Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations

Brian J. Zikmund-Fisher, PhD, Dylan M. Smith, PhD, Peter A. Ubel, MD, Angela Fagerlin, PhD

Background. *In a companion article, the authors describe the Subjective Numeracy Scale (SNS), a self-assessment of numerical aptitude and preferences for numbers that correlates strongly with objective numeracy.* ***Objective.*** *The objective of this article is to validate the Subjective Numeracy Scale using measures of subjects' capacity to recall and comprehend complex risk statistics and to complete utility elicitations.* ***Research Design.*** *The study is composed of 3 general public surveys: 2 administered via the Web and 1 by paper and pencil.* ***Subjects.*** *Studies 1 and 3 surveyed 862 and 1234 people, respectively, recruited via a nationwide commercial Internet survey panel. Study 2 involved 245 people who completed paper-and-pencil surveys in a Veterans Administration hospital.* ***Measures.*** *The authors tested whether one's score on the SNS predicted the likelihood of correct recall and interpretation of risk information (studies 1 and 2A) or the likelihood of effectively completing a time tradeoff or person-tradeoff utility elicitation (studies 2B and 3). In Studies 1 and 2, the authors also tested whether an objective test of quantitative ability would predict performance.* ***Results.*** *In all studies, survey participants with higher SNS scores performed significantly better than other respondents. The predictive ability of the SNS approached that observed for objective numeracy.* ***Conclusions.*** *The SNS effectively predicts both risk comprehension and completion of utility elicitations without requiring survey participants to complete time-consuming and stress-inducing mathematics tests. The authors encourage the use of the SNS in a variety of health services research contexts.* ***Key words:*** *numeracy; risk communication; decision making; literacy; utility measurement.* ***(Med Decis Making 2007;27:663–671)***

The Subjective Numeracy Scale (SNS) is a self-report measure of perceived ability to perform various mathematical tasks and preference for the use of numerical versus prose information.[1] The 8-item scale contains no mathematics questions and has no correct or incorrect answers. Instead, it consists of 4 questions asking respondents to assess their numerical ability in different contexts and 4 questions asking them to state their preferences for the presentation of numerical and probabilistic information. Because the SNS measures perceptions of quantitative ability rather than the ability itself, we expected it to serve as a valuable, but imperfect, proxy for tests of objective numeracy. Our companion article[1] shows that the SNS is both reliable and highly correlated with the most commonly used numeracy measure.

We report the results of 3 studies that test the predictive validity of the SNS. Specifically, Study 1 tested individuals' ability to correctly recall risk information presented textually (as in previous numeracy research[2]) as well as visually in pictographs (image matrices).[3] Study 2 (part of study 2 of our development article[1]) had 2 parts: Study 2A extended our inquiry into numeracy effects in risk communication by assessing individuals' ability to comprehend and interpret risk statistics presented in a survival curve format. Study 2B tested whether subjective numeracy would predict the ability of survey respondents to complete and provide internally consistent information in a paper-and-pencil time-tradeoff (TTO) utility elicitation task, a task

DOI: 10.1177/0272989X07303824

previously shown to be correlated with scores on objective numeracy measures.[4] Studies 1 and 2 also included measures of objective numeracy, enabling some discussion of comparative validity. Study 3 then extended the numeracy literature by examining the association between SNS scores and completion rates for an Internet-administered person-tradeoff (PTO) utility task.

STUDY 1: RECALL OF PICTOGRAPHS AND TEXTUAL RISK INFORMATION

The goal of study 1 was to demonstrate the potential of subjective numeracy measures to predict the ability to comprehend statistical health information. To do so, we included a subset of the subjective numeracy questions as part of an ongoing research project evaluating the use of pictographs (image matrices) as risk communication tools.

Methods

Study 1's primary between-subjects manipulation compared recall of risk information presented in text form only versus in text plus pictographs. Study 1 also tested whether a respondent's numeracy was associated with his or her ability to comprehend the presented statistics. We recruited participants through an Internet survey panel maintained by Survey Sampling International (the Survey Spot Internet panel). This panel is made up of more than 1 million unique member households, recruited via random-digit dialing, banner ads, and other opt-in techniques. The sample was stratified to mirror the U.S. census population based on age, gender, race, education level, and income. Upon completion of the survey, participants were entered into a drawing for cash prizes of up to $10,000. This research design and recruitment method received Institutional Review Board approval, as did each of the subsequent studies.

At the beginning of the study, each participant answered questions 1 through 3 and 5 and 6 of the 8-item SNS (see table 2 in our companion article[1]). Only a subset of the SNS questions, here referred to as the 5-item abbreviated SNS, was used because this study was completed while the development of the SNS was still ongoing. As per the methodology used in our companion article,[1] we used an imputed SNS score, equal to the average rating across all SNS items answered, for all subjects who had missing data yet completed more than half of the SNS question items.

In the next section, participants received information about 2 treatments for angina (coronary bypass surgery and balloon angioplasty) presented either in text form or in text plus pictograph form. The provided information included 3 sets of statistics about each treatment: the likelihood of reducing chest pain, the probability of serious complications, and the likelihood that the patient would require an additional procedure within the next 3 years. In the text conditions, each probability was included in the textual description (e.g., "About 54 out of 100 people who have BALLOON ANGIOPLASTY will need another procedure within 3 years"). In the pictograph conditions, each key probability was also shown visually in a 5×20 matrix of small figures.

After some filler tasks (which were long enough to discourage returning to the probability data), participants were then posed 2 types of recall questions. Regardless of whether they saw risk information in text or pictographs, all participants were asked 4 questions that required them to recall exact numerical risk information (verbatim recall) about 2 of the 3 sets of risk statistics provided. For example, some participants were asked to recall the number of patients out of 100 receiving bypass surgery who would experience significant pain relief. All verbatim recall questions used the same "___ out of 100"

Received 7 February 2005 from the VA Health Services Research & Development Center for Practice Management and Outcomes Research, VA Ann Arbor Healthcare System, Ann Arbor, Michigan (BJZ-F, DMS, PAU, AF); the Center for Behavioral and Decision Sciences in Medicine, Ann Arbor, Michigan (BJZ-F, DMS, PAU, AF); the Division of General Internal Medicine, University of Michigan, Ann Arbor (BJZ-F, DMS, PAU, AF); and the Department of Psychology, University of Michigan, Ann Arbor (PAU). Portions of this research were presented at the annual meeting of the Society for Medical Decision Making, Chicago, Illinois, 22 October 2003. Financial support for this study was provided by grants from the National Institutes for Health (R01 CA87595 and P50 CA101451). Dr. Zikmund-Fisher was supported by an HSR&D postdoctoral fellowship from the U.S. Department of Veterans Affairs, Drs. Smith and Fagerlin are supported by MREP early career awards from the Department of Veterans Affairs, and Dr. Ubel was the recipient of a Presidential Early Career Award for Scientists and Engineers during this research. The funding agreements ensured the authors' independence in designing the study, interpreting the data, and publishing the report. The authors acknowledge the outstanding research assistance of Aleksandra Jankovic throughout the course of this project as well as the insightful comments from 3 reviewers and the associate editor. Revision accepted for publication 5 October 2006.

Address correspondence to Brian J. Zikmund-Fisher, PhD, Center for Behavioral & Decision Sciences in Medicine, 300 North Ingalls Building, Room 7C27, Ann Arbor MI, 48109-0429; e-mail: bzikmund@umich.edu.

response scale. Participants were also asked 3 gist recall questions, which required them only to be able to identify which treatment was better on a given dimension. For example, one gist question asked, "After which treatment would a patient be more likely to experience a serious complication?" Outcomes were then coded as binary correct/incorrect answers, with verbatim recall required to be exact to be coded as correct.

Finally, respondents completed 10 objective numeracy questions at the end of the study. Nine of the questions were worded exactly as in the Lipkus and others[5] measure, and the 10th was an adaptation of another question from the measure.* We calculated an objective numeracy score by counting the number of questions answered correctly, yielding a scale ranging from 0 to 10. Following the methodology of Lipkus and others, objective numeracy scores were calculated considering questions with missing data to have been answered incorrectly (as long as the respondent answered any questions at all).

Data analysis. We used a logistic regression framework to analyze the association of numeracy with recall of risk information, analyzing respondents who were asked to recall textual risk information separately from those asked to recall risk statistics presented in a pictograph format. For both groups, we created a long-format data set, with 1 record for each recall question asked to each survey participant. A single dependent variable indicated whether that particular respondent answered a question correctly or incorrectly. We then ran 2 logistic regressions (1 each for recall of text and pictographs) in STATA 8,[6] including all records corresponding to the 6 possible verbatim recall questions. Each regression included the participants' average score on the abbreviated SNS as well as a set of question-specific binary variables to identify which question was being answered and control for differences in question difficulty. (Question by SNS interaction terms were omitted after initial models found no significant overall interaction, and the linearity of SNS in the models was examined and confirmed using graphical methods.) In addition, we standardized all reported odds ratios ($OR_{Std} = \exp[\beta^* SD_{Numeracy}]$) to show the change in odds due to a 1-standard-deviation change in numeracy score, and we used STATA's cluster option to obtain Huber/White robust standard errors and also account for the clustering caused by the multiple observations from each study participant. We ran similarly structured regressions using our gist recall questions as the dependent variable to test whether the abbreviated SNS would predict qualitative recall of the risk information. In addition, we ran supplemental analyses using only (partially complete) subscales (i.e., the 3 ability questions or the 2 preference questions) to confirm that both types of questions independently contribute to the predictive power of the SNS. Finally, we repeated the main analyses, substituting objective numeracy scores for the SNS.

*The 3rd question in the Lipkus and others scale reads, "In the Acme Publishing Sweepstakes, the chance of winning a car is 1 in 1000. What percentage of tickets to Acme Publishing Sweepstakes win a car?" As a result of a programming error, the question was modified to have the chance of winning the car to be 1 in 100.

Results

A total of 6306 people were invited to participate, and 996 participants began the survey, for a response rate of 15.8%. Of those who began the survey, 862 (87%) completed both a majority of the 5-item abbreviated SNS questions and at least 1 of the verbatim or gist recall questions and so were included in the SNS analyses. By a similar criterion, a total of 829 respondents were included in the objective numeracy analyses. Most (58%) respondents were female, and the average reported age was 49 years, with a range of 18 to 85 years. Education varied substantially among participants, with 34% holding a bachelor's degree or higher but also 15% reporting no education beyond high school. The distribution of SNS scores was roughly normal (with spikes at round number answers 3 and 5). The median SNS composite rating was 4.2 on a scale ranging from 1 to 6; 25th and 75th percentiles were 3.2 and 4.8, respectively.

Respondents correctly recalled certain types of risk information more often than others. Not surprisingly, gist recall was better than verbatim recall. For example, of the respondents who were asked whether they were more likely to need another procedure after angioplasty or after surgery, 73% correctly replied angioplasty. However, the accuracy of verbatim recall of the specific statistics regarding the need for another procedure was much lower (24% for angioplasty, 19% for surgery). Overall, accuracy rates varied from 19% to 54% for the verbatim recall questions and from 45% to 82% for the gist recall questions.

The 1st section of Table 1 presents the odds ratios of correct recall, *P* values, and c-index values derived from the logistic regression analyses. Note that all odds ratios reported in Table 1 are standardized to

Table 1 Logistic Regression Results Showing the Effect of Higher Subjective Numeracy Scale and Objective Numeracy Scores on Various Performance Measures

	Subjective Numeracy Scale			Objective Numeracy		
	Standardized Odds Ratio[a]	*P* Value	c-Index[b]	Standardized Odds Ratio[a]	*P* Value	c-Index[b]
Study 1 (N = 862)[c,d]						
Verbatim recall of						
Textual risk information	1.48	< 0.001	0.695	1.54	< 0.001	0.699
Pictograph risk information	1.38	< 0.001	0.687	1.51	< 0.001	0.704
Gist recall of						
Textual risk information	1.59	< 0.001	0.751	1.89	< 0.001	0.773
Pictograph risk information	1.51	< 0.001	0.738	1.76	< 0.001	0.758
Study 2A (N = 102)[d]						
Correct interpretation of survival curves	1.81	< 0.001	0.707	1.99	< 0.001	0.720
Study 2B (N = 143)						
Provide useable time-tradeoff values	1.70	0.004	0.651	1.97	< 0.001	0.677
Study 3 (N = 1234)						
Complete 3 internet-administered person-tradeoff utility elicitations	1.31	< 0.001	0.577	—	—	—

a. To facilitate comparison across scales, odds ratios have been standardized to show the effect of a 1-standard-deviation increase in Subjective Numeracy Scale or objective numeracy scores.
b. The c-index represents the total area under the receiver-operating characteristic curve.
c. Study 1 used a 5-item abbreviated scale instead of the full 8-item Subjective Numeracy Scale and a modified 10-question version instead of the full 11-item objective numeracy questionnaire of Lipkus and others.[5] See text for details.
d. Repeated-measure regressions in studies 1 and 2A included a record for each relevant question response while controlling for question item difficulty and clustering by respondent.

facilitate comparisons between numeracy scales. As shown in the set of columns on the left, in each case, rating oneself higher on the 5 questions drawn from the SNS was significantly associated with increased odds of accurate gist and verbatim recall for risk information presented textually or in pictographs.

Because it is difficult to interpret odds ratios when applied to nonbinary variables (such as our SNS scores), to illustrate the magnitude of the association, we also divided our sample into tertiles based on their average 5-item SNS ratings and compared the observed differences in the likelihood of giving a correct answer (for both gist and verbatim recall tasks) between the tertiles with the highest and lowest 5-item SNS scores. For gist recall, the lowest difference in correct answer rates was 11.7% (82.3% of participants in the highest tertile correctly recalled from pictographs which treatment reduced the need for an additional procedure the most v. 70.6% of the lowest tertile). The largest observed difference was a 26.4% difference in the percentage of respondents correctly recalling from text statistics which treatment was more likely to reduce pain. For the 12 verbatim recall questions, the average difference in the percentage of questions answered correctly was 17.4%, with 11 of the 12 questions in the range of 8.4% to 29.9% and 1 outlier question showing no effect (difference: –1.5%). Regressions using abbreviated SNS subscales showed strong predictive power for both the self-reported ability and self-reported preference subscales and distinct independent predictive power for each when both are included yet weaker overall power than a single composite. These results confirm our use of the full aggregate rating scale for our primary analyses.

The columns on the right in Table 1 show the corresponding logistic regression results for our abbreviated objective numeracy measure. As with the abbreviated SNS, objective numeracy is significantly associated with the odds of correct recall of risk information presented either textually or in pictographs. The standardized odds ratios and c-index values reported here suggest that objective numeracy measures might have slightly higher predictive power than SNS questions in this type of task, especially for gist recall. Some of this possible difference in predictive power, however, could be due to the fact that study 1 used only the 5-item abbreviated SNS instead of the full 8-item measure. Further research is clearly warranted.

STUDY 2: INTERPRETATION OF SURVIVAL CURVE GRAPHICS AND ABILITY TO COMPLETE TTO UTILITY ELICITATIONS

In study 1, a subset of the questions in the SNS proved to have high discriminatory power, effectively identifying those individuals whose poor quantitative skills limit their ability to recall textual and pictograph risk communications. The goal of study 2 (which was part of study 2 of the SNS development article[1]) was to extend this finding to 2 domains. With approximately half of our study participants (study 2A), we tested whether the complete SNS would predict the ability to correctly interpret risk information presented in a survival curve format. Survival graphs have been used with mixed success in patient risk communications in situations in which health risks extend or vary over time,[7–12] and differences in numeracy might account for some of the observed problems. With the remaining participants (study 2B), we assessed whether the SNS could identify those individuals who either cannot or choose not to complete a paper-and-pencil TTO utility elicitation. The TTO is one of the most commonly used methods to elicit people's preferences (utilities) for different health states,[13] which are key inputs to cost-effectiveness analyses in the health sector.[14,15] Previous research has demonstrated a link between objective numeracy and the ability to effectively complete TTO tasks,[4] a result we hoped to replicate.

Overall Methods

Each participant first completed the full SNS instrument, the full 11-item Lipkus and others'[5] objective numeracy measure, and an unrelated set of filler questions on the respondent's current health. Study participants were then randomized to receive either the survival curve comprehension task or the TTO task.

We used the same methods as in study 1 to create SNS and objective numeracy scores, with the exception that the objective numeracy scores varied from 0 to 11 because of the additional question. We again used logistic regression analyses to test the association of these scores to respondent performance. The specific methods and results for each substudy are described separately below.

Sample

A total of 309 people, recruited from waiting rooms at a Veterans Administration hospital, agreed to complete our survey (155 respondents in study 2A and 154 in study 2B). However, because of the length of the survey and time limits during subject recruitment, some subjects were unable to complete the full instrument. In fact, missing data problems prevented our use of nested-model analyses because too few people completed both the SNS and the objective numeracy measure. In total, 245 people both completed more than half of the SNS questions (our criterion for inclusion in the analysis) and either answered 1 or more of the survival curve comprehension questions (for those in study 2A) or had continued to answer survey questions (including filler tasks) until the start of the TTO elicitation section (study 2B). A complete discussion of the sample demographics of study 2 is presented in our companion article,[1] but we note, in particular, that the respondents' mean age was 58 years (range, 19–85) and that only 18% had a bachelor's or higher degree of education. SNS scores were again roughly normally distributed: The median SNS score was 4.13, with a 25th percentile value of 3.38 and a 75th percentile value of 4.88.

Study 2A: Survival Curves

Methods

The key section of the study 2A was a brief vignette describing a hypothetical disease that could be treated with 2 different drugs. The vignette asked respondents to imagine that 100 patients had been treated with pill A and 100 other patients with pill B. A survival graph showed how many patients out of each set would remain alive at different points in time for up to 50 years, and respondents answered a series of questions based on this graph. The slopes of the 2 survival curves varied over time, and whereas pill A had higher rates of survival initially, pill B had higher overall survival after 50 years.

Because it was easy for survey participants to refer back to the survival graph in a paper-and-pencil survey, our analysis focused on questions that required comparisons and interpretations of the outcomes data presented rather than recall of single statistics. The 4 critical questions asked 1) in what year the difference in total survival between pill A and pill B was the largest, 2) in what year the number of people alive who took pill B became larger than the number of people alive who took pill A, 3) how many people taking pill B died between year 5 and year 15, and 4) which pill had the highest overall survival after 50 years.

Data analysis. We used the same logistic regression analysis framework as in study 1. We created a long data set with 4 records for each respondent and then ran a single regression to analyze all questions simultaneously, using STATA's cluster option to obtain robust standard error estimates and adjust for the clustering of answers by each respondent. Our initial model tested for question by SNS score interactions, but as these were nonsignificant, we report estimates from reduced models containing only SNS or objective numeracy scores and indicator variables to control for question difficulty.

Results

As shown in the study 2A row of Table 1, average ratings on the SNS significantly predicted people's ability to correctly interpret our complex survival graph. To clarify the size of the effect of higher versus lower numeracy scores, we again divided our sample into 3 groups based on their average ratings on the SNS questions and compared the percentage answering each question correctly in the lowest and highest tertiles. The differences in the percentage of questions answered correctly were again large, ranging from 20% (73% of respondents in the highest tertile correctly reported which drug resulted in the highest overall survival after 50 years v. 53% in the lowest tertile) to 31% (96% v. 65% correctly identified the year in which the difference in total survival was the largest). Regressions that replaced the full SNS with individual SNS subscales (i.e., the 4 ability items or the 4 preference items) again showed strong and similarly sized individual associations, supporting our use of the composite score. Thus, the SNS appears to have high predictive value regarding people's ability to interpret and use complex risk information presented in survival graphs.

Table 1 also displays the analogous regression results for the Lipkus and others[5] objective numeracy measure. As in study 1, objective numeracy was significantly associated with task performance, again to a slightly stronger degree than the SNS.

Study 2B: TTO Utility Elicitations

Methods

The TTO task consisted of 8 questions comparing the respondent's current health to perfect health. The first 2 questions approached the tradeoff from opposite extremes by asking participants to choose between 10 additional years of life in their current health followed by painless death and (Q1) 10 additional years of life in perfect health followed by death or (Q2) 1 month of perfect health followed by death. Participants then completed 6 additional questions, which started by comparing 10 years of life in one's current health to 9 years and 10 months of life in perfect health and then progressively decreased the amount of time that the individual would be alive in perfect health to 9 years, 8 years, 7 years, 5 years, and 2 years, respectively. All tradeoffs were described both in text and using parallel horizontal bars to visually represent the amount of time under consideration.

Given the challenging nature of this task, we sought to determine whether subjective numeracy would predict whether survey participants would provide complete and consistent TTO responses. We labeled as consistent all participants who answered sufficient questions to isolate how much time they were willing to give up (if any) and whose responses were not contradictory. For example, an individual reporting that he would choose 10 years of life in his current health over 9 years of life in perfect health yet on the next question reporting that he would prefer 8 years of life in perfect health over 10 years of life in his current health would have been labeled as inconsistent. We used a logistic regression to assess the relationship between respondents' SNS scores and whether or not they provided consistent TTO responses.

Results

The literature on paper-and-pencil utility elicitations suggests that such tasks are very challenging, and ours proved to be no different. Only 51% of our subjects provided responses that yielded a consistent TTO result. We observed a clear association between numeracy and TTO performance, as demonstrated by the significant odds ratios observed for both the SNS and the Lipkus and others[5] objective numeracy measure reported in the 3rd section of Table 1. To make this result more intuitive, note that only 39% of respondents scoring themselves in the lowest tertile of SNS scores provided useable TTO values. However, 69% of respondents in the highest tertile of SNS ratings did so. A comparison of the standardized odds ratios suggests that both the SNS and the objective numeracy measure had similar degrees of association with TTO performance.

We thus confirm that both the SNS and the Lipkus and others[5] objective numeracy measure predict the ability to complete TTO utility elicitations accurately. Of the 2 SNS subscales, the perceived ability subscale had the strongest independent relation to

TTO performance (standardized odds ratio = 1.67 v. 1.48 for the preference subscale).

STUDY 3: WILLINGNESS TO MAKE TRADEOFFS IN PTO UTILITY ELICITATIONS

Given our positive results in studies 1 and 2, we sought to extend the literature on numeracy by testing for similar effects in an Internet-administered PTO utility elicitation. To our knowledge, previous research has not examined the effects of numeracy on the successful completion of the PTO. Like the TTO task, PTO elicitations quantify how people perceive the difference between 2 health states. However, instead of comparing different possible life spans, PTO tasks involve comparing 2 populations of people, and research has shown that people's responses to PTO questions are influenced by equity and fairness considerations in ways that TTO tasks generally are not.[16–18] Specifically, in PTO elicitations, people are asked to estimate the relative benefits of different health improvements by equating 2 different populations of people, each with 1 of 2 conditions. For example, people might be asked how many patients need to be cured of mild shortness of breath to bring the same amount of benefit as curing 100 patients of severe shortness of breath.[19–21] We hypothesized that subjects with lower levels of numeracy would be particularly challenged by the type of thinking required for PTO elicitations, specifically, the generation of relative ratios. Because the computerized elicitation prevented inconsistent answers (our dependent variable in study 2B), participants' difficulty or frustration with the task would most likely manifest itself in increases in the likelihood of dropout from the survey.

Methods

In this study, participants were first asked to complete the full SNS and then to complete a series of 3 PTO elicitations. Each subject imagined that he or she was the executive director of a health care organization allocating funds to groups of patients with different severities of shortness of breath (SOB). Each participant compared 3 combinations of SOB (severe v. mild, severe v. moderate, moderate v. mild), with the order varied randomly. For example, one of the elicitations asked subjects to choose between 2 treatment programs that cured the same number of patients, one that cured patients whose SOB was so severe that they had difficulty walking from the bedroom to the bathroom, and another that cured patients with more moderate SOB (e.g., SOB occurring if they walked a city block). For subjects who chose the treatment program that cured severe SOB, we prompted for an indifference point by asking how many patients would have to be cured of moderate SOB to make that program seem equally good as a treatment program that cured 100 patients of severe SOB. This methodology was approved by the Institutional Review Board.

We again used a simple logistic regression to test whether respondents' SNS ratings affected their willingness and/or ability to complete the PTO tasks and assessed effect size by calculating the percentage of respondents in the lowest and highest tertile of SNS scores who completed all 3 PTO elicitations.

Results

We recruited survey participants for this study using a new, demographically balanced sample drawn from the same commercially administered Internet panel used in study 1. E-mail invitations were sent to 10,966 individuals, and 1270 (a 12% response rate) clicked on the provided link to access the experimental Web site. Of these, 1234 completed most of the SNS questions and proceeded to the PTO elicitation section of the survey. The SNS scores were approximately normally distributed, with a median score of 4.5, 25th percentile at 3.75, and 75th percentile at 5.13.

Because participants who dropped out of the survey before completion did not answer demographics questions, it is impossible to know whether the dropouts differed from survey completers in other important ways. However, it is worth noting that the individuals who did complete the PTO task had a broad range of demographic characteristics. The average age of completers was 42 years, with a range of 18 to 88 years. About 57% of completers were female, with 16% having only a high school education or less but 34% with a bachelor's degree or higher. Slightly less than 19% identified themselves as belonging to a racial or ethnic minority.

Respondents' scores on the SNS were significantly associated with their willingness and/or ability to complete all 3 PTO tasks included in this study, as demonstrated by the regression results reported in the last section of Table 1. When we reran the analysis using the SNS subscales, the preference subscale had the stronger individual association with task completion (standardized odds ratio = 1.33 v. 1.21 for the ability subscale). To assess the magnitude of

the numeracy effect, we again broke our sample into 3 groups based on SNS scores and compared the lowest and highest groups. Whereas only 73% of participants scoring in the lowest tertile of the SNS score distribution were able to complete this utility elicitation task, this percentage increased to 84% for those scoring in the highest tertile. Similarly significant patterns were observed when we relaxed the criterion to identify those participants completing 2 PTO tasks instead of all 3.

GENERAL DISCUSSION

Survey participants' responses to the questions on the SNS, a self-assessment of both numerical aptitude and preferences for numerical information, are significantly related to performance on numerical information–processing tasks relevant to medical decision making. Individuals with low numeracy levels (as measured by both the SNS and the objective numeracy measure) are less likely to 1) recall risk information presented textually or in pictographs, 2) comprehend risk information displayed in survival curves, and 3) effectively complete utility elicitation measures.

It is important to note that in addition to having significant predictive validity, the SNS offers several qualitative improvements over existing objective numeracy measures. As detailed in our companion article,[1] a survey that included the SNS was faster to complete, evoked fewer negative reactions, and produced greater expressed willingness to participate in further research than a survey that contained the objective numeracy questions of Lipkus and others.[5] The SNS consists of questions addressing both perceived ability and preferences for numeric information, subscales that were shown to be relevant in our studies. Furthermore, because the SNS records subjective perceptions rather than accuracy on specific tasks, the SNS both has lower missing data rates and should be less sensitive to potential biases caused by missing data.

Although it was neither our intent nor our expectation that the SNS would match objective tests of numerical ability, the SNS did hold up reasonably well compared with an established objective numeracy measure. Mathematics tests, such as the standard objective numeracy questions, assess one's ability to do a specific task. Our self-perceptions of skill are necessarily 1 step removed from aptitude. Nevertheless, it seems clear that the SNS significantly predicts the same behaviors as objective numeracy measures do. Just how equivalent the SNS and objective numeracy measures will be in a particular application is, of course, context dependent. Similarly, the minimum SNS score required to predict satisfactory performance is likely to vary from situation to situation. Still, we believe that the SNS can effectively proxy for objective numeracy measures in many circumstances, reducing respondent burden while maintaining significant predictive validity.

Several factors limit the strength of our conclusions. The fact that study 1 used only 5 of the 8 SNS questions makes it difficult to compare its results with others and likely reduced its predictive power versus objective numeracy. Although we did stratify the e-mail invitation lists for studies 1 and 3 to mirror the U.S. census population with respect to age, gender, and income, these studies' low initial response rates could affect the generalizability of their conclusions to other contexts. Still, our response rates, while low by mailed survey standards, are similar to those obtained in other Internet-based surveys that use opt-in panels.[22] Studies 2A and 2B used participants recruited from Veterans Administration hospital waiting rooms—again, a particular sample whose unique characteristics (e.g., lower education) might possibly have influenced our results and hence limit their generalizability. In particular, those people who declined to participate or failed to complete the surveys might have had different levels of numeracy than those who did participate, and their absence could have increased or decreased the predictive power of the SNS. Still, despite whatever sample peculiarities might have existed, we found similar effects of self-reported numeracy in each study.

We believe that the SNS is an important research tool for anyone studying risk communication, utility elicitations, and other similar situations in which quantitative aptitude is an essential task component. It represents a different tradeoff in terms of usability versus descriptive precision than is offered by currently available numeracy measures. Objective numeracy measures are likely to yield somewhat higher predictive power in some contexts and may thus be necessary for certain applications. In many cases, however, the extra time required and subject irritation caused by mathematics tests may more than offset their slightly increased statistical efficacy. However, regardless of which method is selected, it is clear that measuring numerical facility can significantly further our understanding of patients' medical decision making, and we encourage broader adoption of numeracy measures in health services research.

REFERENCES

1. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry H, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale (SNS). Med Decis Making. 2007;27:672–80.

2. Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med. 1997;127(11):966–72.

3. Schapira MM, Davids SL, McAuliffe TL, Nattinger AB. Agreement between scales in the measurement of breast cancer risk perceptions. Risk Anal. 2004;24(3):665–73.

4. Woloshin S, Schwartz LM, Moncur M, Gabriel S, Tosteson AN. Assessing values for health: numeracy matters. Med Decis Making. 2001;21(5):382–90.

5. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001;21(1):37–44.

6. Stata Statistical Software. 8th ed. College Station (TX): Stata Corporation; 2003.

7. Mazur DJ, Hickam DH. The effect of physician's explanations on patients' treatment preferences: five-year survival data. Med Decis Making. 1994;14(3):255–8.

8. Mazur DJ, Hickam DH. Patients' and physicians' interpretations of graphic data displays. Med Decis Making. 1993;13(1):59–63.

9. Mazur DJ, Hickam DH. Interpretation of graphic data by patients in a general medicine clinic. J Gen Intern Med. 1990;5(5): 402–5.

10. Mazur DJ, Hickam DH. Five-year survival curves: how much data are enough for patient-physician decision making in general surgery? Eur J Surg. 1996;162(2):101–4.

11. Armstrong K, FitzGerald G, Schwartz JS, Ubel PA. Using survival curve comparisons to inform patient decision making can a practice exercise improve understanding? J Gen Intern Med. 2001;16(7):482–5.

12. Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. Med Decis Making. 2002;22(1):76–83.

13. Torrance GW. Utility approach to measuring health-related quality of life. J Chron Dis. 1987;40(6):593–603.

14. Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-effectiveness in Health and Medicine. JAMA. 1996;276 (14):1172–7.

15. Gold MR, Siegel JE, Russell LB, Weinstein M, eds. Cost-effectiveness in Health and Medicine. New York: Oxford University Press; 1996.

16. Nord E, Pinto-Prades JL, Richardson J, Menzel P, Ubel PA. Incorporating societal concerns for fairness in numerical valuations of health programmes. Health Econ. 1999;8(1):25–39.

17. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Value measurement in cost-utility analysis: explaining the discrepancy between rating scale and person trade-off elicitations. Health Policy. 1998;43(1):33–44.

18. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Individual utilities are inconsistent with rationing choices: a partial explanation of why Oregon's cost-effectiveness list failed. Med Decis Making. 1996;16(2):108–16.

19. Damschroder LJ, Baron J, Hershey JC, Asch DA, Jepson C, Ubel PA. The validity of person tradeoff measurements: a randomized trial of computer elicitation versus face-to-face interview. Med Decis Making. 2004;24(2):170–80.

20. Damschroder LJ, Baron J, Hershey JC, Asch DA, Jepson C, Ubel PA. Does being on the hot seat change people's valuation of health conditions? Med Decis Making. 2003;23(6):557.

21. Damschroder LJ, Miklosovic ME, Ubel PA. Quality of life values change when people are primed to think about how they adapt to difficult situations. Med Decis Making. 2003;23(6):563.

22. Couper MP. The promise and perils of Web surveys. In: Westlake A, Manners T, Rigg M, eds. The Challenge of the Internet. London: Association of Survey Computing; 2001. p ix-194.