Public Opinion Quarterly, Vol. 78, No. 3, Fall 2014, pp. 591–633

RESPONSE HEAPING IN INTERVIEWER-ADMINISTERED SURVEYS
IS IT REALLY A FORM OF SATISFICING?

ALLYSON L. HOLBROOK*
SOWMYA ANAND
TIMOTHY P. JOHNSON
YOUNG IK CHO
SHARON SHAVITT
NOEL CHÁVEZ
SAUL WEINER

Abstract Response heaping (also referred to as rounding or digit preference) occurs when respondents show a preference for rounded numbers (often those divisible by five or 10). Conventional wisdom is that this is the result of taking cognitive shortcuts to make question answering easier, and as such, that it may be a form of survey satisficing. In four studies, we test this conventional wisdom for the first time by exploring whether response heaping occurs for five types of survey questions (behavioral frequency questions, questions that ask about an

Allyson L. Holbrook is an associate professor of public administration and psychology and affiliate faculty member at the Survey Research Laboratory of the University of Illinois at Chicago, Chicago, IL, USA. Sowmya Anand is a project coordinator at the Survey Research Laboratory of the University of Illinois at Urbana-Champaign, Champaign, IL, USA. Timothy P. Johnson is a professor of public administration and director of the Survey Research Laboratory at the University of Illinois at Chicago, Chicago, IL, USA. Young Ik Cho is an associate professor in the Zilber School of Public Health at the University of Wisconsin–Milwaukee, Milwaukee, WI, USA. Sharon Shavitt is the Walter H. Stellner Professor of Marketing, professor of psychology, and research professor at the Survey Research Laboratory and the Institute of Communications Research at the University of Illinois at Urbana-Champaign, Champaign, IL, USA. Noel Chávez is an associate professor of community health sciences and codirector of the Maternal and Child Health Program at the School of Public Health of the University of Illinois at Chicago, Chicago, IL, USA. Saul Weiner is professor of medicine, pediatrics, and medical education at the University of Illinois at Chicago and staff physician at the Jesse Brown VA Medical Center, Chicago, IL, USA. An earlier version of this paper was presented at the 2010 annual meeting of the American Association for Public Opinion Research. This work was supported by grants from the National Science Foundation [0648539 to A.L.H., T. P. J., Y.I.C., S.S., N.C., and S.W.]; and the National Institutes of Health [1RO1HD053636-01 to A.L.H., T. P. J., Y.I.C., S.S., N.C., and S.W.].
*Address correspondence to Allyson Holbrook, 412 S. Peoria St., Sixth Floor, Chicago, IL 60607, USA; e-mail: allyson@uic.edu.

doi:10.1093/poq/nfu017 Advance Access publication August 13, 2014

individual's personal characteristics, questions that ask about an individual's age at the time of an event, questions that ask the respondent to report a percentage, and feeling-thermometer attitude reports) under the conditions thought to foster survey satisficing (e.g., among respondents lower in ability and motivation, when the task of question-answering is difficult, and later in a long questionnaire) and whether heaped responses show effects of survey satisficing (e.g., shorter response latencies, less accuracy, and lowered predictive validity). We also examine the prevalence of response heaping and the extent to which heaping is associated across questions. Heaping above chance levels was found for most types of questions (although the prevalence of heaping varied systematically across different types of questions), but we found little evidence that heaping for most types of questions is more common under conditions thought to foster satisficing. In fact, heaping for some questions may actually reflect more thoughtful processes and result in higher data quality.

Survey questions often ask respondents to give integer numeric responses. Response heaping (larger than expected proportions of rounded answers, such as those divisible by five)[1] has been observed in reports of age (Zelnik 1964), food expenditures (Battistin, Miniaci, and Weber 2003), wages and hours worked (Hirsch 2005), job tenure (Diebold, Neumark, and Polsky 1997), time allocation (Hobson 1976), dates of events and how long ago events occurred (Huttonlocher, Hedges, and Bradburn 1990), frequency of hunting trips (Vaske and Beaman 2006), number of days spent fishing (Tarrant and Manfredo 1993), and weight (Krueger et al. 2004). Heaping is a concern because for most variables, heaping at numbers divisible by five indicates that some of these responses are inaccurate (Battistin, Miniaci, and Weber 2003).

What Leads to Heaping?

Heaping for questions about the frequency of behaviors may be more likely when respondents use estimation processes (Burton and Blair 1991) and less likely when they use counting or episodic enumeration (e.g., Conrad, Brown, and Cashman 1998). Estimation processes are particularly likely when the number of behaviors is high (which could vary as a function of both the respondent and the question) and for questions about long time frames and events that are similar and occur regularly (Burton and Blair 1991), suggesting that heaping

1. Researchers also have observed heaping at other values when relevant (e.g., heaping at six-month intervals for time estimates).

should be more likely under these conditions. Estimation processes may also be more likely for questions about events for which respondents do not have episodic memories, such as proxy reports (Barbieri and Hertrich 2005; West, Robinson, and Bentley 2005).

Research examining demographic correlates of heaping suggests that it may be more likely among low-income (Myers 1976) and illiterate respondents (Budd and Guinnane 1991) and more prevalent in less modernized countries (Nagi, Stockwell, and Snavley 1973). There is also some evidence that women heap more than men (Boyle and Gráda 1986), that non-Whites heap more than Whites (Coale 1955), and that heaping varies by culture and country (Barbieri and Hertrich 2005).

Heaping as a Response Strategy

Research on conversational logic in survey interactions suggests that rounded responses communicate uncertainty (Zhang and Schwarz 2012). People may also heap because it eases the task of answering questions. As Schaeffer and Presser (2003, 68) argued:

> "Estimation strategies lead to heaping at common numbers, such as multiples of 5 or 10. Many of these strategies can be considered techniques for 'satisficing,' that is, for conserving time and energy and yet producing an answer that seems good enough for the purposes at hand."

Similarly, Tourangeau, Rips, and Rasinski (2000, 254) concluded that

> "…respondents are likely to attempt to make the task of reporting their answer as easy as they can. They will use ranges or round values to report numeric quantities…To the extent that they are tired, uninterested, or generally unable to cope with the demands of the items, respondents will be more prone to use strategies that reduce the burden that the questions impose."

If heaping is a cognitive shortcut more likely to occur among respondents who cannot or do not want to answer the questions carefully, heaping may be a form of survey satisficing (Walejko 2010).

Satisficing Theory

Satisficing theory (Krosnick 1991) suggests that the process of answering questions accurately involves substantial cognitive work. To "optimize," respondents must interpret the question meaning, retrieve relevant information from memory, integrate that information into a judgment, and report that judgment. Some respondents may become fatigued and less motivated to optimize

as they go through the survey. Others may be unwilling or unable to go through the steps necessary to optimize. These respondents may satisfice by looking for ways to give satisfactory answers to survey questions without going fully through all the steps necessary to optimize.

Satisficing is more likely among people with less ability (e.g., those with fewer cognitive skills, such as those low in educational attainment) or less motivation (e.g., those less interested in the survey) and when the response task is difficult (e.g., for questions that ask about novel stimuli; see Narayan and Krosnick [1996]; Krosnick [1999]). If heaping is a form of survey satisficing, it should be more likely under these conditions, and heaped responses may also be provided more quickly and be associated with greater measurement error than nonheaped responses.

The Current Research

We evaluate whether heaping is the result of survey satisficing in four interviewer-administered surveys. We first compared heaping across several types of questions (studies 1, 2, and 3). Second, we tested whether heaping across items is related (studies 1–4). If heaping is a form of satisficing, respondents who heap on one question should be more likely to do so on other questions. Third, we tested whether heaping is greater among respondents low in ability and motivation and for difficult questions (studies 1–4). Fourth, we tested whether heaped responses were less accurate (study 2) and had lower predictive validity (study 4) than unheaped responses and whether they were reported faster than unheaped responses (studies 2 and 3). Finally, our studies allow us to try to replicate past findings that heaping in behavioral frequency questions is less likely when respondents count behaviors than when they use estimation strategies (study 1), more likely when the number of events being recalled is larger (studies 1, 2, and 3), and more likely for proxy reports than for self-reports (study 2).

Study 1

In this study, we examined the prevalence of heaping across four types of questions, whether the tendency to heap was clustered within respondents across questions (a response style), and whether heaping was greater among respondents low in cognitive skills. We also tested whether heaping is lower for behavioral frequency questions when respondents use counting strategies rather than estimation strategies and when the number of events is large.

METHODS

Respondents: Respondents were 423 adults from Chicago, selected so that there were approximately equal numbers of African Americans, Mexican Americans, Puerto Ricans, and non-Hispanic Whites. Respondents within each group were stratified by sex, age, and education.

Procedures: The data were collected July 1993–May 1994. Respondents were recruited from the community via newspaper ads and flyers. Those screened as eligible (in the selected race/ethnic groups) via telephone were invited to the University of Illinois at Chicago Survey Research Laboratory (UIC-SRL) to complete an interview of approximately 50 health-related and demographic questions, followed by probes asking about their cognitive processing of each question. Respondents also completed a brief self-administered questionnaire. The interviews averaged an hour.

Measures: Respondents were asked to answer nine behavioral frequency questions (BFQs), two personal characteristic questions (PCQs), two questions about their age at the time of an event (AEQs), and one question about a percentage (PERCQs). For each, a heaping variable was computed: coded 1 if the response was divisible by five and 0 if not divisible by five.[2] For seven of the nine BFQs, a probe was included to assess the cognitive process by which respondents answered (counting or some other process). Education, a proxy for cognitive skills, was used as a measure of respondent ability (Narayan and Krosnick 1996).

For two of the BFQs, respondents were asked to choose one of two time frames when answering how often "per month or week" they perform a behavior. This was not designed for studying heaping, but was based on the expected range of responses (some respondents would do the behavior frequently and others infrequently). This variable was not experimentally manipulated, but our analyses controlled for the time frame selected because it may influence heaping. See appendix A for question wordings and coding.

ANALYSIS

We examined the proportion of heaped responses to each question and compared this proportion to 20 percent (the proportion of responses one would expect to be divisible by five by chance if no heaping occurred). Second, we used the Stata xtlogic command to estimate a series of multilevel models predicting heaping as the dependent variable with questions clustered within respondent (see appendix B for more information about these analyses).[3]

2. Values of 0 were not coded as heaping. We also did not count the highest response as heaping when there was an upper boundary (e.g., 100 for percentage questions), so as not to confound heaping with extreme response style (e.g., Bachman and O'Malley 1984).

3. Because the data are cross-classified by respondents and questions (i.e., questions are nested within respondent and respondents are nested within question), we also estimated cross-classified models for all the multilevel models reported. These analyses simultaneously account for both clustering within respondent and within question. However, given the number of question characteristics in the models and the very limited number of questions in each of the models, there was not enough power to estimate random variance between questions, and the results of cross-classified models were essentially not different from those reported. Therefore, we do not report the results of the cross-classified models.

RESULTS

Prevalence of response heaping: The proportion of heaped responses was significantly greater than 20 percent (chance level) for four of the nine BFQs, for the single PERCQ, and for the two PCQs (table 1). Neither of two AEQs had heaping levels above 20 percent.

Intraclass correlations: Only the rho for PCQs was significantly different from 0 (likelihood-ratio test of rho = 0: BFQs: rho = .02, chibar2(01) = 1.01, p = .16; PCQs: rho = .17, chibar2(01) = 4.90, p = .01; AEQs: rho = .16, chibar2(01) = 1.91, p = .08; see the bottom row of table 2), and both PCQs were concerned with respondent's weight. This suggests that there was a significant within-respondent tendency to heap only for PCQs.

Predictors of heaping: Confirming the evidence from table 1 that the prevalence of heaping varied by question type, PCQs (lrc[4] = 3.15, p < .001), PERCQs (lrc = 3.21, p < .001), and AEQs (lrc = .34, p = .02) showed significantly greater heaping than BFQs (see column 1 of table 2). Across all question types, education was unassociated with heaping. Consistent with Burton and Blaire (1991), the response value was positively associated with heaping (lrc = .62, p < .001) and respondents who used a process other than counting for behavioral frequencies showed significantly greater heaping than those who used counting (lrc = 1.85, p < .001).

Main effects in the overall model were qualified by several interactions between question type and other predictors. A significant interaction emerged between the dummy variable for less than a high school education and the dummy variable for PERCQs (lrc = –1.84, SE = .62, p = .003), such that the effect of the education variable was negative and significant for PERCQs (lrc = –1.74, p = .003; see row 1 in column 5 of table 2), but not for any of the other question types (see row 1 in columns 2–4 of table 2). The interaction between the dummy variable for high school degree and the dummy variable for PCQs (lrc = .66, SE = .30, p = .03) was also significant, such that the effect of the education dummy variable was positive and significant for PCQs (lrc = .66, p = .01) but not for the other three types of questions. The direction of the effect of education was not consistent across question type. For PERCQs, there was less heaping among lower-education respondents than among those with a college degree, but for PCQs, there was more heaping among lower-education respondents.

The interactions between response value and each of the question type dummy variables were also significant (PCQs: lrc = –.78, SE = .13, p < .001; AEQs: lrc = –1.11, SE = .13, p < .001; PERCQs: lrc = –1.95, SE = .20, p < .001). Consistent with Burton and Blair (1991), the response value was

4. Throughout the text, "lrc" indicates "logistic regression coefficient."

Table 1. Percentage of Heaped Responses in Survey Questions by Judgment Type and Study (range of responses shown in parentheses)

Type of judgment		Heaping
Behavioral frequencies		% Heaped responses
Study 1	Doctor visits in past year (0–96)	18.2
	Physical activity (1–100)	16.6
	Talking to friends or family on the phone (1–100)	31.0**
	(If smokes) number of cigarettes per day (1–80)	56.9**
	(If drinks) days in the last month *R* had a drink (0–30)	20.0
	(If smoked marijuana) times smoked (1–122,640)	43.5**
	(If had sex) sexual partners in the past 5 years (1–90)	13.8
	(If talked about HIV) times talked about HIV (0–200)	27.3*
	(If drinks) drinks per day *R* drinks (0–60)	9.4
Study 2	Days in last 30 *R*'s health was not good (0–30)	16.3
	(If *R* smokes) number of cigarettes per day (1–30)	38.4*
	Number of times thought about getting cancer (0–1,000)[a]	38.0*
	Days in last 30 *R*'s mental health was not good (0–30)	22.2
	Close friends talked to in past 30 days (0–20)[b]	12.2
	(If *R* had ever placed a bet) number of bets made in lifetime (0–1,000)[c]	61.0**
	Number of times *R* talked with proxy about cancer (0–995)	11.7
	(PROXY) Days not good health (0–30)	14.8
Study 3	Times *R* felt angry at anyone past 7 days (0–60)	9.0
	Times *R* felt angry at family member past 7 days (0–30)	3.5
Personal characteristics		% Heaped responses
Study 1	Weight in pounds (90–350)	69.1**
	Weight when *R* was 16 years old (70–250)	79.2**

(*Continued*)

Table 1. *Continued*

Type of judgment		Heaping
Personal characteristics		% Heaped responses
Study 2	Systolic blood pressure (40–210)	66.0**
	Diastolic blood pressure (20–140)	69.3**
	Weight in pounds (93–389)	63.3**
	Weight when *R* was 12 years old (30–200)[d]	83.6**
	Number of close friends (0–20)[e]	34.0*
	(PROXY): Age (18–87)[f]	23.9
	(PROXY): Weight (19–385)	80.1**
Study 3	Number of years *R* has lived in present community (0–78)	31.8**
Age at the time of an event		% Heaped responses
Study 1	(If *R* ever smoked regularly) age s/he started smoking (8–35)	16.2
	(If *R* ever drank alcohol) age s/he had first drink (3–35)	17.7
Study 2	(If *R* had ever smoked) age first thought about smoking (1–40)	19.6
	(If *R* had ever had an alcoholic drink) age had first alcoholic drink (6–60)	22.4
	Age first treated at emergency room (0–66)	20.7
	(If *R* had ever experienced a stressful event) age experienced first stressful event (3–65)	29.6**
	(If *R* had ever had a serious conflict with employer) age first time serious conflict with employer (14–66)	25.1
	(If *R* had ever placed a bet) Age first time placed a bet (2–60)	33.1**
	(PROXY: If *P* ever smoked) Age first smoked cigarette (6–40)	24.7
	(PROXY: If *P* ever drank) Age first had alcoholic drink (0–40)	22.4

(*Continued*)

Table 1. *Continued*

Type of judgment		Heaping
Percentages		% Heaped responses
Study 1	(If *R* sex in the past 5 years) percent of time condom used (0–100)	73.9**
Study 3	People willing to pay more taxes to pay for universal health care (0–100)	86.5**
	Voter turnout in the 2004 presidential election (8–96)	66.5**
Attitudes toward groups		Range % heaped responses
Study 3	14 feeling-thermometer items (all significant at $p < .001$)	56.0–90.3**
Study 4	(24 feeling-thermometer items)	
	Telephone interviews (all significant at $p < .001$)	69.6–91.2**
	In-person interviews (all significant at $p < .001$)	65.5–94.5**

NOTE.—Statistical tests assessed whether percentages significantly differed from 20 percent (the percent of responses that would be divisible by 5 by chance alone if no response heaping occurred). Percentages marked as significant indicate that significantly more than 20 percent of responses to the question were divisible by 5.

*$p < .05$; **$p < .01$

[a]There were 12 respondents who said 1,001 or more times and 24 who said "all the time." These respondents were excluded from our analysis.

[b]There were 3 respondents who said they had talked to 21 or more friends. These respondents were excluded from our analysis.

[c]There were 56 respondents who said 1,001 times or more. These respondents were excluded from our analysis.

[d]Five respondents who gave unrealistic responses (1 and 8) were excluded from our analysis.

[e]There were 16 respondents who said 21 or more friends. These respondents were excluded from our analysis.

[f]Two respondents reported imprecise answers ("in her 20s"; "in his 60s"), which were counted as heaped responses.

strongly and positively associated with heaping for BFQs (lrc = 1.19, $p < .001$; see row 4 of column 2 in table 2). The response value was also significantly but less strongly associated with heaping for PCQs (lrc = .42, $p < .001$; see row 4 of column 3 in table 2), not associated with heaping for AEQs (lrc = .10, $p = .40$; see row 4 of column 4 in table 2), and negatively associated with heaping for PERCQs (lrc = –.75, $p < .001$; see row 4 of column 5 in table 2).

Table 2. Multilevel Model Predicting Heaping for Study 1 (standard errors in parentheses)

Predictor	All Items (ALL)	Behavioral Frequencies (BFQ)	Personal Characteristics (PCQ)	Age for An Event (AEQ)	Percentages (PERCQ)
Respondent ability[a]					
Education					
Less than HS degree	.14	.04	.29	.16	**–1.74****
	(.15)	(.19)	(.30)	(.40)	(.58)
HS degree	.15	–.02	**.66***	.06	–.36
	(.13)	(.16)	(.27)	(.35)	(.45)
Some college	.06	.002	.21	–.004	–.27
	(.13)	(.16)	(.25)	(.35)	(.45)
Replication variables					
Response value	.62**	1.19**	**.42****	**.10**	**–.75****
	(.04)	(.08)	(.11)	(.12)	(.18)
Count[b]					
Process other than counting Used	2.08**	1.85**	NA	NA	NA
	(.20)	(.21)			
No follow-up probe	.29	.13	NA	NA	NA
	(.21)	(.22)			
Time frame[c]					
Longer time frame	1.00**	.86**	NA	NA	NA
	(.19)	(.20)			
Time frame not Varied	1.48**	1.37**	NA	NA	NA
	(.17)	(.17)			

(*Continued*)

Table 2. *Continued*

Predictor	All Items (ALL)	Behavioral Frequencies (BFQ)	Personal Characteristics (PCQ)	Age for An Event (AEQ)	Percentages (PERCQ)
Question type[d]					
Personal characteristics	3.15**				
	(.13)				
Age for an event	.34*				
	(.15)				
Percentages	3.21**				
	(.18)				
N (respondents)	420	420	420	398	268
N (observations)	4,509	2,782	823	636	268
Rho	.03	.0000001	.08	.16	NA
Rho (no predictors)	.000003	.02	.17	.16	NA

NOTE.—Analyses conducted controlling for gender, race/ethnicity, age, and age squared. Unstandardized coefficients shown (standard errors in parentheses) from xtlogit analysis in Stata with heaping across questions clustered within respondents. Bolded values in columns 3–5 indicate effects that were significantly different from those in column 2.

*$p < .05$; **$p < .01$

[a]At least a four-year degree was the comparison group.

[b]Respondents who indicated in response to a probe that they used a counting or episodic enumeration process were the comparison group.

[c]Respondents who selected the shorter time frame when given the option were the comparison group.

[d]Behavioral frequencies were the comparison group.

CONCLUSION AND LIMITATIONS

These findings indicate that heaping varies systematically across question type. There was little evidence of a respondent-level tendency to heap (except for the two PCQs about respondents' weight). Replicating Burton and Blair (1991), counting was associated with less heaping than other strategies, and heaping was greater when the number of events reported was high for behavioral frequency questions.

There was little evidence of more heaping for less-educated respondents. Only PCQs showed some evidence of this: Respondents with only a high school degree showed more heaping than those with at least a four-year college degree. However, this study provided a limited opportunity to study heaping as a cognitive shortcut because education was the only measure of the conditions thought to foster satisficing. In three additional studies, we analyzed data that included many more measures of the conditions thought to foster satisficing. These studies also allowed us to test whether heaped responses would be reported faster and be less accurate than unheaped responses. Because the variables and analyses for studies 2 and 3 are very similar, we next describe the methods for these two studies and report analyses of data combined across them.

Studies 2 and 3

As in study 1, studies 2 and 3 examined the prevalence of heaping in a variety of question types and tested whether heaping was clustered across questions within respondent. Furthermore, we tested whether heaping was more common among respondents low in ability (e.g., those with less education and those rated as less intelligent by interviewers), respondents low in motivation (e.g., those low in the need for cognition or motivation to think, those less interested in the survey, and those who reported that they put less effort into answering questions), and respondents who rated the question answering process as more difficult. We also tested the hypothesis that heaping would be greater for questions asked later in a survey interview. Further, these data allowed us to test whether heaping was associated with faster response latencies and less accurate responses to PCQs measuring blood pressure and weight.

STUDY 2 METHODS

Respondents: Respondents were 603 adults from Chicago. Approximately equal numbers of non-Hispanic Whites, Mexican Americans, African Americans, and Korean Americans were recruited. Among Mexican Americans and Korean Americans, half of the interviews were in English and half were in Spanish and Korean, respectively.

Procedure: The data were collected July 2009–June 2010. Respondents were recruited via screening telephone samples (for membership in the

relevant race/ethnic categories), and in the case of Korean Americans, snowball sampling and electronic ads. They were invited to complete a 90-minute interview at UIC-SRL and offered a $40 incentive to participate in an interviewer-administered computer-assisted personal-interview (CAPI) survey composed of approximately 270 health items about themselves and (when possible) proxy reports for a member of their household. They also completed a self-administered questionnaire that included respondent demographics. Finally, respondents were offered additional incentives to allow the interviewer to measure their height, weight, and blood pressure. A response rate of 9.7 percent was estimated for the purchased sample (AAPOR response rate 3).[5] This response rate represents the proportion of eligible numbers that resulted in completed interviews in the laboratory (i.e., in order to be counted as complete, an eligible respondent had to go through the screener, make an appointment to come to the lab, and complete the interview and questionnaires at the lab).

Measures: Respondents were asked six BFQs about themselves, one about a selected proxy, and one about the frequency of a specific type of interaction between the respondent and the selected proxy. Respondents also were asked five PCQs about their own characteristics, two PCQs about characteristics of the selected proxy, six AEQs about themselves, and two AEQs about the proxy. For each of these 23 questions, a heaping indicator was coded as in study 1.

This survey also included two measures of respondent ability (education and interviewer rating of respondent intelligence), five variables associated with respondent motivation (need for cognition, self-reported respondent effort, interviewer rating of respondent interest in the interview, and where in the survey interview the question was asked—sections of the questionnaire were rotated across respondents).[6] It also included a measure of the perceived difficulty of the question-answering task. See appendix A for specific question wordings and coding.

5. The goal of this study was not to draw a representative sample, but to recruit individuals from each of the four racial and ethnic groups. In order to maximize the number of individuals in one or more of these groups, samples were purchased in strata. Strategies included purchasing numbers from geographic areas with high numbers of these individuals, purchasing samples from neighborhoods near UIC to minimize respondent burden to travel to the lab, and purchasing samples based on last names in order to find Korean respondents. Korean Americans were also recruited as part of a convenience sample. This response rate is difficult to interpret because not all numbers in the sample were contacted the maximum number of times before being finalized. In addition, greater efforts were put into contacting and gaining cooperation from households in the Korean sample than samples in the other strata. The reported response rate includes Korean cases where we attempted to interview people in the household other than the selected respondent and cases where selected respondents or households recommended other respondents or households to be interviewed (snowball sampling). This response rate estimate excludes 99 Korean American convenience sample cases.
6. Interviewers are routinely asked to rate respondents' intelligence and interest, and no special training was provided to interviewers on how to do these ratings. Interviewers did not report difficulty with the ratings.

Response latencies (amount of time to answer a question) were assessed for all questions used to measure heaping using a modified version of Bassili's (1996) procedure. We transformed this variable by taking the inverse as recommended for response latency data (see Fazio 1990) and standardized the variable for each question.

STUDY 3 METHODS

Respondents: Respondents were 400 adults from Chicago. The sample was stratified by race, ethnicity, and language.

Procedure: The data were collected August 2008–April 2010. Respondents were recruited via screening telephone samples, as before. Eligible respondents were invited to complete a 90-minute interviewer-administered CAPI survey (comprising approximately 100 social and political items) at UIC-SRL and offered a $40 incentive. They also completed the same self-administered questionnaires as in study 2. During the CAPI survey, response latencies were measured for all but one of the questions used to assess heaping.[7] A response rate of 7.6 percent was estimated for the purchased sample (AAPOR response rate 3), as was done in study 2.[8]

Measures: In this survey, there were two BFQs, one PCQ, and two PERCQs. Respondents also were asked to evaluate 14 groups in society on 14 101-point feeling-thermometer questions (FTQs). For each question, a heaping indicator was coded as in study 1. This survey included the same measures of the conditions thought to foster satisficing as study 2 (see appendix A). Response latencies were assessed as in study 2.

ANALYSIS

As in study 1, we examined the proportion of heaped responses to each question and compared it to 20 percent. We used the Stata xtlogic command to estimate a series of multilevel models predicting heaping as the dependent variable, with questions clustered within respondent. Additionally, we used the Stata xtreg command to estimate a series of multilevel models predicting reciprocalized response latencies as the dependent variable (see appendix B for more information about these analyses). Finally, for reports of systolic and diastolic blood pressure and weight, we compared the error (i.e., the absolute

7. Response latencies were not measured for the question asking the age of another adult in the household.

8. The same caveats about the response rate from study 2 apply here. This response rate estimation excludes 4 Korean American convenience sample cases and 2,649 cases that were released in error. The response rate is 5.9 percent if the latter cases are included in the response rate estimation.

difference between self-reports and measures taken by the interviewer) in respondents' self-reports for heaped and unheaped responses. Sample sizes for these accuracy comparisons were smaller than the full sample because not all respondents agreed to allow biophysical measures to be taken after the interview or reported their weight or blood pressure the last time it was taken.

RESULTS

Prevalence of response heaping: Three of the 10 BFQs had significantly more than 20 percent heaped responses (rows 10–19 of table 1). Seven of the eight PCQs had significantly more than 20 percent heaped responses (rows 22–29 of table 1). Only two of the eight AEQs showed heaped responses significantly greater than 20 percent (rows 32–39 of table 1). Both of the PERCQs (rows 31–32 of table 1) and all the FTQs (row 33 of table 1) showed significantly greater than 20 percent heaped responses. Although there was variation in the amount of heaping observed for different question types, all question types showed heaping greater than would be expected by chance alone.

Intraclass correlations: The rho statistic across question types was moderate (rho = .19) and significantly different from 0 (chibar2(1) = 1,470.05, $p < .001$). Rho was significantly different from 0 for BFQs (rho = .05, chibar2(1) = 8.28, $p = .002$), AEQs (rho = .08, (chibar2(1) = 19.96, $p < .001$), PERCQs (rho = .14, chibar2(1) = 2.60, $p = .05$), and FTQs (rho = .48, chibar2(1) = 1,026.98, $p < .001$), but not for PCQs (rho = .01, chibar2(1) = .27, $p = .30$; see the bottom row of table 3, suggesting that there was a within-respondent tendency to heap for all question types except PCQs (this tendency was particularly strong for FTQs).

Predictors of heaping: As in study 1, different types of questions demonstrated different levels of heaping. Replicating study 1, PCQs (lrc = 1.51, $p < .001$) and PERCQs (lrc = 4.01, $p < .001$) showed significantly greater heaping than did BFQs. Unlike study 1, AEQs did not show more heaping than BFQs (lrc = –.08, $p = .18$). FTQs also showed significantly greater heaping than BFQs (lrc = 4.05, $p < .001$). Across all question types (see column 1 of table 3), none of the dummy variables representing education were significantly associated with heaping. However, interviewer ratings of respondent intelligence were positively and significantly associated with heaping (lrc = .48, $p = .003$), suggesting that heaping was actually more likely among respondents rated as more intelligent. Two of the variables associated with respondent motivation also were associated with heaping. Respondents' self-reported effort was negatively associated with heaping, such that greater effort was associated with less heaping (lrc = –.81, $p < .001$). Interviewer ratings of respondents' interest in the survey were also marginally significantly related to heaping (lrc = –.20, $p = .06$), such that greater interest also was associated with less heaping. As in study 1, greater response values were associated with more heaping (lrc = .17, $p < .001$). Somewhat unexpectedly, proxy judgments showed less heaping than non-proxy judgments (lrc = –.40, $p < .001$).

However, these main effects were qualified by several significant interactions between question type and other predictors. The interaction between the dummy

Table 3. Multilevel Model Predicting Heaping for Studies 2 and 3 (standard errors in parentheses)

Predictor	All Items	Behavioral Frequencies	Personal Characteristics	Age for An Event	Percentages	Feeling Thermometer
Respondent ability						
Education						
Less than HS degree	–.05	–.64**	–.30*	**.13**	–.13	**.22**
	(.12)	(.19)	(.15)	(.20)	(.44)	(.33)
HS degree	–.11	–.21	–.21+	.01	.02	.05
	(.09)	(.14)	(.11)	(.15)	(.31)	(.24)
Some college	–.08	–.22+	–.07	–.03	**.82***	–.27
	(.08)	(.12)	(.11)	(.14)	(.35)	(.23)
R's intelligence (*I*'s report)	.48**	.19	.12	.34	**1.38***	**1.72****
	(.16)	(.26)	(.22)	(.28)	(.62)	(.45)
Respondent motivation						
Need for cog.	.01	.06	–.17	.10	**–.77***	.05
	(.08)	(.11)	(.10)	(.12)	(.34)	(.24)
Effort	–.81**	–.17	.05	–.35	.03	**–1.68****
	(.20)	(.30)	(.25)	(.33)	(.77)	(.57)
R's interest (*I*'s report)	–.30+	–.09	–.31	.01	.06	–.70
	(.16)	(.26)	(.21)	(.28)	(.60)	(.45)
Asked later in the questionnaire	–.09	–.49*	**–.02**	**–.06**	**.06**	**.09**
	(.06)	(.13)	(.13)	(.06)	(.22)	(.16)
Question difficulty	.22	.43	–.19	–.18	–.27	.45
	(.02)	(.34)	(.28)	(.36)	(.69)	(.51)

(*Continued*)

Table 3. *Continued*

Predictor	All Items	Behavioral Frequencies	Personal Characteristics	Age for An Event	Percentages	Feeling Thermometer
Replication variables						
Response value	.17**	1.45**	**.29****	**.34****	**–1.88***	**–3.83****
	(.02)	(.07)	(.04)	(.05)	(.20)	(.16)
Proxy judgment	–.40**	–.91**	–.48**	**.14**	NA	NA
	(.08)	(.19)	(.12)	(.17)		
Control variables						
Question type	1.51**					
Personal characteristic	(.05)					
Age for an event	–.08					
	(.06)					
Percentage	4.01**					
	(.14)					
Feeling thermometers	4.05**					
	(.12)					
Study 2	1.63**	3.46**	1.26**	NA	NA	NA
	(.11)	(.36)	(.12)			
N (respondents)	978	978	978	586	397	397
N (observations)	16,650	4,045	3,392	3,258	794	5,161
Rho	.11	.000001	.000001	.07	.00001	.30
Rho (no predictors)	.19	.05	.01	.08	.14	.48

NOTE.—Analyses also controlled for respondent gender, age, age squared, and race/ethnicity (coded via three dummy variables to represent Mexican Americans, African Americans, and Korean Americans (Whites were the control group), and language of interview (dummy variables were included for Korean and Spanish, with English as the comparison group). Unstandardized coefficients shown (standard errors in parentheses) from xtlogit analysis in Stata with heaping across questions clustered within respondents. Bolded values in columns 3–6 indicate effects that were significantly different from those in column 2.

*$p < .05$; **$p < .01$

variable for less than high school education and the dummy variable for AEQs was significant (lrc = .73, SE = .26, p = .006), as was the interaction between this education dummy variable and the variable indicating FTQs (lrc = .69, SE = .29, p = .02). Respondents with less than a high school education showed significantly less heaping than those with at least a four-year college degree for BFQs (lrc = –.64, p = .001), but not for AEQs (lrc = .13, p = .52) or FTQs (lrc = –.22, p = .51). The interaction between the dummy variable for some college education and the dummy variable for PERCQs was also significant, such that respondents with some college heaped somewhat less than those with at least a four-year degree for BFQs (lrc = .22, p = .07) but significantly more for PERCQs (lrc = .82, p = .02). The interaction between interviewer ratings of intelligence and the dummy variable for PERCQs was marginally significant (lrc = 1.14, SE = .70, p = .10), and the interaction between ratings of intelligence and the dummy variable for FTQs was significant (lrc = 1.00, SE = .39, p = .01) such that ratings of intelligence were positively associated with heaping only for these two question types (PERCQs: lrc = 1.38, p = .03; FTQs: lrc = 1.72, p < .001). Therefore, although the effects of respondent ability on heaping varied across question type, when these effects were significant, they were in the opposite direction from what one might expect if heaping is a form of satisficing—that is, more heaping among respondents with more ability.

The effects of three of the four variables associated with respondent motivation also varied by question type. Need for cognition was negative and significant only for PERCQs (lrc = –.77, p = .02); which differed significantly from the effect for BFQs (lrc = .06, p = .58; interaction between PERCQ dummy variable and need for cognition: lrc = –.86, SE = .37, p = .02). The effect of respondents' self-reported effort was negative and significant only for FTQs (lrc = –1.68, p = .003), which differed from the nonsignificant relationships for BFQs (lrc = –.17, p = .57; interaction between the FTQ dummy variable and respondent self-reported effort: lrc = –1.15, SE = .48, p = .02). Need for cognition and self-reported effort were each associated with heaping in the expected direction for one question type (although not the same question type), such that greater motivation was associated with less heaping. The effect of question location also varied by question type—but contrary to predictions, if heaping is a form of satisficing, being asked later in the questionnaire was associated with *less* heaping for BFQs (lrc = –.49, p < .001) but not for any of the other four question types (all interactions between question type dummy variables and questionnaire location were significant).[9]

As in study 1, the effect of the response value varied across question type. This effect was positive and highly significant for BFQs (lrc = 1.45, p < .001), significantly weaker but still positive and significant for PCQs (lrc = .29,

9. These findings suggest that heaping is not a form of satisficing. Consistent with this, an index of heaping across questions from the study 3 data was weakly negatively correlated with a measure of acquiescent response bias (measured as the proportion of six agree-disagree questions to which the respondent said "agree": r = –.09, N = 405, p = .07) and uncorrelated with a measure of no-opinion responding (the proportion of no-opinion responses to three questions with explicitly offered "no opinion" responses minus the proportion of such responses to three questions with this response option omitted: r = –.03, N = 405, n.s.).

$p < .001$; interaction between PCQ dummy variable and response value: lrc = –1.10, SE = .08, $p < .001$) and for AEQs (lrc = .34, $p < .001$; interaction between AEQ dummy variable and response value: lrc = –1.08, SE = .08, $p < .001$). In contrast, the relationship between the response value and heaping was negative and significant for PERCQs (lrc = –1.88, $p < .001$; interaction between PERCQ dummy variable and response value: lrc = –3.34, SE = .21, $p < .001$) and FTQs (lrc = –3.83, $p < .001$; interaction between FTQ dummy variable and response value: lrc = –5.26, SE = .16, $p < .001$), such that respondents who gave smaller responses were more likely to heap.

Finally, inconsistent with past research (Barbieri and Hertrich 2005; West, Robinson, and Bentley 2005), proxy judgments showed significantly less heaping than non-proxy judgments for BFQs (lrc = –.91, $p < .001$) and PCQs (lrc = –.48, $p < .001$) but not AEQs (lrc = .14, $p = .14$; interaction between AEQ dummy variable and proxy dummy variable: lrc = 1.03, SE = .26, $p < .001$). Proxy PERCQs and FTQs were not asked.

Response latencies and response heaping:
Across questions, heaping was negatively associated with reciprocalized response latencies (lrc = –.18, $p < .001$; row 1 of column 1 in table 4), such that heaping was associated with longer response times. However, this main effect was qualified by interactions between heaping and the dummy variable for PCQs (lrc = .48, SE = .06, $p < .001$), heaping and the dummy variable for AEQs (lrc = .29, SE = .06, $p < .001$), and heaping and the dummy variable for PERCQs (lrc = .25, SE = .11, $p = .02$). When models were estimated separately for each question type, heaping was negatively and significantly associated with reciprocalized response latencies for BFQs and FTQs, suggesting that respondents who heaped actually took longer to respond than those who didn't (row 1 of columns 2 and 6 in table 4). Heaping was unassociated with response latencies for AEQs and PERCQs (see row 1 of columns 4 and 5 in table 4). Finally, for PCQs, heaping was associated with shorter response latencies (lrc = .11, $p = .003$).

Response heaping and response accuracy:
We next compared error in self-reported systolic and diastolic blood pressure and respondent weight for heaped and unheaped responses.[10] The average absolute difference between self-reported and measured diastolic blood pressure was significantly greater for heaped responses ($x = 12.77$, SD = 13.83) than for nonheaped responses ($x = 8.13$, SD = 7.19; $t(207) = 2.07$, $p = .04$), but this difference was not significant for systolic blood pressure (heaped $x = 20.26$, SD = 18.61; unheaped $x = 17.40$, SD = 15.03; $t(221) = 1.14$, $p = .26$). The average absolute difference between self-reported and measured

10. There is strong evidence that average blood pressure is stable, both from tracking (Jyothinagaram et al. 1990) and from twin studies (Hottenga et al. 2005).

Table 4. Multilevel Model Predicting Reciprocalized Response Latencies for Studies 2 and 3 (standard errors in parentheses)

Predictor	All Items	Behavioral Frequencies	Personal Characteristics	Age for An Event	Percentages	Feeling Thermometer
Heaped response	–.18**	–.32**	.11**	–.07	.09	–.32**
	(.02)	(.05)	(.04)	(.05)	(.12)	(.04)
Respondent ability						
Education						
Less than HS degree	–.07	–.13	–.27**	–.13	–.09	.18
	(.06)	(.09)	(.10)	(.11)	(.22)	(.12)
HS degree	.02	–.01	–.08	.01	–.08	.13
	(.05)	(.07)	(.07)	(.08)	(.15)	(.08)
Some college	.02	–.05	–.12+	.02	.11	.16+
	(.04)	(.06)	(.07)	(.07)	(.14)	(.08)
R's intelligence (*I*'s report)	–.09	–.24+	–.17	–.21	.35	.14
	(.09)	(.13)	(.13)	(.15)	(.28)	(.16)
Respondent motivation						
Need for cog.	.05	.09	–.02	–.08	.12	.07
	(.04)	(.06)	(.06)	(.06)	(.15)	(.08)
Effort	–.18	–.05	–.06	.01	–.001	–.52**
	(.11)	(.15)	(.16)	(.17)	(.36)	(.20)
R's interest (*I*'s report)	–.04	–.15	.02	.03	–.89**	–.13
	(.09)	(.13)	(.13)	(.15)	(.27)	(.16)
Asked later in the questionnaire	.06*	.03	.02	.08	.05	.04
	(.02)	(.05)	(.06)	(.08)	(.10)	(.06)

(*Continued*)

Table 4 *Continued*

Predictor	All Items	Behavioral Frequencies	Personal Characteristics	Age for An Event	Percentages	Feeling Thermometer
Question difficulty	–.29**	–.64**	–.16	–.52**	–.51	–.24
Control variables	(.11)	(.16)	(.17)	(.19)	(.33)	(.18)
Response value	–.09**	–.17**	–.10**	–.12**	.10+	–.04*
	(.01)	(.02)	(.02)	(.02)	(.05)	(.02)
Proxy judgment	–.03	–.04	–.01	–.05	NA	NA
	(.03)	(.06)	(.06)	(.08)		
Question type						
Personal characteristics	.01					
	(.03)					
Age for an event	.01					
	(.02)					
Percentage	.12*					
	(.06)					
Feeling thermometers	.11**					
	(.04)					
Study 2	–.002	.10	.11+	.64**	NA	NA
	(.05)	(.06)	(.06)	(.24)		
N (respondents)	978	923	919	577	326	397
N (observations)	12,824	2,924	2,451	2,471	474	4,504
Rho	.17	.14	.25	.14	.23	.25
Rho (no predictors)	.18	.14	.21	.17	.22	.28

NOTE.—Analyses also controlled for respondent gender, age, age squared, and race/ethnicity (coded via three dummy variables to represent Mexican Americans, African Americans, and Korean Americans (Whites were the control group), and language of interview (dummy variables were included for Korean and Spanish, with English as the comparison group). Regression coefficients shown (standard errors in parentheses) from xtreg analysis in Stata with latencies across questions clustered within respondents.

+$p < .10$; *$p < .05$; **$p < .01$

weight was significantly greater for heaped responses (x = 7.47, SD = 10.01) than for nonheaped responses (x = 5.56, SD = 7.60; $t(411) = 2.08$, $p = .04$). These findings suggest that error in responses was greater for heaped than for nonheaped responses for two of the three personal characteristics where accuracy could be assessed.

CONCLUSIONS AND LIMITATIONS

These analyses demonstrated that heaping occurs for FTQs and is more prevalent for FTQs than for BFQs. These findings replicate study 1's finding that PCQs and PERCQs show more heaping than BFQs. They also replicate the association between the response value and heaping for BFQs, PCQs, and PERCQs. Only the findings regarding AEQs did not replicate across studies.

As in study 1, there is little evidence that heaping occurs under conditions thought to foster satisficing. Heaped responses for some of the PCQs were less accurate than nonheaped responses. For PCQ questions, heaping was also associated with faster responses, suggesting that heaping may sometimes allow respondents to answer faster but with greater error. In contrast, for BFQs and FTQs, heaping was associated with longer response latencies.

Indeed, FTQs were different from other question types in a number of ways. In addition to being associated with longer response latencies, the clustering of heaping across questions within respondent was quite strong. There was also weak evidence that heaping for FTQs was associated with both high ability (interviewer rating of respondent intelligence) and effort (respondent self-reported effort). These findings raise the possibility that heaping for responses to attitude items such as FTQs may be qualitatively different than heaping in responses to other question types. Specifically, heaping for these questions may result from an effortful strategy employed by respondents to make sense of the choice among an overwhelming 101 possible responses.

These findings are limited, however, because this study's FTQs were presented with a show card with some of the points labeled (see figure 1). Most respondents pick points that are labeled (Alwin and Krosnick 1991), and most of the labeled responses are heaped responses, so heaping may be influenced by show-card presentation. This limitation was addressed in study 4, where half of respondents were asked FTQs without a show card.

Study 4

This study involved parallel face-to-face and telephone surveys and included FTQs about groups in society and candidates for president. Respondents interviewed face-to-face were asked these questions with a show card, as in study 3. Respondents interviewed via telephone did not see show cards, although verbal labels were provided for 0, 50, and 100. As in prior studies, we examined prevalence of heaping, whether it was a response style, and whether heaping

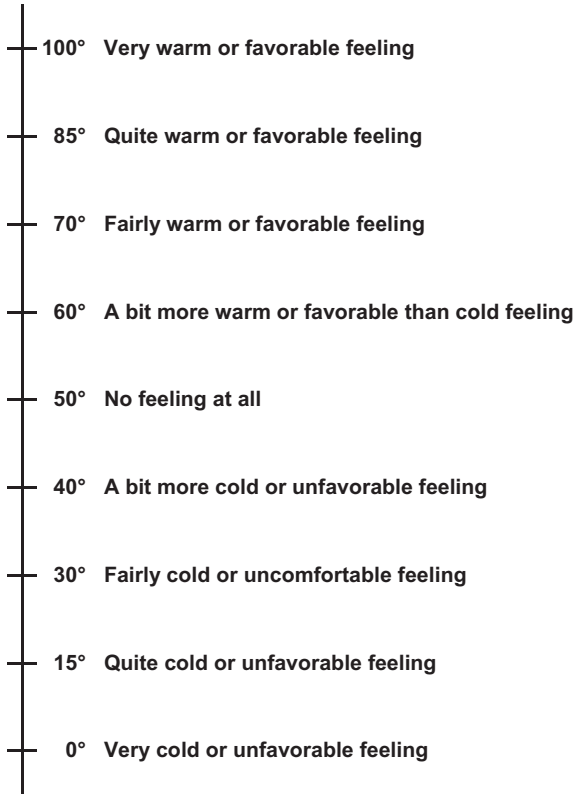

Figure 1. Feeling-Thermometer Showcard.

was more common among respondents low in ability and motivation. We also used FTQs about candidates to test the hypothesis that heaped responses would be lower in predictive validity than nonheaped responses.

METHODS

Data from the 2000 American National Election Studies survey (conducted September–November 2000) were used. This included an area probability sample with 1,006 face-to-face interviews (response rate = 64.8 percent) and an RDD sample with 801 telephone interviews (response rate = 57.2 percent). The ANES calculates the response rate as the ratio of completed interviews to the total number of potential respondents, which is similar to AAPOR response rate 3 (more methodological details can be found at http://electionstudies.org/studypages/2000prepost/2000prepost.htm). After the election, 1,555 respondents were reinterviewed during November–December 2000 (693 face-to-face

and 862 telephone). Our analysis is restricted to 693 interviewed face-to-face and 693 interviewed via telephone both pre- and postelection.

Measures: Respondents were asked FTQs about the Republican and Democratic candidates for president and 24 groups in society. For each FTQ, a heaping indicator was coded as previously. Respondent ability was measured using education, interviewer ratings of respondent intelligence, and two measures of political knowledge (the percentage of 14 knowledge questions a respondent answered correctly and interviewer ratings of respondent knowledge). This survey also included several measures of respondent motivation (e.g., need for cognition and interviewer ratings of respondent interest in the survey) as well as data about contact attempts and refusals, which were used to create two variables identifying "reluctant" respondents who might not be as motivated to optimize (Wedeking and Miller 2004).

To examine predictive validity, heaping indicators were coded for FTQs about the two major presidential candidates (Gore and Bush). During the preelection survey, respondents saying they intended to vote were asked for whom they planned on voting, and two variables, indicating intention to vote for Gore and for Bush, were created. During the postelection survey, respondents were asked for whom they had voted, and two variables, indicating voting for Gore and for Bush, were created (see appendix A).

ANALYSIS

We first tested whether the proportion of respondents who gave heaped values was significantly greater than 20 percent, whether heaping varied by mode, and the extent to which heaping across questions was clustered within respondent. We also used the xtlogit command in Stata to estimate a multilevel model in which heaping was regressed on measures of respondent ability and motivation, survey mode, and control variables (e.g., the response value given by the respondent; see appendix B for more information about these analyses). Finally, we tested whether the validity of FTQs toward the presidential candidates in predicting preelection intentions to vote for the candidate and actual voting for the candidate (postelection) varied for heaped and nonheaped responses. This was done by regressing the dependent variable (vote intention or actual voting for a candidate) on ratings of the candidate, the heaping indicator for the candidate's FTQ, and their interaction.

RESULTS

Prevalence of response heaping: All the FTQs showed significant heaping for both telephone and in-person interviews (last 2 rows of table 1). The mean number of heaped responses for telephone respondents (x = 19.4, SE = 5.70) and in-person respondents (x = 19.8, SE = 5.11) were not significantly different

($t(1,384) = 1.40$, $p = .16$). Mode was included as a control variable in subsequent analyses.

Intraclass correlation: The intraclass correlation was very large and significantly different from 0 (rho = .79, chibar2(01) = 23,000, $p < .001$), suggesting that there was a strong respondent-level heaping tendency.

Predictors of heaping: Table 5 shows a multilevel model predicting heaping across FTQs. All dummy variables capturing education were significant, such that lower-education respondents heaped *less* than those with at least a four-year college degree (less than high school degree: lrc = –1.10, $p < .001$; high school degree: lrc = –.71, $p < .001$; some college: lrc = –.38, $p = .02$). This is the opposite of what one might expect if heaping is a form of satisficing. None of the other variables associated with respondent ability or motivation was associated with heaping (see rows 4–10 in table 5). When controlling for these other variables, mode had a significant effect, such that there was greater heaping on the telephone than in person (lrc = .24, $p = .05$). Response value was also negatively and significantly associated with heaping (lrc = –.74, $p < .001$).

Response heaping and predictive validity: Heaped responses from FTQs predicted behavioral intentions and vote choice more strongly than nonheaped responses from FTQs in all four analyses (interaction between heaping indicator and FTQ response: Bush behavioral intentions: coefficient = .05, SE = .01, $p < .001$; Bush vote choices: coefficient = .06, SE = .01, $p < .001$; Gore behavioral intentions: coefficient = .03, SE = .01, $p = .004$; Gore vote choice: coefficient = .03, SE = .01, $p = .006$). Thus, heaped responses to these FTQs showed greater predictive validity than unheaped responses.

DISCUSSION AND CONCLUSION

Our research is among the first to systematically examine heaping and the processes that lead to heaping across a variety of question types. We observed high levels of heaping for percentage questions, personal characteristic questions, and feeling-thermometer questions, whereas behavioral frequency questions and questions that asked respondents their age at the time of an event showed lower levels of heaping. It is particularly interesting that behavioral frequency questions showed relatively low levels of heaping, because much of the research on the processes that lead to heaping has focused on these questions.

Consistent with past research (e.g., Burton and Blair 1991), heaping for behavioral frequency questions was greater for higher response values and lower among respondents who used counting to answer than among those who used other strategies. We did not replicate the finding of greater heaping for proxy questions than

Table 5. Multilevel Model Predicting Heaping for Study 4 (standard errors in parentheses)

Predictor	Residual Heaping index
Ability	
Education	
Less than HS degree	–1.10**
	(.26)
HS degree	–.71**
	(.18)
Some college	–.38*
	(.16)
Interviewer's ratings of respondent's intelligence	.41
	(.47)
Interviewer's ratings of respondent's political knowledge	–.15
	(.40)
Index of objective political knowledge questions	.10
	(.39)
Motivation	
Need for cog.	–.29
	(.19)
R's interest	–.34
	(.34)
Refusal conversion	.38
	(.51)
Contact attempts	.01
	(.02)
Mode	
Telephone	.24*
	(.12)
Control variables	
Response value	–.74**
	(.02)
N (respondents)	1,325
N (observations)	29,690
Rho	.50
Rho (no predictors)	.79

NOTE.—Analyses conducted controlling for gender, race/ethnicity, mode, age, and age squared. Unstandardized coefficients shown (standard errors in parentheses) from xtlogit analysis in Stata with heaping across questions clustered within respondents.

*$p < .05$; **$p < .01$

for self questions, but only a few proxy questions were examined, and they did not perfectly parallel the questions respondents answered about themselves.

For most types of questions examined, there was little evidence that heaping was a response style or was a strategy respondents used to satisfice. Questions about respondents' personal characteristics were a partial exception to this. Although heaping for these questions did not show strong intraclass correlations and was not more likely under the conditions thought to foster satisficing, heaped responses were faster than unheaped responses and contained greater error when compared to objective measures. This provides weak evidence that heaping in questions about personal characteristics results from respondents taking cognitive shortcuts.

Feeling-thermometer questions were the only subjective questions for which heaping was examined and showed substantially different patterns than other question types. Heaping in feeling-thermometer questions showed consistent and strong intraclass correlations but was weakly associated with respondent ability and motivation, such that respondents with greater ability and less motivation heaped more. Furthermore, heaping in these questions was associated with longer response latencies and greater predictive validity.

These findings suggest that the processes that lead to response heaping may be very different for questions about objective constructs than for those about subjective phenomena. For objective questions, heaping may reflect estimation strategies and a lack of confidence that one can generate an accurate response, whereas for subjective questions heaping may reflect a thoughtful, engaged process of making sense of a difficult task. Therefore, it may not be reasonable to assume that heaping always reflects lower data quality (Couper et al. 2006), particularly for questions about subjective phenomena.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

Although this research makes an important contribution to our understanding of heaping, there are limitations to consider. First, our studies included a relatively small number of questions and question type was confounded with the judgment respondents were asked to make; it would be difficult to separate these dimensions (e.g., to hold all other features of a question constant and manipulate whether it requires a respondent to report a behavioral frequency versus a feeling thermometer). The range and specific distribution of responses also varied across items and question types. Therefore, we cannot draw strong conclusions about why different question types showed different levels of heaping and different associations with predictors. One direction for future research would be to examine heaping for a more extensive set of survey items.

A second limitation is that we measured rather than manipulated many of the heaping predictors. Future research could manipulate factors associated with satisficing (e.g., respondent ability or motivation) and test their effects on response heaping. This type of experiment is rare in the literature exploring survey satisficing.

Third, we were unable to assess measurement error for some question types. Future studies could compare self-reported responses to BFQs to official records of behavioral frequencies or data from another source, such as diaries. Our results

suggest that heaping may be linked to higher data quality for subjective questions but to lower data quality for objective questions. Additional research is necessary to determine the conditions under which heaping poses a problem for researchers.

Fourth, our examination of heaping was limited to interviewer-administered surveys. Although heaping has been observed in self-administered surveys (e.g., Couper et al. 2006), we could not test whether heaping differs for self-administered and interviewer-administered modes of data collection. Future research could explore this more directly.

Finally, these studies largely confirm the null hypothesis that heaping does not reflect a strategy on the part of respondents to satisfice, which some might see as a limitation. The scientific process is often biased against seeing value in these kinds of null results, and null findings are all too often relegated to the "file drawer," where they make no contribution to accumulated knowledge (Greenwald 1975), although significant results are no more true than null findings (Nosek, Spies, and Motyl 2012). We argue that in this case, our finding that response heaping does not appear to be a form of survey satisficing is important because it suggests conventional wisdom is wrong. Providing evidence that conventional wisdom is incorrect tells us that we need to look in other directions to understand why and under what conditions respondents heap, and then to understand the consequences of heaping for data quality.

Appendix A. Question Wordings and Codings

Study 1

Response heaping (all items were asked in open-ended format, and responses were coded as numeric values; that is, number of days per month was coded between 0 to 30):

Behavioral frequencies:

(1) During the last year, how many times did you see or talk to a medical doctor?
(2) How many times per week or per month did you take part in [PHYSICAL ACTIVITY IDENTIFIED IN PREVIOUS QUESTION] during the past month?
(3) In a typical week, how many times do you talk on the telephone with family, friends, or neighbors?
(4) (If *R* currently smokes) On average, about how many cigarettes do you smoke per day?
(5) (If *R* ever smoked marijuana) About how many times in your life have you used marijuana or hash?
(6) (If *R* had sex in the past five years) In the past five years, how many different people have you had as sexual partners?

(7) (If *R* ever had an alcoholic drink) On about how many different days did you have one or more drinks during the past 30 days?
(8) (If *R* had an alcoholic drink in the last month) About how many drinks did you usually have in a day on the days that you drank during the past 30 days?
(9) (If *R* has talked with anyone about getting AIDS by eating in a restaurant where the cook has the AIDS virus) How many times do you remember talking with someone about this?

Personal characteristics:

(1) About how much do you weigh without shoes?
(2) How much did you weigh when you were 16 years old?

Percentages

(1) (If *R* had sex in the past five years) What percent of the time would you say [you/your partner(s)] used a condom?

Age at event:

(1) (If *R* ever smoked daily) About how old were you when you first started smoking daily?
(2) (If *R* ever drank alcohol) How old were you the first time you had a glass of beer or wine or a drink of liquor, such as whiskey, gin, scotch, and so forth? Do not include sips that you might have had from an older person's drink.

Cognitive process involved in answering behavioral frequency questions:

Immediately after the first seven behavioral frequency questions above, respondents were asked, "How did you arrive at this number?" or "How did you remember the number of times you did this?" Responses were coded to reflect whether or not respondents used counting (or episodic enumeration). For each variable, responses to this structured probe were coded 1 if the respondent indicated they used counting or episodic enumeration and 0 if they indicated they used any other strategy.

Question time frame:

For the question asking respondents for the number of times per week or month they had engaged in physical activity, a variable was coded 0 if the respondent chose to report per week and 1 if *R* chose to report per month. For the question asking respondents to report the number of times per day or week that they talk to friends or family on the phone, a variable was coded 0 if the respondent chose to report per day and 1 if the respondent chose to report per week.

Respondent ability (education):

What is the highest level of formal education you have completed? (Please include high school and college but do not include vocational or technical training.) Eight years or less, Some high school, High school diploma or GED, Some college, College degree (from four-year college), or Graduate school/degree.

Coding: Responses were coded using three dummy variables. The first was coded 1 for respondents with less than a high school degree and 0 for everyone else; the second was coded 1 for respondents with a high school degree or GED and 0 for everyone else; and the third was coded 1 for respondents with some college and 0 for everyone else. Respondents with a four-year college degree or more were used as the comparison group.

Study 2

Response heaping:

Behavioral frequencies (all items were asked in open-ended format, and responses were coded as numeric values; that is, number of days per month was coded between 0 to 30):

(1) Now, thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
(2) (IF *R* SMOKES) On average, how many cigarettes do you now smoke per day?
(3) How many times have you thought about your chances of getting cancer?
(4) Thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
(5) (IF *R* REPORTED AT LEAST ONE CLOSE FRIEND) Excluding electronic social networking like email and Facebook, how many of these close friends did you spend time with or talk to during the past seven days?
(6) The next question is about how often you have ever bet or gambled for money. In answering, think about all the times you ever made a bet of any sort, from betting on sports in an office pool to playing cards with friends, buying lottery tickets, playing bingo, speculating on high-risk stocks, playing pool or golf for money, playing slot machines, betting on horse races, and any other kind of betting or gambling. Taking all these things together, what's your best estimate of how many times you ever made a bet of any kind in your entire life?
(7) Now, thinking about [PROXY]'s physical health, which includes physical illness and injury, for how many days during the past 30 days was [PROXY]'s physical health not good?
(8) How many times has [PROXY] spoken with you about his chances of getting cancer?

Personal or proxy characteristics (all items were asked in open-ended format, and responses were coded as numeric values; for example, weight was coded as the number of kilograms or pounds):

(1–2) Blood pressure is usually given as one number over another. What was your most recent blood pressure in numbers? (Interviewers then entered separate reports of both systolic and diastolic blood pressure.)
(3) How much do you weigh without shoes?
(4) How much did you weigh at age 12?
(5) In general, how many close friends do you have? By "close friends," I mean relatives or non-relatives that you feel at ease with, can talk to about private matters, and can call on for help.
(6) Can you tell me the age of the adult(s) in the household? To maximize sample size, age reports for only the first adult in the household were used.
(7) How much does [PROXY] weigh without shoes?

Age at an event (all items were asked in open-ended format, and responses were coded as numeric values):

(1) Now I would like to ask you some questions about tobacco use. How old were you the first time you ever thought about trying to smoke a cigarette?
(2) (IF EVER HAD ALCOHOLIC DRINK) How old were you the first time you had a drink of an alcoholic beverage? Please do not include any time when you had only a sip or two from a drink.
(3) How old were you the first time you were ever treated in a hospital emergency room?
(4) (IF *R* SAID THERE HAD BEEN A TIME IN THEIR LIFE WHEN THEY HAD BEEN UNDER A GREAT AMOUNT OF STRESS) How old were you the first time you felt under a great amount of stress?
(5) (IF *R* REPORTED EVER HAVING A SERIOUS CONFLICT WITH EMPLOYER OR COWORKER) How old were you the first time this happened?
(6) (IF *R* REPORTED EVER PLACING A BET) How old were you the very first time you placed a bet or gambled for money?
(7) (IF *R* REPORTED THE PROXY HAD EVER SMOKED A CIGARETTE) How old do you think [PROXY] was the first time [PROXY] ever smoked a cigarette?
(8) (IF R REPORTED THE PROXY HAD EVER HAD AN ALCOHOLIC DRINK) How old was [PROXY] the first time s/he had a drink of an alcoholic beverage? Please do not include any time when [PROXY] may have had only a sip or two from a drink.

Respondent ability:
Respondent education:
What is the highest degree you have completed?

Coding: Respondent education was coded into two dummy variables for respondents with a high school degree or less and for respondents with some college education (but not a four-year degree). The comparison group was respondents with a four-year degree or more.

Respondents' intelligence:

Interviewers were asked to rate respondent's apparent intelligence as very high, fairly high, average, fairly low, or very low.

Coding: Responses were coded to range from 0 (very low) to 1 (very high).

Respondent motivation:

Need for cognition:

(1) Some people like to have responsibility for handling situations that require a lot of thinking, and other people don't like to have responsibility for situations like that. What about you? Do you like having responsibility for handling situations that require a lot of thinking, do you dislike it, or do you neither like nor dislike it? [IF LIKE: Do you like it a lot or just somewhat? IF DISLIKE: Do you dislike it a lot or just somewhat?] (Responses were coded 0 for "dislike it a lot," .25 for "dislike it a little," .50 for "neither like nor dislike it," .75 for "like it a little," and 1.0 for "like it a lot.")
(2) Some people prefer to solve simple problems instead of complex ones, whereas other people prefer to solve more complex problems. Which type of problems do you prefer to solve: simple or complex? (Responses were coded 0 for respondents who reported that they preferred simple problems and 1 for respondents who reported that they preferred complex problems.)

Index: Responses to these two questions were highly correlated ($\alpha = .64$), and responses to these two items were averaged to form a need for cognition index.

Self-reported effort:

(1) How carefully did you think when answering this survey? Would you say extremely carefully, very carefully, somewhat carefully, not too carefully, or not carefully at all?
(2) How thoroughly did you search your memory when deciding on your answers to the questions? Would you say extremely thoroughly, very thoroughly, somewhat thoroughly, not too thoroughly, or not thoroughly at all?
(3) How hard did you work at interpreting the meaning of each question in the questionnaire? Would you say extremely hard, very hard, somewhat hard, not too hard, or not hard at all?

Index: Responses to each of the three questions were coded to range from 0 (least effort, thoroughness, care) to 1 (most effort, thoroughness, care) and averaged to form an index ($\alpha = .54$).

Respondent interest:

Interviewers were asked, "Overall, how great was *R*'s interest in the interview? Very high, fairly high, average, fairly low, or very low?" Responses were coded to range from 0 (very low) to 1 (very high).

Questionnaire location:

The order of two halves of the questionnaire was rotated across respondents so that some respondents were randomly assigned to receive sections 1 and 2 of the questionnaire before sections 3 and 4, and other respondents were randomly assigned to receive sections 3 and 4 before sections 1 and 2. All the questions used to assess heaping were in these four sections (a few questions, such as the questions measuring respondent self-reports of interest, effort, and difficulty, were always asked after all the other items). A variable was coded 0 for respondents who were asked a given heaping question in the first half of the questionnaire and 1 for respondents who were asked the question in the second half of the questionnaire.

Task difficulty:

Respondent ratings:

(1) How difficult was it to answer the questions in this survey? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not difficult at all?
(2) How difficult was it to remember information relevant to the questions I asked you? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not difficult at all?
(3) How difficult was it to select an answer to each question in this survey? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not at all difficult?
(4) How clear were your memories about the types of information asked about in this survey? Would you say extremely clear, very clear, somewhat clear, not too clear, or not at all clear? (Item was reverse coded.)
(5) How easy was it to understand the meaning of each question in this survey? Would you say extremely easy, very easy, somewhat easy, not too easy, or not at all easy? (Item was reverse coded.)

Coding: Responses to all five questions were coded to range from 0 (least difficulty or most ease) to 1 (most difficulty or least ease) and averaged to form an index of difficulty ($\alpha = .73$).

Response latencies:

In order to ensure that the response latency time was accurately captured, the instrument was set up with three screens for each item:

1. The first screen was the "Q screen" (question screen). It contained only the question with the response options included. Interviewers did not enter a respondent's answer on this screen. After they read the question, pressing "Enter" took them to the response screen.
2. The second screen, or the "R screen" (response screen), contained the text of the question in parentheses and the response options with their values next to them. Interviewers read the question again only if the respondent asked them to repeat the question. Otherwise, when the respondent provided an answer, the interviewer selected the proper response option value and was automatically taken to the third screen. The only valid keystrokes were the response option values.
3. The third screen was the "L screen" (response latency screen). This screen was the same for every item in the questionnaire, and it contained an option for a Valid Latency, as well as a number of options for issues that might have affected the response latency. This screen was not to be read aloud.

The instrument was programmed so that a timer would begin as soon as an interviewer hit the "Enter" key after reading the question and went to the response screen. This ensured the response latency would not be affected by how long it took the interviewer to read the question. The timer was stopped by the interviewer entering the respondent's answer as soon as the respondent began to answer. Immediately after this, the interviewer indicated whether the response latency was valid or invalid (e.g., due to an event like the respondent answering the question before it was completely read).

Study 3

Response heaping:

Behavioral frequencies (all items were asked in open-ended format, and responses were coded as numeric values):

(1) How many times in the past seven days have you felt angry at anyone?
(2) How many times in the past seven days have you felt angry at a member of your family?

Personal characteristics (all items were asked in open-ended format, and responses were coded as numeric values; for example, length of time in community was coded as number of years):

(1) How long have you lived in your present community?

Percentages (all items were asked in open-ended format, and responses were coded as numeric values ranging from 0 to 100):

(1) What percentage of people living in the United States do you believe would be willing to pay more taxes to support a federal program to provide universal health-care coverage for all people living in the United States?
(2) In the presidential election of 2004, what percent of eligible voters in the United States cast a ballot?

Feeling thermometers (all items were asked in open-ended format, and responses were coded as numeric values ranging from 0 to 100):

Please look at CARD C1. I'd like to get your feelings toward some groups. I'll read the name of a group, and I'd like you to rate that group using something we call the feeling thermometer. Ratings between 50 and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. You would rate the group at the 50-degree mark if you don't feel particularly warm or cold toward the group. If we come to a group you don't recognize, you don't need to rate that group. Just tell me and we'll move on to the next one.

Procedure and groups: Respondents were then given a show card (see figure 1) and asked to rate the following groups: (1) Jews; (2) middle-class people; (3) labor unions; (4) poor people; (5) the military; (6) big business; (7) people on welfare; (8) working-class people; (9) older people (the elderly); (10) environmentalists; (11) gay men and lesbians; (12) Southerners; (13) young people; and (14) rich people.

Respondent motivation:
Self-reported effort:

(1) How carefully did you think when answering this survey? Would you say extremely carefully, very carefully, somewhat carefully, not too carefully, or not carefully at all?
(2) How thoroughly did you search your memory when deciding on your answers to the questions? Would you say extremely thoroughly, very thoroughly, somewhat thoroughly, not too thoroughly, or not thoroughly at all?
(3) How hard did you work at interpreting the meaning of each question in the questionnaire? Would you say extremely hard, very hard, somewhat hard, not too hard, or not hard at all?
(4) How much effort did you spend to make sure the answer you gave to each question best represented your views? Would you say a great deal of effort, a lot of effort, some effort, a little effort, or no effort at all?

Index: Responses to each of the four questions were coded to range from 0 (least effort, thoroughness, care) to 1 (most effort, thoroughness, care) and averaged to form an index ($\alpha = .69$).

Questionnaire location:

The questionnaire was separated into four sections. Respondents were randomly assigned to receive the questionnaire sections in one of three orders (sections 1, 2, 3, 4; sections 2, 3, 1, 4; or sections 3, 1, 2, 4). Section 1 asked questions about the respondent's physical health; section 2 asked questions about physical- and mental-health knowledge; section 3 asked about the respondent's mental health; and section 4 asked questions about the selected proxy and demographic questions and was always asked last. Questions that were asked in the first part of the questionnaire were coded 0; questions asked in the second part were coded 1/3; questions that were asked in the third part were coded 2/3; and questions that were asked in the final part were coded 1.

Task difficulty:

(1) How difficult was it to answer the questions in this survey? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not difficult at all?

(2) How difficult was it to remember information relevant to the questions I asked you? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not difficult at all?

(3) How difficult was it to select answers that best represented your views? Would you say extremely difficult, very difficult, somewhat difficult, not too difficult, or not at all difficult?

Index: Responses to all three questions were coded to range from 0 (not difficult at all) to 1 (extremely difficult) and averaged to form an index of difficulty ($\alpha = .69$).

Study 4

Response heaping

Feeling thermometers (all items were asked in open-ended format, and responses were coded as numeric values ranging from 0 to 100):

Respondents interviewed face-to-face were told:

Please look at page 1 of the booklet. I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person, and I'd like you to rate that person using something we call the feeling thermometer. Ratings between 50 and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50-degree mark if you don't feel particularly

warm or cold toward the person. If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one. How would you rate…? [LATER IN THE INTERVIEW] Still using the thermometer, how would you rate…?

Respondents interviewed via the telephone were told:

I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person, and I'd like you to rate that person using something we call the feeling thermometer. The feeling thermometer can rate people from 0 to 100 degrees. Ratings between 50 and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 and 50 degrees mean that you don't feel favorable toward the person. Rating the person at the midpoint, the 50-degree mark, means you don't feel particularly warm or cold toward the person. If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one. How would you rate….? [LATER IN THE INTERVIEW] Still using the thermometer, how would you rate…?[11]

Targets: All respondents were asked to evaluate the Democratic and Republican presidential candidates (Al Gore and George W. Bush, respectively) and 24 groups: (1) the Supreme Court; (2) Congress; (3) the military; (4) the federal government in Washington, DC; (5) Blacks; (6) Whites; (7) conservatives; (8) liberals; (9) labor unions; (10) big business; (11) poor people; (12) people on welfare; (13) Hispanics; (14) Christian fundamentalists; (15) women's movement,; (16) older people; (17) environmentalists; (18) homosexuals; (19) Christian coalition; (20) Catholics; (21) Jews; (22) Protestants; (23) feminists; and (24) Asian Americans.

Respondent ability:

Respondent education:

What is the highest degree that you have earned?

Coding: Education was coded via three dummy variables. The first was coded 1 for respondents with less than a high school education and 0 for everyone else; the second was coded 1 for respondents with a high school

11. Although telephone respondents did not see a show card with scale points labeled, they were provided with verbal labels for three of the scale points (0, 50, and 100). It is necessary to do this in cases like this, where a respondent is being asked to make a subjective rating, since the numbers used have no inherent meaning. One might be concerned that the verbal labels provided to telephone respondents contributed to heaping in this mode and that respondents might not have heaped had these verbal labels not been provided. However, the values of 0 and 100 were not counted as heaping (to avoid confounding heaping and extreme response style). Furthermore, in the telephone mode, 97 percent of respondents who chose a response other than 0, 50, or 100 chose a response that was divisible by 5. So, even among respondents who did not select one of these labeled scale points, heaping was extremely common.

degree and 0 for everyone else; and the third was coded 1 for respondents with some college and 0 for everyone else. Respondents with at least a four-year degree were the baseline comparison group.

Interviewers' ratings of respondents' intelligence:

Interviewers were asked to rate each respondent's intelligence as very high, fairly high, average, fairly low, or very low. Responses were coded to range from 0 (very low) to 1 (very high).

Objective political knowledge:

(1) Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers, and the like. The first name is TRENT LOTT. (What job or political office does he NOW hold?)
(2) WILLIAM REHNQUIST [PRON: Renn-kwist] (What job or political office does he NOW hold?)
(3) TONY BLAIR (What job or political office does he NOW hold?)
(4) JANET RENO (What job or political office does she NOW hold?)
(5) Next, I'd like to ask you about the candidates who ran for president and their running mates. We're interested in some of the things that people may have heard about these candidates. The first candidate I'd like to ask you about is George W. Bush. What U.S. state does George W. Bush live in now?
(6) What is George W. Bush's religion?
(7) Now take Al Gore. What U.S. state is Al Gore from originally?
(8) What is Al Gore's religion?
(9) What about Dick Cheney? What U.S. state does Dick Cheney live in now?
(10) What is Dick Cheney's religion?
(11) And Joseph Lieberman. What U.S. state does Joseph Lieberman live in now?
(12) What is Joseph Lieberman's religion?
(13) Do you happen to know which party had the most members in the House of Representatives in Washington BEFORE the election (this/last) month? (IF NECESSARY: WHICH ONE?)
(14) Do you happen to know which party had the most members in the U.S. Senate BEFORE the election (this/last) month? (IF NECESSARY: WHICH ONE?)

Index: A political knowledge index was calculated as the percentage of these 14 knowledge questions that each respondent answered correctly.

Subjective political knowledge:

Interviewers were asked to rate respondents' political knowledge as very high, fairly high, average, fairly low, or very low. Responses were coded to range from 0 (very low) to 1 (very high).

Respondent motivation:

Need for cognition:

Need for cognition was measured and coded as in studies 2 and 3.

Interviewers' ratings of respondents' interest:

Interviewers rated respondents' interest in the interview as very high, fairly high, average, fairly low, or very low. Responses were coded to range from 0 (very low) to 1 (very high).

Reluctant respondents:

Number of contact attempts and whether or not respondents had ever refused were assessed during the data-collection period. A refusal conversion variable was coded 1 for respondents who were successfully interviewed preelection after an initial refusal and 0 for respondents who did not refuse before agreeing to be interviewed. A contact attempts variable indicated the number of times respondents were contacted before an interview was completed.

Respondent behavioral intentions:

(IF THE RESPONDENT REPORTED THAT S/HE PLANNED TO VOTE IN THE NOVEMBER ELECTION) "Who do you think you will vote for in the election for president?"

Coding: One variable was coded 1 if the respondent said they planned to vote for Al Gore and 0 otherwise. A second was coded 1 if the respondent said they planned to vote for George W. Bush and 0 otherwise.

Respondent vote choice:

(IF THE RESPONDENT REPORTED THAT S/HE HAD VOTED FOR PRESIDENT IN THE ELECTION) "Who did you vote for?"

Coding: One variable was coded 1 if the respondent said they had voted for Al Gore and 0 otherwise. A second was coded 1 if the respondent said they had voted for George W. Bush and 0 otherwise.

Appendix B. Additional Explanation of Multilevel Modeling

Multilevel models are used to examine the effects of question characteristics and respondent characteristics, adjusting for responses being nested within respondents (Stata 2011). Models to predict heaping are based on random-intercept logistic regression models and employ the Stata xtlogic command. A generic equation of such models is defined as

$$\begin{aligned}\log\Big[\, p\left(Heaping_{ij}=1\right)/(1-p(Heaping_{ij}=1)\Big] \\ = B_{00}+B_{1j}X_{1ij}+B_{2j}X_{2ij}+\ldots B_{nj}X_{nij} \\ {}_{+}B_{01}Z_{1j}+B_{02}Z_{2j}+\ldots B_{0n}Z_{nj}+u_{0j},\end{aligned} \tag{1}$$

where B_{nj} ($n=1, 2,\ldots, k$) are coefficients of the k explanatory variables of X_{nij} and Z_{nj} at the question and respondent level, respectively, predicting the probability of having a heaping problem, $p(Heaping=1)$, with question i by respondent j. The probability is modeled using a logit link function with an assumption of a Bernoulli-distributed outcome. u_0j is the random effect at the respondent level, which is assumed to be normally distributed with the expected value of 0 and the variance of ψ.

The between-subject heterogeneity for heaping is quantified as the intraclass correlation coefficient (ρ), specified as $\psi/(\psi+ (\pi^2/3))$, a proportion of between-subject variance (ψ) out of the total variance (between- and within-subject variance ($\pi^2/3$)). The unconditional intraclass correlation coefficient (ρ, rho) was estimated without any predictors (as presented in the bottom row of the tables), and the conditional intraclass correlation was calculated after controlling for all the preceptors in the model (as presented in the penultimate row of the tables). This was done across all questions and separately for each question type of which there were at least two questions (studies 1 and 3). Subsequently, a model was estimated in which heaping was regressed on education, the response value (the numeric answer each respondent gave to the question, standardized within question), a series of dummy variables indicating the response strategy used by the respondent, a series of dummy variables reflecting the time frame respondents chose to use to report their response, and a series of dummy variables indicating question type. A third model tested the interactions between all other predictor variables and question type. Finally, models predicting heaping were estimated separately for each question type.

For a continuous outcome of our study (i.e., response latencies), the Stata xtreg command is employed to estimate a series of multilevel models predicting reciprocalized response latencies as the dependent variable (a model where response latencies were regressed on heaping, predictors of heaping, and question type; a model in which reciprocalized response latencies were regressed on heaping, predictors of heaping, question type, and interactions between heaping and question type; and models in which reciprocalized

response latencies were regressed on heaping and predictors of heaping separately for each question type). Equations for the models are specified as

$$\begin{aligned} Latencies_{ij} &= B_{00} + B_{1j}X_{1ij} + B_{2j}X_{2ij} + \ldots B_{nj}X_{nij} \\ &\quad + B_{01}Z_{1j} + B_{02}Z_{2j} + \ldots B_{0n}Z_{nj} + e_{ii} + u_{0j}, \end{aligned} \tag{2}$$

where $Latencies_{ij}$ is the outcome variable for the *i*th question, responded by *j*th subject. B_{00} is the intercept, and explanatory variables are at the question level (X_{nij}) and respondent level (Z_{nj}). The residual errors at the response level, e_{ii}, are assumed to have a mean of 0 and a variance of σe2. The respondent-level random variance terms, u_{0j}, are assumed to be normally distributed with a mean of 0 and a variance of τ, and to be independent from the residual errors e_{ii}.

As in the case of heaping, the between-subject heterogeneity for latencies is quantified as the intraclass correlation coefficient (ρ), which is specified as τ/(σe2 + τ), and unconditional and conditional intraclass correlation coefficients were estimated.

References

Alwin, Duane F., and Jon A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20:139–81.

Bachman, Jerry G., and Patrick M. O'Malley. 1984. "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles." *Public Opinion Quarterly* 48:491–509.

Barbieri, Magali, and Véronique Hertrich. 2005. "Age Difference between Spouses and Contraceptive Practice in Sub-Saharan Africa." *Population* 60:617–54.

Bassili, John N. 1996. "The 'How' and 'Why' of Response Latency Measurement in Survey Research." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by Norbert Schwarz and Seymour Sudman, 319–46. San Francisco: Jossey-Bass Publishers.

Battistin, Erich, Raffaele Miniaci, and Guglielmo Weber. 2003. "What Do We Learn from Recall Consumption Data?" *Journal of Human Resources* 38:354–85.

Boyle, Phelim P., and Cormac Ó. Gráda. 1986. "Fertility Trends, Excess Mortality, and the Great Irish Famine." *Demography* 23:543–62.

Budd, John W., and Timothy Guinnane. 1991. "Age-Misreporting, Age-Heaping, and the 1908 Old Age Pension Act in Ireland." *Population Studies* 45:497–518.

Burton, Scot, and Edward Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys." *Public Opinion Quarterly* 55:50–79.

Coale, Ansley. 1955. "The Population of the United States in 1950 Classified by Age, Sex, and Color—A Revision of Census Figures." *Journal of the American Statistical Association* 50(269):16–54.

Conrad, Frederick G., Norman R. Brown, and Erin R. Cashman. 1998. "Strategies for Estimating Behavioural Frequency in Survey Interviews." *Memory* 6:339–66.

Couper, Mick P., Roger Tourangeau, Frederick G. Conrad, and Eleanor Singer. 2006. "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment." *Social Science Computer Review* 24:227–45.

Diebold, Franci X., David Neumark, and Daniel Polsky. 1997. "Job Stability in the United States." *Journal of Labor Economics* 15:206–33.

Fazio, Russell H. 1990. "A Practical Guide to the Use of Response Latency in Social Psychological Research." In *Review of Personality and Social Psychology*, vol. 11, *Research Methods in Personality and Social Psychology*, edited by Clyde A. Hendrick and Margaret S. Clark, 74–97. Thousand Oaks, CA: Sage Publications.

Greenwald, Anthony G. 1975. "Consequences of Prejudice against the Null Hypothesis." *Psychological Bulletin* 82:1–20.

Hirsch, Barry T. 2005. "Why Do Part-Time Workers Earn Less? The Role of Worker and Job Skills." *Industrial and Labor Relations Review* 58:525–51.

Hobson, Richard. 1976. "Properties Preserved by Some Smoothing Functions." *Journal of the American Statistical Association* 71(355):763–66.

Hottenga, Jouke-Jan, Dorret I. Boomsma, Nina Kupper, Danielle Posthuma, Harold Snieder, Gonneke Willemsen, and Eco J. de Geus. 2005. "Heritability and Stability of Resting Blood Pressure." *Twin Research and Human Genetics* 8:499–508.

Huttonlocher, Janellen, Larry V. Hedges, and Norman M. Bradburn. 1990. "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16:196–213.

Jyothinagaram, Sathya G., Leanne Rae, Adam V. Campbell, and Paul L. Padfield. 1990. "Stability of Home Blood Pressure over Time." *Journal of Human Hypertension* 4:269–71.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

———. 1999. "Survey Research." *Annual Review of Psychology* 50:537–67.

Krueger, Patrick M., Richard G. Rogers, Robert A. Hummer, and Jason D. Boardman. 2004. "Body Mass, Smoking, and Overall and Cause-Specific Mortality among Older U.S. Adults." *Research on Aging* 26:82–107.

Myers, Robert J. 1976. "An Instance of Reverse Heaping of Ages." *Demography* 13:577–80.

Nagi, M. H., E. G. Stockwell, and L. M. Snavley. 1973. "Digit Preference and Avoidance in the Age Statistics of Some Recent African Censuses: Some Patterns and Correlates." *International Statistical Review* 41:165–74.

Narayan, Sowmya, and Jon A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60:58–88.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability." *Perspectives on Psychological Science* 20:1–17.

Schaeffer, Nora Cate, and Stanley Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29:65–88.

StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Tarrant, Michael A., and Michael J. Manfredo. 1993. "Digit Preference, Recall Bias, and Nonresponse Bias in Self-Reports of Angling Participation." *Leisure Sciences* 15:231–38.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Vaske, Jerry, and Jay Beaman. 2006. "Lessons Learned in Detecting and Correcting Response Heaping: Conceptual, Methodological, and Empirical Observations." *Human Dimensions of Wildlife* 11:285–96.

Walejko, Gina K. 2010. "The Effectiveness of an Interactive Web Survey in Decreasing Satisficing and Social Desirability Bias." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, USA.

Wedeking, Justin P., and Joanne M. Miller. 2004. "Measuring Public Opinion: Examining the Impact of the Refusal Conversions and Callbacks on Data Quality." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, USA.

West, Kirsten K., J. Gregory Robinson, and Michael Bentley. 2005. "Did Proxy Respondents Cause Age Heaping in the Census 2000?" Proceedings of the Survey Research Methods Section of the American Statistical Association, 3658–65.

Zelnik, Melvin. 1964. "Errors in the 1960 Census Enumeration of Native Whites." *Journal of the American Statistical Association* 59(306):437–59.

Zhang, Charles, and Norbert Schwarz. 2012. "How and Why 1 Year Differs from 365 Days: A Conversational Logic Analysis of Inferences from the Granularity of Quantitative Expressions." *Journal of Consumer Research* 39:248–59.