and Language, 33*, 1-18

Holbrook, A., Green, M., & Krosnick, J. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparison of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67,* 79-125.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., & Thun, M. J. (2009). Cancer statistics, 2009.*CA: A Cancer Journal for Clinicians,* 59, 225-249

Johnson, T., Cho, Y., Holbrook, A., O'Rourke, D., Warnecke, R., & Chavez, N. (2006). Cultural Variability in the Effects of Question Design Features on Respondent Comprehension *Annals of Epidemiology*, *16*, 661-668.

Jonides, J., & Naveh-Benjamin, M. (1987). Estimating Frequency of Occurrence *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 230-240.

Kahneman, D., Ritov, I., Jacowitz, K. E., & Grant, P. (1993). Stated willingness to pay for public goods: A psychological perspective. *Psychological Science*, 4, 310-315.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk *Econometrica, 47,* 263-292.

Kelson, S. R., Pullella, J. L., & Otterland, A. (1975). The growing epidemic: A survey of smoking habits and attitudes toward smoking among students in grades 7 through 12 in Toledo and Lucas County Ohio public schools-1964 and 1971. *American Journal of Public Health*, *65,* 923-938.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism, *Cognition, 83, B35-B42.*

Kirsch, L. S., Jungeblut, A., Jenkins, L. & Kolstad, A. (2002). *Adult literacy in America: A first look at the findings of the National Adult Literacy Survey* (3rd ed.) Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kish, L. (1965). *Survey Sampling.* John Wiley and Sons, Inc.: New York.

Kristiansen, C. M., Harding, C. M., & Eiser, J. R. (1983). Beliefs about the relationship between smoking and causes of death. *Basic and Applied Social Psychology*, *4,* 253-261.

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. (2001). Americans' Perceptions of the Health Risks of Cigarette Smoking: A New Opportunity for Public Education Paper Presented at the Conference of Survey Research on Household Expectations and Preferences, November 2 - 3, 2001, in Ann Arbor, MI, USA.

Kumar, R. (2005). *Research Methodology: A step-by-step guide for beginners.* London: Sage Publications

Lee, C. (1989). Perceptions of immunity to disease in adult smokers. *Journal of Behavioral Medicine*, *12,* 267-277.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling Errors in Surveys.* New York: John Wiley and Sons.

Levine, S. C., Jordan, N. C., & Hottenlocher, J. (1992). Development of calculation abilities in young children. *Journal of Experimental Child Psychology, 53, 72-103*

Lichtenstein S., & Newman, J. R. (1967). "Empirical scaling of common verbal phrases associated with numerical probabilities." *Psychonomic Science9*, 563-564.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 551-578.

Linton, M. (2000). Flashbulb memories. In U. Neisser & I. E. Hyman (Eds.), *Memory observed: Remembering in natural contexts* (2d ed.) (pp. 107-118). New York: Worth.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples *Medical Decision Making, 21, 37-44.*

Lohr, S. L. (2010). *Sampling: Design and Analysis, 2nd edition.* Pacific Grove, CA: Duxbury Press.

Lu, M., Safren S. A., Skolnik, P. R., Rogers, W. H., Coady, W., Hardy, H., & Wilson, I. B. (2008). Optimal Recall Period and Response Task for Self-Reported HIV Medication Adherence *AIDS and Behavior, 12*, 86-94.

Luke, D. A. (2004). *Multilevel Modeling* Newbury Park, CA: Sage.

Marincic, J. L. (2011). *Vague Quantifiers of Behavioral Frequency: An Investigation of the Nature and Consequences of Differences in Interpretation* Doctoral Dissertation: University of Nebraska- Lincoln.

McEvoy, C. L., & Nelson, D. L. (1982). “Source Category Name and Instance Norms for 106 Categories of Various Sizes, *The American Journal of Psychology*, *95*, 581-634.

Means, B., & Loftus, E. F. (1991). When Personal History Repeats Itself: Decomposing Memories for Recurring Events *Applied Cognitive Psychology, 5,* 297-318.

Menon, G. (1993). The Effects of Accessibility of Information in Memory Judgments of Behavioral Frequencies *Journal of Consumer Research, 20*, 431-440.

Menon, G. (1997). Are the Parts Better than the Whole? The Effects of Decompositional Questions on Judgments of Frequent Behaviors *Journal of Marketing Research, 34,* 335-346.

Menon, G., Raghubir, P., &Agrawal, N. (2008). Health Risk Perceptions and Consumer Psychology, in *The Handbook of Consumer Psychology*, C. Haugtvedt, P. Herr and F. Kardes, eds., Lawrence Erlbaum and Associates, 981-1010.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22*, 212-228.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215, 1519-1520.*

National Survey of Student Engagement. (2011). About NSSE http://nsse.iub.edu/html/about.cfm, accessed on 07/08/2011.

Nelson Laird, T. F., Korkmaz, A., & Chen, D. (2009) How often is “often” revisited: The meaning and linearity of vague quantifiers used on the National Survey of Student

Engagement. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Nuerk, H., Kaufmann, L., Zoppoth, S., &Willmes, K. (2004). On the Development of the Mental Number Line: More, Less, or Never Holistic with increasing age. *Developmental Psychology, 40, 1199-1211.*

Olson, M. J., & Budescu, D. V. (1997). Patterns of Preference for Numerical and Verbal Probabilities *Journal of Behavioral Decision Making,102,* 117-131.

Peters, E., Hibbard, J., Slovic, P., & Dieckmann, N. (2007). Numeracy Skill and the Communication, Comprehension, and Use of Risk-Benefit Information *Health Affairs*, *26,* 741-748.

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17, 407-413.*

Pohl, N. F. (1981). Consideration in Using Vague Quantifiers. *The Journal of Experimental Education, 49*, 235-240.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A.S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical Linear and Nonlinear Modeling* SSI Scientific Software International, Lincolnwood, IL.

Reyna, V. F. (2004). How people make decisions that involve risk: A dual process approach. *Current Directions in Psychological Science, 13*, 60-66.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice. Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4, 249-262.*

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18, 89-107.*

Reyna, V. F., & Farley, F. (2006). Risk and rationality in adolescent decision making: Implications for theory, practice, and public policy. *Psychological Science in the Public Interest*, *7,* 1-44.

Romer, D., & Jamieson, P. (2001). Do adolescents appreciate the risks of smoking? Evidence from a national survey. *Journal of Adolescent Health,* 29, 12-21.

Roper Organization (The). (1977, June). *A four part survey about the American Cancer Society and the American Lung Association.* Report prepared for the Tobacco Institute.

Roper Organization (The). (1980, May). *A study of public attitudes toward cigarette smoking and the tobacco industry.* Report prepared for the Tobacco Institute.

Sakshaug, J. W., Yan, T., & Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection. *Public Opinion Quarterly 74*, 907-933.

Salber, E. J., MacMahon, B., &Welsh, B. (1962). Smoking habits of high school students related to intelligence and achievement. *Pediatrics*, *29*, 780-787.

Sanford, A. J., Moxey, L. M., & Paterson, K. (1994). Psychological studies of quantifiers. *Journal of Semantics*, *11,* 153-170.

Sanford, A. J., Moxey, L. M., & Paterson, K. (1996). Attentional focusing with quantifiers in production and comprehension. *Memory& Cognition, 24,* 144-155.

SAS Institute, Inc. (2010). *SAS/STAT® 9.22 User's Guide.* Cary, NC: SAS Institute Inc.

Schaeffer, N. C. (1991). Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers. *Public Opinion Quarterly*, *55*, 395-423.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Schwarz, N., & Bienias, J. (1990). What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology, 4*, 61-72.

Schwarz N., Hippler H. J., Deutsch, B., & Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly 49*, 388-395.

Schwartz, L. M, Woloshin, S., Black, W. C, & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127, 966-972.*

Slovic, P (1987). Perception of Risk. *Science, 236,* 280-285.

Slovic, P. (1998). Do adolescent smokers know the risks? *Duke Law Journal, 47,* 1133-1141.

Sutton, S. (1998). How ordinary people in Great Britain perceive the health risks of smoking. *Journal of Epidemiology and Community Health*, *52*, 338-339.

Teigen, K. H., & Brun, W. (2003). Verbal Probabilities: A Question of Frame? *Journal of Behavioral Decision Making 16,* 53–72.

Thun, M. J., Henley, S. J., Burns, D., Jemal, A., Shanks, T. G., & Calle, E. E. (2006). Lung cancer death rates in lifelong nonsmokers. *Journal of the National Cancer Institute*, *98*, 691-699.

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

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context *Public Opinion Quarterly, 60,* 275-304.

Turconi, E., Campbell, J. I. D., & Serona, X. (2006). Numerical order and quantity processing in number comparison. *Cognition, 98,* 273-285.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1130.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and psychology of choice. *Science, 211*, 453-458.

United States Department of Health and Human Services, Public Health Service. (1989). *Reducing the Health Consequences of Smoking: 25 Years of Progress: A Report of the Surgeon General.* Washington, DC: Author.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50,* 289–335.

Viscusi, W. K. (1990). Do smokers underestimate risks? *Journal of Political Economy, 98,* 1253-1268.

Viscusi, W. K. (1992). *Smoking: Making the Risky Decision*. Oxford: Oxford University Press.

Viscusi, W. K. (2002). *Smoke Filled Rooms: A Post-mortem on the tobacco deal.* Chicago: University of Chicago Press.

Viscusi, W. K., & Hakes, J. (2008). Risk beliefs and smoking behaviour. *Economic Inquiry*, *46,* 45-59.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 312,* 135-138.

Wanke, M. (2002). Conversational norms and the interpretation of vague quantifiers. *Applied Cognitive Psychology, 16,* 301-307.

Watkins, M. J., & LeCompte, D. C. (1991). Inadequacy of Recall as a Basis for Frequency Knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1161-1176.

Weinstein, N. D., & Diefenbach, M. A. (1997). Percentage and verbal category measures of risk likelihood. *Health Education Research, 121*, 139-141.

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15,* 19-35.

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%--right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25,* 1514-1533.

Windschitl, P. D., & Wells, G.L. (1996). Measuring Psychological Uncertainty: Verbal Versus Numeric Methods. *Journal of Experimental Psychology: Applied, 2,* 343-364.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). 'How much is quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology, 85,* 479-496.

Yamigishi, K. (1997). When a 12.86% mortality is more dangerous than 24.14%. *Implications for Risk Communication Applied Cognitive Psychology*, *11,* 495-506.

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies*, *20*, 121-134.

Appendix 1

Questions about Smoking Risks Used in Literature Review

Questions about Smoking Risks Used in Literature Review

Gallup Poll, November 1949 (n = 1500):

- Do you think smoking is harmful or not? Yes: 60%, No: 33%, No Opinion: 7%

Gallup Poll Jan. 1954 (n = 1500):

- Have you heard or read anything recently that cigarette smoking may be a cause of cancer of the lung? Yes, have heard or read: 83%, No, have not: 17%
- What is your own opinion—do you think cigarette smoking is one of the causes of lung cancer, or not? Yes: 41%, No: 31%, No opinion: 28%

Gallup Poll Jun. 1954 (n = 1435):

- Have you heard or read anything recently to the effect that cigarette smoking may be a cause of cancer of the lung? Yes, have heard or read: 90%, No, have not: 10%
- What is your own opinion—do you think cigarette smoking is one of the causes of lung cancer, or not? Yes: 42% No: 30%, No opinion: 28%

Gallup Poll Jun. 1981 (n = 1535):

- Do you think that cigarette smoking is or is not one of the causes of lung cancer? Is: 83%, Is not: 10%, Don't Know: 7%

Risk Perception Measures used in Table 1:

Roper, 1964:

- According to the report, a person who smokes a pack or more a day has about ten times as great a chance of getting lung cancer as a non-smoker, but what does that mean to you in terms of the likelihood of the pack a day smoker getting lung cancer? Out of 100 pack a day smokers how many would you say would get lung cancer – 5 out of 100, 25 out of 100, 50, 75, 95 out of 100, or how many?

Roper, 1977, 1980:

- Out of every one hundred people who have been cigarette smokers, how many would you estimate get lung cancer at some time in their lives?

Viscusi, 1985:

- Among 100 smokers, how many of them do you think will get lung cancer because they smoke?

Viscusi, 1991:

- Among 100 smokers, how many of them do you think will die from lung cancer because they smoke?

Sutton, 1995:

- On average, out of 1000 20 year olds in Britain who smoke cigarettes regularly and who carry on smoking,..how many do you think will be killed by smoking before the age of 70?

Viscusi, 1997:

- Among 100 smokers, how many of them do you think will develop lung cancer because they smoke?

Viscusi, 1998:

- Out of 100 smokers, how many do you think will die from lung cancer because they smoke?

Annenberg 2, 1999:

- Now I would like you to imagine 100 cigarette smokers, both men and women, who smoked cigarettes for their entire adult lives. How many of these 100 people do you think will die from lung cancer?

Krosnick, 2000:

- Next, I'd like to turn to a different topic: what you personally think about the effect of cigarette smoking on people's health. I'm going to read these next two questions very slowly to let you think about each part of them, and I can repeat each question as many times as you like before you answer, so you can be sure they are clear to you. First, if we were to randomly choose one thousand American adults who never smoked cigarettes at all during their lives, how many of those one thousand people do you think would get lung cancer sometime during their lives?

Appendix 2

Target words and Exemplars for use in Study 1

Target words and Exemplars for use in Study 1

1. Weapon
    1) Gun
    2) Knife
    3) Sword
    4) Bat
    5) Bomb
    6) Fist
    7) Rifle
    8) Arrow
    9) Rope
    10) Mace
    11) Ax
    12) Grenade
    13) Missile
    14) Club
    15) Spear
    16) Bazooka

2. Fruit
    1) Apple
    2) Orange
    3) Banana
    4) Grape
    5) Pear
    6) Strawberry
    7) Peach
    8) Kiwi
    9) Pineapple
    10) Watermelon
    11) Raspberry

12) Plum
13) Grapefruit
14) Mango
15) Lemon
16) Cherry

3. City
    1) New York
    2) Los Angeles
    3) Denver
    4) Miami
    5) Paris
    6) London
    7) Toronto
    8) Moscow
    9) Charlotte
    10) Dallas
    11) Orlando
    12) Houston
    13) Seattle
    14) Boston
    15) Chicago
    16) Baltimore

4. State
    1) Oregon
    2) California
    3) Kansas
    4) Ohio
    5) Texas
    6) Maryland
    7) Utah

8) Iowa
9) Missouri
10) Oklahoma
11) Michigan
12) Georgia
13) Florida
14) Colorado
15) Idaho
16) Nevada

5. Country
    1) United States
    2) Canada
    3) Mexico
    4) France
    5) Russia
    6) Germany
    7) Iraq
    8) Chile
    9) Japan
    10) China
    11) Spain
    12) Italy
    13) Brazil
    14) Iran
    15) India
    16) Peru

6. Instrument
    1) Drum
    2) Guitar
    3) Flute

4) Piano
5) Trumpet
6) Clarinet
7) Saxophone
8) Violin
9) Trombone
10) Tuba
11) Oboe
12) Harp
13) Cello
14) Organ
15) Cymbal
16) Banjo

7. Job

1) Doctor
2) Teacher
3) Lawyer
4) Nurse
5) Fireman
6) Professor
7) Accountant
8) Dentist
9) Psychologist
10) Secretary
11) Manager
12) Cook
13) Policeman
14) Banker
15) Engineer
16) Scientist

8. Sport
    1) Football
    2) Basketball
    3) Soccer
    4) Baseball
    5) Hockey
    6) Tennis
    7) Swimming
    8) Golf
    9) Volleyball
    10) Rugby
    11) Lacrosse
    12) Running
    13) Polo
    14) Wrestling
    15) Bowling
    16) Skiing

9. Clothing
    1) Shirt
    2) Pants
    3) Socks
    4) Dress
    5) Coat
    6) Jeans
    7) Gloves
    8) Sweatshirt
    9) Blouse
    10) Undershirt
    11) Scarf
    12) Underpants

13) Shorts
14) Bra
15) Skirt
16) Jacket

10. Tool

1) Hammer
2) Nail
3) Saw
4) Screwdriver
5) Drill
6) Wrench
7) Screw
8) Ruler
9) Level
10) Pliers
11) Hoe
12) Shovel
13) Chisel
14) Rake
15) Spade
16) Trowel

11. Bird

1) Eagle
2) Robin
3) Bluejay
4) Parrot
5) Hawk
6) Cardinal
7) Crow
8) Sparrow

9) Hummingbird
10) Falcon
11) Owl
12) Dove
13) Pigeon
14) Parakeet
15) Duck
16) Chicken

12. Insect

1) Fly
2) Ant
3) Caterpillar
4) Bee
5) Mosquito
6) Beetle
7) Ladybug
8) Grasshopper
9) Butterfly
10) Wasp
11) Roach
12) Moth
13) Gnat
14) Flea
15) Cricket
16) Hornet

13. Animal

1) Dog
2) Cat
3) Lion
4) Horse

5) Bear
6) Tiger
7) Elephant
8) Cow
9) Deer
10) Mouse
11) Pig
12) Giraffe
13) Rat
14) Rabbit
15) Goat
16) Donkey

14. Fish

1) Salmon
2) Trout
3) Goldfish
4) Bass
5) Catfish
6) Tuna
7) Shark
8) Flounder
9) Swordfish
10) Herring
11) Carp
12) Cod
13) Marlin
14) Piranha
15) Halibut
16) Minnow

15. Drug

1) Marijuana
2) Cocaine
3) Heroin
4) Ecstasy
5) Alcohol
6) LSD
7) Crack
8) Acid
9) Mushrooms
10) Speed
11) Nicotine
12) Caffeine
13) Opium
14) Morphine
15) Tylenol
16) Aspirin

Appendix 3

Numeracy Test from Galesic and Garcia-Retamero (2010)

Numeracy Test from Galesic and Garcia-Retamero (2010)

1. Imagine that we flip a fair coin 1000 times. What is your best guess about how many times the coin will come up heads in 1000 flips?

   _____ times out of 1000

2. In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket to BIG BUCKS?

   _____ person(s) out of 1000

3. In the Daily Times Sweepstakes, the chance of winning a car is 1 in 1,000. What percent of tickets to Daily Times Sweepstakes win a car?

   _____ % of tickets

4. Imagine that we rolled a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?

   _____ times out of 1000

5. Which of the following numbers represents the biggest risk of getting a disease?

   1 in 100 1 in 1000 1 in 10

6. Which of the following represents the biggest risk of getting a disease?

   1% 10% 5%

7. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?

   _____ person(s) out of 1000

8. If the chance of getting a disease is 20 out of 100, this is the same as having what percentage chance of getting the disease?

   _____ % chance

9. If person A's chance of getting a disease is 1 in 100 in 10 years and person B's risk is double that, what is B's risk?

   _____

Appendix 4

Questions Used from the NSSE

Questions used from the NSSE

Vague quantifier/translation questions (Questions asked in a list, grouped under one question stem. Question stem appears as it did in the Web survey with those under it included in that grouping on one page. Not all questions used in translations that were asked this way. All questions given the response options, in order they appeared on the page, going left to right: Very Often, Often, Sometimes, Never. See example Web page below in Figure 2.)

Active and Collaborative Learning Scale

In your experience at your institution during the current school year, about how often have you…

- Asked questions in class or contributed to class discussions
- Made a class presentation
- Worked with other students on projects during class
- Worked with classmates outside of class to prepare class assignments
- Tutored or taught other students (paid or voluntary)
- Participated in a community-based project (e.g., service learning) as part of a regular course
- Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Student-Faculty Interaction Scale

In your experience at your institution during the current school year, about how often have you…

- Discussed grades or assignments with an instructor
- Talked about career plans with a faculty member or advisor
- Received prompt written or oral feedback from faculty on your academic performance
- Worked with faculty members on activities other than coursework (committees, orientation, student life activities, etc.)
- Discussed ideas from your readings or classes with faculty members outside of class

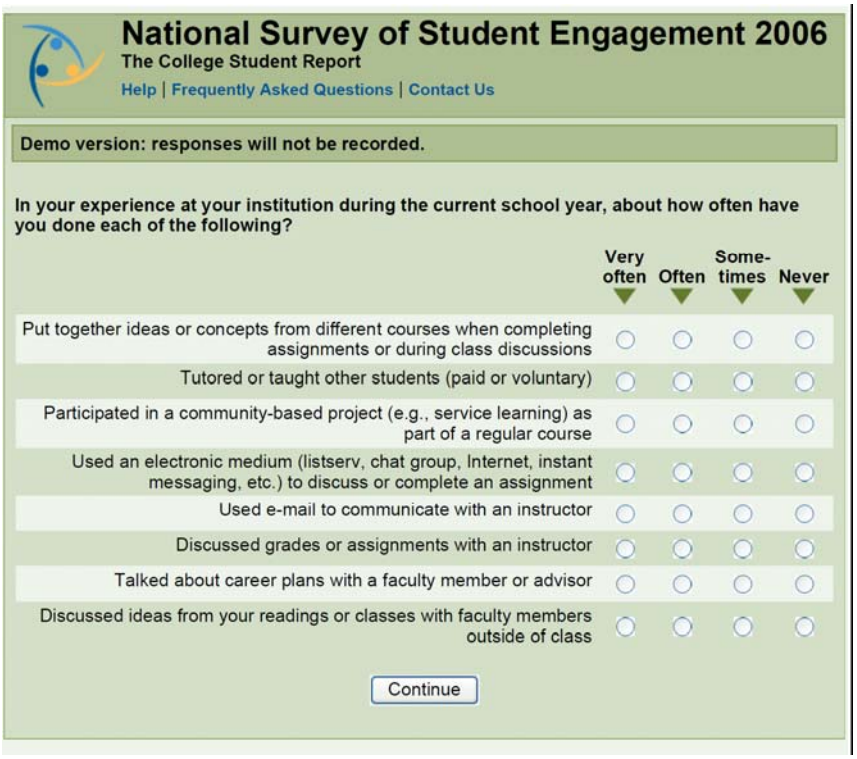

Figure 2. Example Screen of How Vague Quantifier Questions Asked.

Theoretically Related Variables of Interest

How would you evaluate your entire educational experience at this institution?

Poor

Fair

Good

Excellent

If you could start over again, would you go to the *same institution* you are now attending?

Definitely no

Probably no

Probably yes

Definitely yes

What have most of your grades been up to now at this institution?

C- or lower

C

C+

B-

B

B+

A-

A

Appendix 5

Questions Used from Annenberg 2 Surveys.

Questions used from Annenberg 2 Surveys.

Risk Perception Measures

- Now I would like you to imagine 100 cigarette smokers, both men and women, who smoked cigarettes for their entire adult lives. How many of these 100 people do you think will die from lung cancer?
- I just asked you about smokers. Now I would like you to imagine 100 non-smokers, both men and women, who never smoked and don't live with smokers. How many do you think will die from lung cancer?
- In your opinion, would smoking everyday be very risky for your health, somewhat risky, a little risky or not at all risky for your health?

Theoretically Related Variables of Interest

How frequently did you smoke cigarettes in the past 30 days? Just tell me when I get to the right amount. (Read responses 1-7:)

1 Less than one cigarette a day
2 One to five a day
3 Six to ten a day
4 Eleven to fourteen a day
5 Fifteen to nineteen a day
6 Twenty a day
7 More than twenty a day
8 (Don't know)
9 (Refused)

The above variable will be recoded to be such that anyone smoking one or more cigarettes a day is considered a smoker.