

The Numeracy Understanding in Medicine Instrument: A Measure of Health Numeracy Developed Using Item Response Theory

Marilyn M. Schapira, MD, MPH, Cindy M. Walker, PhD, Kevin J. Cappaert, MA, Pamela S. Ganschow, MD, Kathlyn E. Fletcher, MD, MA, Emily L. McGinley, MS, MPH, Sam Del Pozo, MA, Carrie Schauer, BSW, Sergey Tarima, PhD, Elizabeth A. Jacobs, MD, MAPP

Background: Health numeracy can be defined as the ability to understand and apply information conveyed with numbers, tables and graphs, probabilities, and statistics to effectively communicate with health care providers, take care of one's health, and participate in medical decisions. **Objective:** To develop the Numeracy Understanding in Medicine Instrument (NUMi) using item response theory scaling methods. **Design:** A 20-item test was formed drawing from an item bank of numeracy questions. Items were calibrated using responses from 1000 participants and a 2-parameter item response theory model. Construct validity was assessed by comparing scores on the NUMi to established measures of print and numeric health literacy, mathematic achievement, and cognitive aptitude. **Participants:** Community and clinical populations in the Milwaukee and Chicago metropolitan areas. **Results:** Twenty-nine percent of the 1000 respondents were Hispanic, 24% were non-Hispanic white, and 42% were non-Hispanic black. Forty-one percent had no more than

a high school education. The mean score on the NUMi was 13.2 ($s = 4.6$) with a Cronbach α of 0.86. Difficulty and discrimination item response theory parameters of the 20 items ranged from -1.70 to 1.45 and 0.39 to 1.98 , respectively. Performance on the NUMi was strongly correlated with the Wide Range Achievement Test-Arithmetic (0.73 , $P < 0.001$), the Lipkus Expanded Numeracy Scale (0.69 , $P < 0.001$), the Medical Data Interpretation Test (0.75 , $P < 0.001$), and the Wonderlic Cognitive Ability Test (0.82 , $P < 0.001$). Performance was moderately correlated to the Short Test of Functional Health Literacy (0.43 , $P < 0.001$). **Limitations:** The NUMi was found to be most discriminating among respondents with a lower-than-average level of health numeracy. **Conclusions:** The NUMi can be applied in research and clinical settings as a robust measure of the health numeracy construct. **Key words:** decision aids; shared decision making; risk communication; risk perception; health literacy. (*Med Decis Making* 2012;32:851-865)

Received 28 May 2011 from the Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (MMS); Department of Educational Psychology, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin (CMW, KJC); Department of Medicine, John H. Stroger Jr. Hospital of Cook County and Rush University Medical Center, Chicago, Illinois (PSG, SDP); Department of Medicine, Clement J. Zablocki VA Medical Center, and Medical College of Wisconsin, Milwaukee, Wisconsin (KEF); Center for Patient Care and Outcomes Research, Medical College of Wisconsin, Milwaukee, Wisconsin (ELM, CS); Institute of Health and Society, Medical College of Wisconsin, Milwaukee, Wisconsin (ST); Department of Medicine, University of Wisconsin, Madison, Wisconsin (EAJ). This study was funded by the National Cancer Institute of the National Institutes of Health (NCIR01CA115954). The work was presented at the 2010 annual meeting of the Society for Medical Decision Making, Toronto, Canada. The RED-Cap database used in this work was supported by a Clinical and Translational Science Institute grant (1UL1-RR031973-01). The study was funded by the National Cancer Institute. The funder did not have a role in the collection, analysis, or interpretation of the data. Revision accepted for publication 15 March 2012.

Health numeracy can be defined as the ability to understand and apply information conveyed with numbers, tables and graphs, probabilities, and statistics to effectively communicate with health care providers, take care of one's health, and participate in medical decisions.¹⁻⁵ Numbers and numeric-based concepts are integrated throughout the spectrum of health-related communication and decision making. Knowledge and understanding regarding the cause, incidence, and natural history of disease are associated with health

Address correspondence to Marilyn M. Schapira, MD, MPH, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104; e-mail: mschap@upenn.edu.

DOI: 10.1177/0272989X12447239

numeracy.⁶⁻⁸ Furthermore, numeric skills such as risk perception, estimates of probabilistic outcomes, and the ability to weigh risks and benefits are central to theoretical frameworks of health behavior, such as the health belief model and normative theories of medical decision making.⁹⁻¹¹ A growing body of evidence supports the role of health numeracy in the adoption of health protective behaviors.^{7,8,12-15} Although the mechanism has not been fully delineated, health numeracy has been associated with increased self-efficacy,¹² improved self-management of chronic disease,¹³⁻¹⁵ and the assessment of values and preferences in the context of shared decision making.¹⁶⁻¹⁸

The ability to measure health numeracy among individuals or populations has both research and clinical applications in the field of health communication and medical decision making. A valid measurement of health numeracy supports the potential to tailor communication and shared decision making to the level of understanding of a given patient or population.^{2,19} Existing health numeracy measures have primarily been developed using classical test theory (CTT) and in majority populations.²⁰⁻²⁷ These measures have been helpful in moving the field forward, as they have supported an association between health numeracy and outcomes associated with informed decision making.^{7,8,12-14,16} However, existing measures emphasize only components of the full construct of health numeracy, be it number sense, risk communication and probability, or the interpretation of medical study results.^{20-23,26,27} For some purposes, a measure that reflects the full spectrum of health numeracy skills may be optimal. Furthermore, existing measures have not been developed using cross-cultural approaches, making it unclear if these measures are valid for certain populations such as Hispanics. The use of IRT scaling methods offers several useful features in scale development.²⁸⁻³⁰ First, IRT psychometric methods support the development of computer-adaptive test (CAT) modalities, an approach that can decrease respondent burden while increasing the accuracy of the measure.^{31,32} In addition, IRT methods allow for the assessment of measurement bias through use of differential item functioning (DIF) analyses. This approach is valuable in the development and evaluation of cross-cultural measurement tools.^{33,34} The objectives of this study are 1) to develop a measure of health numeracy, the Numeracy Understanding in Medicine Instrument (NUMi), which uses IRT scaling methods, is based on an empirically derived framework, and is cross-culturally equivalent across Hispanic and non-Hispanic populations, and

2) to create a robust item bank for use in a CAT version of the NUMi.

METHODS

Overview

The first stage of the study was the development of a framework for the health numeracy construct and the generation and calibration of a large item bank ($n = 110$) to assess the health numeracy construct. The second stage involved the formation of a 20-item paper-and-pencil test through purposeful selection of items from the full item bank. Content and construct validity of the 20-item measure was evaluated and a scoring system proposed. An overview of the use of the study population for various stages of the study is provided (Figure 1).

Development of Theoretical Framework

A theoretical framework for the construct of health numeracy was developed drawing on previous work of our group and others.¹⁻⁶ The definition of health numeracy that emerged from this work was the following: the ability to understand and apply information conveyed with numbers, tables and graphs, probabilities, and statistics to effectively communicate with health care providers, take care of one's health, and participate in medical decisions. The framework was expanded to include cross-cultural considerations through qualitative studies in Hispanic clinical and community populations.³⁵ This formative work in the Hispanic population highlighted the importance of several key concepts for patients, including the desire for health information to be specific to one's ethnic group and community and the desire to understand the meaning behind numbers. The theoretical framework used in the development of the measure includes 4 domains of numeric skills that are widely applied in health: number sense, tables and graphs, probability, and statistics. These domains were used as the basis of scale development; the operational definitions of each skill area are provided in Table 1.

Item Generation

A test specification table was developed to represent the set of skills composing the health numeracy construct (Table 1) and the health care context in

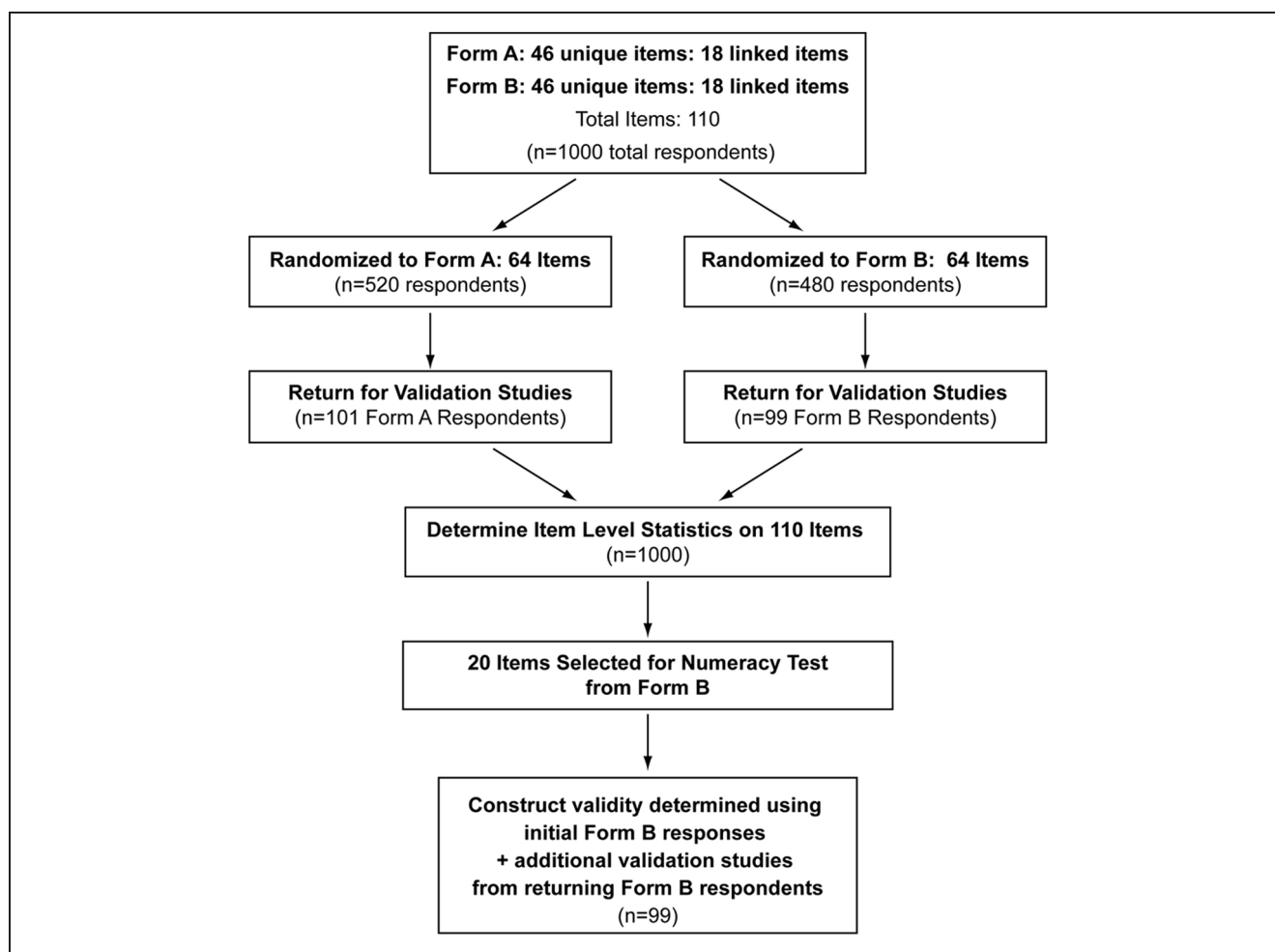


Figure 1 Flow chart of study population used to obtain psychometric data.

which the skills are applied. An expert panel was convened to review the health numeracy framework, the test specification table, and the initial items generated. The expert panel consisted of 5 members: 1 clinician who was bilingual and practiced in the Chicago community, 2 clinician-investigators (1 with research expertise in the area of patient-physician communication and 1 with expertise in health print and numeric literacy), an expert in the field of adult education, and an expert in the field of cross-cultural survey research. The 110 items generated were then evaluated by conducting cognitive interviews with a sample of 48 English-speaking Hispanic and non-Hispanic participants who were recruited from community and clinical populations in the Milwaukee and Chicago metropolitan areas. The interview used think-aloud techniques and probe questions to ascertain the respondent's understanding of the question,

interpretation of the question, and understanding of the response options.³⁶ Each of the 110 items underwent a cognitive interview process with at least 2 participants. The interviews were individually conducted by members of the research team. Responses were reviewed by the investigative team and items modified accordingly. Items that required significant modification were then retested in a subsequent cognitive interview.

Study Protocol for Obtaining Psychometric Data

The final set of items composing the item bank ($n = 110$) were divided into 2 parallel forms (A and B) with 64 items per form to reduce the respondent burden (Figure 1). Each form contained 18 common linking items and 46 unique items. Unique items were selected by ensuring that the content domain and

Table 1 Definition of Health Numeracy and Identified Skills Within Each Health Numeracy Domain

Construct	Definition
Health numeracy	The ability to understand and apply information conveyed with numbers, tables and graphs, probabilities, and statistics to effectively communicate with health care providers, take care of one's health, and participate in medical decisions
Number sense	The ability to represent, order, compute, and estimate numbers and to understand how fractions and decimals relate to each other and how each can best be used to describe a particular health-related situation <ul style="list-style-type: none"> ● Understand a percentage as a representation of risk and risk reduction ● Rank percentages in order of magnitude ● Rank fractions in order of magnitude ● Understand the relationship of numbers across whole number, fraction, and percentage formats ● Count and estimate whole numbers and numbers with decimals ● Use time and dates ● Interchange metrics, such as milligrams and grams ● Understand the concept of class inclusion judgments
Tables and graphs	The ability to read and use 1- and 2-dimensional graphic forms, such as tables, charts, and graphs, and to apply these skills in a health situation <ul style="list-style-type: none"> ● Use a table to identify the appropriate information, given 2 determinants ● Use a chart to abstract health goals and information ● Use a pie graph to identify proportions ● Use a histogram to compare magnitudes ● Interpret a line graph ● Interpret a pictograph
Probability	The ability to understand concepts related to probability distributions, independent events, and conditional probability and to compute the probability of an event occurring <ul style="list-style-type: none"> ● Convert a risk from a probability statement to a frequency statement ● Understand the probability of 2 independent events ● Understand the probability of 2 nonindependent events ● Understand the importance of pretest probability in diagnostic testing ● Understand that the range of probability is between 0.0 and 1.0
Statistics	The ability to understand descriptive statistics, concepts of inference, random sampling, experimental design, and measures of uncertainty and to interpret such data to make informed decisions about health <ul style="list-style-type: none"> ● Understand the concepts relating to the following aspects of scientific study design and interpretation: sample size, placebo, randomization, causality, inference ● Understand what measures of central tendency and variation convey; general understanding of normal variation among populations ● Understand the meaning of statistical significance and the concept that a reported finding may be due to chance alone ● Understand the concept of uncertainty in estimates and uncertainty as part of the scientific process; general understanding of confidence intervals

Note: These items define the overall construct of health numeracy in each of the 4 domains. The table also lists a select number of skills that compose the test specification table within each domain and guide item generation.

perceived level of difficulty were similar on both forms to create 2 parallel forms of the test. The use of linked items increased the ability to calibrate a large number of items without requiring all respondents to answer all items.

Items were tested among 1000 respondents across the 2 forms. A purposeful sample was obtained from community and clinical populations in the

Milwaukee and Chicago metropolitan areas. Recruitment approaches included newspaper advertisements in community papers, flyer postings at local colleges and community centers, and recruitment booths in community centers and clinical settings. Recruitment booths were staffed by bilingual study personnel. Inclusion criteria included an age of 21 years or older and the ability

to read English. Exclusion criteria included poor eyesight as indicated by a Snellen chart eye test with a corrected vision of less than 20/50. Participants who met enrollment criteria and wished to participate were given a date, time, and location for the testing session.

The items were administered in a classroom setting in groups of up to 60 persons. Baseline assessments included sociodemographic information and the print literacy version of the Short Test of Functional Health Literacy in Adults (S-TOFHLA) and the Wonderlic Cognitive Ability Test. The S-TOFHLA is a reading comprehension test with a potential score of 0 to 36 and classifies respondents as being of inadequate functional health literacy (0–16), marginal functional health literacy (17–22), and adequate health literacy (23–36).³⁷ The Wonderlic is a 50-item test that evaluates respondents' aptitude by assessing their ability to reason and use logic through a series of multiple-choice problems that they are asked to solve.³⁸ Participants were instructed to leave the numeracy test items blank rather than guess if they did not know the answer.

All participants were given the option to have items read aloud, in order to include participants with low reading literacy. A separate classroom setting was used for those participants. Each participant in these sessions was given a copy of the items so that the graphic illustrations and text could be viewed as items were read aloud. Participants who had items read to them did not take the Wonderlic.

An informed consent was read aloud and a printed copy provided to the participants prior to the session. Participants were given \$50 in cash at the conclusion of the session to compensate them for their time. The protocol was approved by the institutional review boards at the Medical College of Wisconsin, University of Wisconsin–Milwaukee, and Cook County Health and Hospital System.

Calibration of Items

Responses to the items were exported to a REDCap database and downloaded into a SAS software file (SAS Institute, Cary, NC) for analysis.³⁹ The IRT software BILOG was used to calibrate items using a unidimensional dichotomous 2-parameter IRT model,⁴⁰ which estimates both a difficulty parameter (beta) and a discrimination parameter (alpha). The 2-parameter IRT model differs from the 1-parameter model (similar to the Rasch model) in that it allows the discrimination parameter to vary between items, providing greater flexibility in allowing the model

to fit the data. A theoretical disadvantage of using the 2-parameter model in comparison to the 1-parameter or Rasch model is the lack of a one-to-one correspondence between the number correct on the test and the estimate of θ because each item is weighted somewhat differently according to its level of discrimination.^{29,30} The mathematical representation of the 2-parameter model is presented in equation 1, where θ represents the ability of respondents, a represents the discrimination of an item, and b represents the difficulty of an item.

$$P(X=1|\theta) = \frac{1}{1 + e^{-1.7a(\theta-b)}} \quad (1)$$

Estimated a priori latent trait scoring was used to obtain parameters. In addition to the IRT parameters, CTT estimates of difficulty (as measured by the proportion of examinees that obtained the correct answer to the item) and discrimination (as measured by the item-total correlation) were calculated, as was Cronbach α for the total scale. All item-level statistics were considered to identify items with a range of difficulty level and a high degree of discrimination.

Test Formation

The test was formed by identifying 20 items from form B that had a range of difficulty, high discrimination, and a range of content. Form B was used because it had a higher number of items with desirable characteristics compared to form A. Choosing all items from one form was necessary in order to use response data from the baseline assessment to obtain a total score in the validation analyses. Difficulty was assessed with the IRT beta parameter, which typically ranges from –3.0 to 3.0, with increasing values indicating a harder item. Discrimination was assessed with the IRT alpha parameter, which typically ranges from 0 to 3, with increasing values indicating a more discriminating item. Attempts were made to choose items with higher levels of discrimination (0.80 or above) whenever possible. The final version of the 20-item NUMi includes 5 items from each of the 4 content areas: number sense, tables and graphs, probability, and statistics.

Evaluation of Validity

A random sample of 200 of the initial 1000 participants was recruited to complete a validation component of the study. Total scores and ability, as determined by responses previously provided (at the first study visit) to the 20 items that composed

the NUMi, were compared to existing measures of health print literacy, aptitude, and numeracy. The sample of participants returned for a second study visit and responded to the following additional validation measures: the Wide Range Achievement Test–Arithmetic (WRAT-A), consisting of 40 math problems⁴¹; the Lipkus Expanded Numeracy Scale, consisting of 11 items²¹; and the Medical Data Interpretation Test (MDIT), consisting of 18 items.²³ Responses from these measures were linked to the data from the first study visit. These assessments were not included in the first study visit due to concerns about respondent burden. Of the 200 respondents recruited for the additional validation measures, 99 had originally responded to form B and thus had data available for use to calculate the NUMi score (Figure 1).

We hypothesized that if performance on the NUMi were a valid measure of health numeracy a positive correlation would be found with existing measures of numeracy, including the WRAT-A, Lipkus, and MDIT. Furthermore, we expected to see a positive correlation to cognitive aptitude as measured by the Wonderlic. We also hypothesized that divergent validity would be demonstrated by a weaker correlation between the NUMi and print health literacy as measured by the S-TOFHLA. Although print literacy and numeric health literacy are correlated, the skills required for print literacy represent a different component of health literacy than those required to process and apply numerical information.⁴² Participants with a S-TOFHLA score of less than 17 ($n = 43$) were excluded from the validation sample because it was required that respondents be able to read the items for the remaining measures (WRAT-A, Lipkus, MDIT).

The NUMi underwent additional evaluation for content validity. The expert panel (original panel with the addition of panelists with health numeracy expertise) was asked to provide feedback on the measure. The purpose of this level of evaluation was to ascertain whether the reduction of the test to 20 items was successful in creating a measure that captured the scope of the health numeracy construct as we had defined it. Respondents were asked whether the items in each domain reflected the theoretical definition of the domain that was presented for each content area. Moreover, respondents were asked to comment on whether the NUMi reflected the overall definition of health numeracy.

Further analyses were conducted to evaluate for DIF across groups, evaluate for unidimensionality

of the measure, and test whether the model fit was improved using the 2-parameter compared to the 1-parameter IRT model. Standard IRT methods^{29,30,33,34,43,44} were used for these additional analyses with details presented with the results in sections below.

RESULTS

Study Population

One thousand participants were recruited to obtain item-level psychometric data on the full item bank. Participants self-identified as 45% white and 44% black. Twenty-nine percent were Hispanic. Sixty percent were female. Eight percent had inadequate or marginal health literacy as measured by the S-TOFHLA. Forty-one percent had no more than a high school–level education. The Wonderlic score (with a potential range of 0 to 50) had a lower mean score for the study population ($\bar{x} = 17.5$, $s = 8.7$) than the published norms of working adults in the United States ($\bar{x} = 21.7$, $s = 7.6$, $P < 0.01$).³⁸ The items were read aloud to 46 (4.6%) respondents. A sample of 200 responded to additional validation measures (99 of which had initially responded to form B) (Table 2).

Psychometric Properties of the NUMi

Item parameters for the NUMi were calculated using CTT and IRT statistics (Table 3). Typically, IRT difficulty parameters range from -3.0 to 3.0 , and IRT discrimination parameters range from 0 to 3.0, with a higher number indicating a more difficult or discriminating item, respectively. As the table illustrates, only a small number of the items on the NUMi would be considered very difficult items. Items 10 (probability domain, understanding risk reduction) and 13 (statistics domain, interpreting a P value) are the most difficult items, with IRT difficulty parameters of 1.45 and 1.19, respectively. Most items were highly discriminating. The least discriminating items were items 7 (probability domain, understanding the relationship of short- and long-term risk of mortality) and 10 as described above, with IRT discrimination parameters of 0.42 and 0.39, respectively.

The test information function (TIF) is a function of ability, θ , and provides a summary of information provided by the full test.^{29,30} It is obtained by summing the item information function, across all items

Table 2 Characteristics of the Study Population

Demographics	Total (n = 1000) n (%)	Form A (n = 520) n (%)	Form B (n = 480) n (%)	Validation Sample (n = 99) n (%)
Sex				
Male	399 (39.9)	204 (39.2)	195 (40.6)	34 (34.3)
Female	599 (59.9)	315 (60.6)	284 (59.2)	65 (65.7)
Missing	2 (0.2)	1 (0.2)	1 (0.2)	0
Age, years				
< 45	536 (53)	277 (53.3)	259 (54)	51 (51.5)
45–59	329 (33)	178 (34.2)	151 (31.4)	36 (36.4)
60–74	119 (12)	58 (11.2)	61 (12.7)	11 (11.1)
≥ 75	16 (2)	7 (1.4)	9 (1.9)	1 (1.0)
Race				
White	448 (44.8)	212 (40.8)	236 (49.2)	63 (63.6)
Black and/or African American	438 (43.8)	252 (48.5)	186 (38.7)	29 (29.3)
American Indian and Alaska Native	14 (1.4)	5 (1.0)	9 (1.9)	1 (1.0)
Asian	38 (3.8)	22 (4.2)	16 (3.3)	4 (4.0)
Native Hawaiian and other Pacific Islander	3 (0.3)	2 (0.4)	1 (0.2)	0
Multiple races	17 (1.7)	8 (1.5)	9 (1.9)	0
Missing	42 (4.2)	19 (3.7)	23 (4.8)	2 (2.0)
Ethnicity				
Hispanic/Latino	290 (29)	137 (26.4)	153 (31.9)	38 (38.4)
Missing	7 (0.7)	3 (0.6)	4 (0.8)	1 (1.0)
Test of Functional Health Literacy in Adults score				
0–16 (inadequate literacy)	43 (4)	19 (3.7)	24 (5)	2 (2.0)
17–22 (marginal literacy)	43 (4)	25 (4.8)	18 (3.8)	2 (2.0)
23–36 (adequate literacy)	914 (92)	476 (91.5)	438 (91.3)	95 (96.0)
Education				
Up to 12 years	409 (40.9)	215 (41.4)	194 (40.4)	31 (31.3)
Some college	274 (27)	149 (28.7)	125 (26)	23 (23.2)
4-year college or more	316 (32)	155 (29.8)	161 (33.5)	45 (45.5)
Missing	1 (0.1)	1 (0.2)	0	0
Wonderlic Cognitive Ability Test				
< 10	175 (17.5)	95 (18.3)	80 (16.7)	14 (14.1)
10–19	435 (43.5)	252 (48.4)	183 (38.1)	42 (42.4)
20–29	244 (24)	108 (20.8)	136 (28.3)	22 (22.2)
30–39	102 (10)	46 (8.8)	56 (11.7)	19 (19.2)
40–50	7 (1)	5 (1)	2 (0.4)	2 (2.0)
Missing	37 (4)	14 (2.7)	23 (4.8)	0
Item administration				
Read aloud to respondent	46 (4.6)	16 (3.1)	30 (6.3)	98 (99.0)
Self-administered by respondent	954 (95.4)	504 (96.9)	450 (93.7)	1 (1.0)

Note: The total sample includes 290 Hispanics (29%), 418 non-Hispanic Blacks (42%), and 239 non-Hispanic White (24%) participants. The Form B sample includes 153 Hispanics (32%), 175 non-Hispanic Blacks (36%), and 127 non-Hispanic White (26%) participants. These 3 groups from the Form B sample were used in the differential item functioning analyses.

on the test, using the following mathematical formula:

$$I(\theta) = D^2 a_i^2 P_i Q_i, \quad (2),$$

where $D = 1.7$, a_i = the discrimination parameter of item i , P_i = the probability of obtaining the correct response to item i , as a function of θ , as expressed in equation 1, and $Q_i = 1 - P_i$. The TIF demonstrates the relationship between ability on the x -axis and

the information provided by the test on the y -axis. A feature of IRT mathematical models is that the ability level of the respondent and the difficulty level of an item are represented on the same scale (represented by the x -axis on the TIF). The TIF for the NUMi peaks at an ability level of -1.0 , indicating that the test is providing the most information (and is most discriminating) at an ability level that is below average for our study population (Figure 2).

Table 3 Item-Level Analysis of the Numeracy Understanding in Medicine Instrument Questions

	Classical Test Theory		Item Response Theory		
	Difficulty, 0 to 1	Discrimination, 0 to 1	Difficulty, -3.0 to 3.0	Discrimination 0 to 3.0	
Number sense					
1	Range / blood sugar goal in diabetic	.85	.38	-1.28	1.37
2	Scale / reporting pain	.86	.51	-1.09	1.40
3 ^a	Frequency format / side effect	.76	.51	-0.85	0.87
4	Ordering numbers / test results	.76	.42	-0.71	1.32
5	Measurement / dosing medication	.52	.40	-0.18	0.73
Probability					
6	Randomization / study participation	.86	.39	-1.27	1.27
7 ^{a,b}	Class inclusion judgments / life expectancy	.57	.33	-.59	0.42
8	Small risks / side effects	.49	.53	0.20	0.62
9 ^a	Calculating probability / screening tests	.46	.47	0.002	0.81
10	Relative risk reduction / cancer recurrence	.30	.33	1.45	0.39
Statistics					
11	Uncertainty / 95% confidence interval of treatment efficacy	.64	.68	-.84	1.40
12 ^a	Statistical significance / treatment efficacy	.58	.52	0.06	0.63
13 ^a	<i>P</i> value / interpretation of study results	.27	.51	1.19	0.85
14	Sample size implications / interpretation of study results	.77	.33	-1.70	0.53
15	Causation v. association / interpretation of study results	.68	.50	-0.71	0.57
Tables and graphs					
16	Bar graphs / interpretation of population statistics	.86	.43	-1.22	1.33
17	Interpreting decimals / reading a digital thermometer	.92	.45	-1.20	1.98
18 ^{a,b}	Reading a table / interpreting a nutrition label	.82	.29	-1.31	0.73
19	Interpreting survival curve / survival estimates	.77	.55	-0.80	1.12
20	Small risk formats / pictogram	.48	.48	0.55	0.68

Note: This table presents item-level statistics for the 20 items included on the Numeracy Understanding in Medicine Instrument. Psychometric data reflecting the difficulty and discrimination of each item were determined using both classical test theory (CTT) statistics and item response theory (IRT) methods. The statistics were based on 520 respondents for form A items and 480 respondents for form B items. A total of 1000 respondents answered the linked questions that were on both forms. In CTT statistics, the difficulty parameter is determined as the percentage who answered the item correctly, with higher values indicating easier items. In IRT, the difficulty parameter is determined from IRT models and represents the difficulty level at which 50% of respondents are anticipated to answer the question correctly. Higher-level IRT difficulty parameters indicate harder questions. In CTT, item discrimination is determined by the correlation between a correct item response and the total score. In IRT, item discrimination is determined by IRT models and represents the ability of the item to discriminate between those with and without the ability level that equals the difficulty of the given item. In a 2-parameter model such as that used in this case, the discrimination value can vary between items. This table illustrates that the items have a range of difficulty and discrimination, as illustrated by both CTT and IRT scaling methods.

a. Linked items: the linked items were administered to respondents in both form A and form B and therefore had a larger sample size from which to obtain item statistics.

b. Revised items: these items were revised and retested in the validation sample of 200 respondents.

Results of Validity Evaluation

The construct validity of the NUMi is supported by the strong Pearson correlations of performance on the NUMi with the WRAT-A (0.73, $P < 0.001$), the Lipkus (0.69, $P < 0.001$), the MDIT (0.75, $P < 0.001$), and the Wonderlic (0.82, $P < 0.001$). As hypothesized, the NUMi demonstrates a more moderate correlation with the print literacy measured by the S-TOFHLA (0.43, $P < 0.001$). The correlations were similar whether the total score (0 to 20) or ability level (θ) was evaluated (Table 4). Validity of the NUMi is also supported by the association observed between socioeconomic characteristics

and performance on the NUMi, with increasing levels of education associated with greater ability on the NUMi (Table 5).

Review of the 20-item NUMi by the expert panel indicated that the items selected adequately represented the domain of health numeracy with the exception that 2 items in the statistics subdomain were noted to have some redundancy in content. Therefore, a replacement item was identified based on content and discrimination and difficulty parameters. Minor modifications were also made to the wording of some items on the final version based on feedback from the expert panel. The substitution did not affect the shape of the TIF.

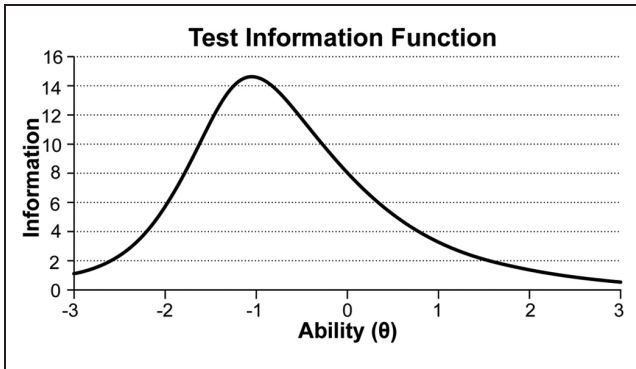


Figure 2 Test information function of the numeracy understanding in medicine instrument. The x-axis represents the degree of ability level—that is, degree of latent trait of the respondent. The y-axis represents the information, or discrimination, that the test provides at each level of respondent ability.

Table 4 Construct Validity of the Numeracy Understanding in Medicine Instrument

	Classical Test Theory Measure: NUMi Score	Item Response Theory Measure: θ
NUMi total score	—	0.98
Estimated ability	.98	—
Lipkus	.69	.69
MDIT	.72	.75
WRAT-A	.70	.73
S-TOFHLA	.43	.43
Wonderlic	.80	.82

Note: Theta (θ) is the latent trait ability of the respondent as determined by responses to the items and the item response theory model. The table presents Pearson correlation coefficients between the following measures: NUMi, Numeracy Understanding in Medicine Instrument; Lipkus: Lipkus Expanded Numeracy Scale²¹; MDIT, Medical Data Interpretation Test²³; WRAT-A, Wide Range Achievement Test in Arithmetic⁴¹; S-TOFHLA, Short Test of Functional Health Literacy in Adults³⁷; Wonderlic, Wonderlic Cognitive Ability Test.³⁸

DIF Analysis

Two sets of exploratory DIF analyses were conducted using SIBTEST (simultaneous item bias), a nonparametric statistical procedure, to test for bias on the 20 items selected for the NUMi.^{33,34,43} In the first set of analyses, SIBTEST was used to compare Hispanics ($n = 153$) to a combined group of non-Hispanics ($n = 202$) who responded to form B. In the second set of analyses, SIBTEST was used to compare blacks ($n = 175$) to a combined group of Hispanic and non-Hispanic whites ($n = 280$) who responded to form B. For both sets of analyses, a 2-step process was undertaken, as is recommended for exploratory DIF analyses.³⁴ In both analyses, DIF

Table 5 Construct Validity of the Numeracy Understanding in Medicine Instrument

Sociodemographic Factor	NUMi Score \bar{x} (s)	Ability Score (θ) \bar{x} (s)
Race/ethnicity ^a		
Non-Hispanic white	16.8 (3.3)	0.95 (0.80)
Non-Hispanic black	10.4 (3.4)	-0.47 (0.64)
Hispanic	11.8 (4.1)	-0.19 (0.82)
Education ^a		
Up to 12 years	9.8 (3.7)	-0.56 (0.72)
Some college	11.3 (4.0)	-0.25 (0.77)
4-year college or more	16.4 (2.9)	0.80 (0.77)

Note: Theta (θ) is the latent trait ability of the respondent as determined by responses to the items and the IRT model. The table demonstrates association of sociodemographic factors and performance on the Numeracy Understanding in Medicine Instrument (NUMi).

a. Race/ethnicity and education levels were significantly associated with health numeracy as assessed by number correct scoring ($P < 0.001$) and θ ($P < 0.001$) using analysis of variance.

was not observed for any of the 20 items using an adjusted P value of 0.05/20 (0.003).

Dimensionality Assessment

Unidimensionality of the latent trait is an underlying assumption of IRT methods. Tests of essential unidimensionality for the 20-item NUMi were conducted using Stout’s test of essential unidimensionality (DIMTEST) in a confirmatory manner.⁴⁴ Data from the 480 respondents for form B were used for the analysis. Two hypotheses were tested: 1) the null hypothesis of unidimensionality between the statistics items and the remaining items (number sense, tables and graphs, and probability) and 2) the null hypothesis of unidimensionality between the probability and statistics items combined compared to the remaining items (number senses and tables and graphs). Neither of these hypotheses yielded significant findings ($t = 0.00, P = 0.50; t = -0.37, P = 0.64$, respectively). Therefore, it was concluded that the items on the 20-item NUMi demonstrated essential unidimensionality.

Model Fit

Log likelihood ratio tests were conducted to compare model fit for the 1- and 2-parameter IRT models for our data.^{29,30} The χ^2 statistic was large, with a rejection of the null hypothesis of no difference between the 1- and 2-parameter model at $P < 0.001$. Therefore, the 2-parameter model results in a statistically significant improvement in fit compared to the 1-parameter model.

Scoring the NUMi

Scoring a measure developed with IRT can be accomplished using IRT software to estimate an examinee's latent trait, θ . However, given that IRT software may not be widely available in practice, another option is to calculate the number correct. The NUMi total score was determined by counting the number of correct items on the 20-item test (potential range of 0 to 20). Items left blank were scored as incorrect. The mean (standard deviation) of scores in the validation sample was 13.2 (4.6) with a range of 2 to 20. The number correct is an examinee's estimate of ability using CTT and is strongly correlated with θ . The correlation between number correct and θ on the NUMi was 0.98 (Table 4). In addition, the correlation of total score with the external validation measures are comparable to those obtained between θ and the external measures (Table 5). Both scoring approaches demonstrate the expected convergent validity, with moderate correlation to existing numeracy measures, and divergent validity, as a lesser degree of correlation is observed with the print literacy measure.

We propose using categories that correspond to cutoff values determined by being more or less than 1 standard deviation from the mean score in our study population. Given a mean score of 13.2 and a standard deviation of 4.6, this scoring approach would be as follows: The category of low numeracy would be defined as a score of 0 to 7; low-average numeracy, 8 to 12; high-average numeracy, 13 to 17; and high numeracy, 18 to 20. The score could also be used as a continuous measure in analyses. Finally, a descriptive presentation of the number correct in each domain may provide the clinician with valuable specific information regarding patient numeracy skills. This score would range from 0 to 5 for each of the following components: number sense, tables and graphs, probability, and statistics.

DISCUSSION

We report on the development and evaluation of a new measure of health numeracy called the NUMi (appendix). This measure is a 20-item paper-and-pencil test that assesses the construct of health numeracy across the areas of number sense, tables and graphs, probability, and statistics. The results of this research indicated that the NUMi has both content and construct validity for use in English-speaking non-Hispanic and Hispanic populations

making it a valuable addition to other existing measures of health numeracy.

There is an emerging consensus in the literature regarding the scope of the health numeracy construct. Health numeracy is generally thought to include a set of skills that range from basic computational skill in arithmetic to the interpretation of table and graph forms of data, to the more conceptual skills required to understand concepts related to probability and statistics.¹⁻⁶ Numeracy is a separate component of health literacy than print literacy.⁴² As with print literacy skills, numerical ability may be related to general cognitive function and intelligence,⁴⁵⁻⁴⁷ a relationship supported by our findings.

Existing measures of health numeracy typically focus on specific components of the health numeracy construct. For example, some measures focus on the application of basic principles of arithmetic, counting, and use of calendar in performing aspects of disease self-management.²⁷ Others focus on aspects of risk communication, including concepts of probability and formats of communicating risk.^{20,21} Still other assessments focus on understanding the results of medical studies as may be communicated by health professionals or through other communication channels²³ or statistical literacy.²⁶ The approach taken in the development of the NUMi was to achieve a comprehensive assessment of skills relevant to the health numeracy framework.

The NUMi demonstrates a moderate level of correlation with existing validated health numeracy instruments, including the Lipkus Expanded Numeracy Scale and the MDIT. This moderate level of correlation suggests both overlap and differences in the skills being assessed. For some purposes, it would be reasonable to use any of these measures for health numeracy. However, depending on one's research or clinical goals, the NUMi may offer advantages to other existing measures that should be considered. In particular, scores obtained from the NUMi will conceptually represent a measure of the content areas of number sense, tables and graphs, probability, and statistics. The total NUMi score thus represents a conceptual measure of this full construct and the preferred measure for some clinical settings. The NUMi may be an appropriate test to use, for example, prior to a cancer treatment consultation or recommendation of use of a decision aid. In both these clinical scenarios, patients may be presented with a range of number-based information, including basic risk and probability information as well as data related to the efficacy of alternative treatment options. The use of the NUMi could indicate the degree to which a patient

has the skills to process and use such information optimally. Information given to a clinician prior to a consultation also has the potential to help the clinician to develop an appropriate communication strategy for the individualized patient during the valuable time they spend together in the consultation.¹⁹

The NUMi was developed to be cross-culturally equivalent across populations. Cultural background may well influence how people think about and use numbers in the context of health.^{48,49} Qualitative work has highlighted important concepts relevant to numeracy among the Mexican American population.³⁵ Cross-cultural methods were used in the qualitative work supporting the theoretical framework as well as steps including item generation, item testing, and item calibration.⁵⁰ This foundation of cross-cultural methods in scale development will support future efforts to translate the NUMi into Spanish and validate its use in Spanish-speaking populations.

The use of IRT methods to develop the NUMi will enable the development of a CAT version of the NUMi that can take advantage of the full bank developed in this work. A CAT uses a computer-generated algorithm to determine which items to administer to respondents based their estimated ability. This approach greatly decreases the response time burden to respondents by using responses to initial items to estimate ability level and identify which remaining items should be administered based on this estimate of ability determined by initial responses. Using a computer-administered modality also offers the advantage of allowing respondents to have items read aloud to them and in their language of choice.^{31,32} A brief CAT of statistical and risk literacy designed for highly educated samples has been developed and demonstrates how ability assessments can be obtained with less respondent burden.⁵¹

The NUMi can be scored using an IRT computer program to determine ability level, θ , or through determination of an examinee's total score on the 20 items. We propose an approach to scoring that categorizes scores into 4 levels: low, low-average, high-average, and high levels of numeracy as determined by the distribution of scores in the study population as detailed above. Future studies are required to correlate performance on the NUMi as measured by both scoring approaches to meaningful outcomes related to informed decision making in the context of medical care.

Our study has some limitations. Using the NUMi to assess numeracy skill may be confounded by levels of reading ability. We used several approaches in the development of the NUMi to minimize confounding

that could occur between print literacy and the performance on the numeracy items. Each respondent was offered the opportunity to have questions read aloud. Furthermore, purposeful sampling was conducted to include data from approximately 5% of respondents who responded to items that were read aloud. Thus, we advise that respondents be given the opportunity to have the items read to them while viewing the graphics and response items. Although this does not exclude the confounding of print and numeric literacy, it offers an approach for those with low reading ability to be assessed for numeric skills. Second, only 8% of study participants demonstrated inadequate or marginal print literacy as measured by the S-TOFHLA, and those with inadequate health literacy were excluded from the validation study. This may raise questions regarding the generalizability of our findings. However, 41% of our participants had only up to a high school-level education. General aptitude, as measured by the Wonderlic, was lower than that of the working US population. Our study population was therefore diverse in not only race and ethnicity but also level of education and cognitive aptitude. In many ways, it is representative of an urban primary care population. Finally, the TIF of the NUMi indicates that the NUMi is most discriminating among respondents that have a lower-than-average level of numeracy. The reasons for this are likely multifactorial. Although efforts were made to develop easy, moderate, and hard items, IRT parameters indicated that many items were easier than originally intended. Furthermore, difficult items were generally found to have poorer discrimination than easier items and were therefore not strong candidates for use in the NUMi. The finding that the NUMi discriminates best at a lower-than-average level of numeracy limits the ability to distinguish skill level at the higher end of numeracy, which might be desirable to identify those who understand more conceptually complex statistical concepts from those that understand only basic statistical concepts. However, it strengthens the ability of the test to identify those at risk due to low numeracy. The future development and addition of difficult items and use of a CAT modality will help to address this limitation. Finally, the NUMi is currently available in English only. However, the cross-cultural methods used and the anticipated translation and development of a CAT modality will support the oral administration of items and the ability to administer the test in English and Spanish versions.

In summary, we developed and validated the NUMi, the first health numeracy test that we are

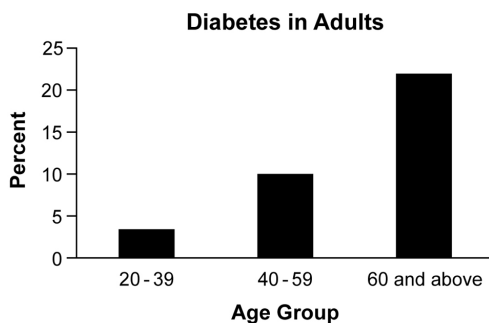
aware of, which was developed using principles of IRT and the full scope of the theoretical definition of health numeracy. The use of IRT offers theoretical and practical advantages in comparison to measures developed using CTT. From a theoretical perspective, IRT measures a latent trait that represents ability relating to a defined set of skills. From a practical point of view, IRT methods support the assessment of item response bias through the use of DIF analyses and the ability to develop a CAT modality that has the potential to reduce respondent burden. Further studies of the use of the NUMi and the relationship of scores to clinically meaningful outcomes will further validate the scoring procedures. We recommend the use of the NUMi for research and clinical settings that seek to assess the overall level of skill across a spectrum of skills reflecting the health numeracy construct.

**APPENDIX
NUMERACY UNDERSTANDING IN MEDICINE
INSTRUMENT**

1. James has diabetes. His goal is to have his blood sugar between 80 and 150 in the morning. Which of the following blood sugar readings is within his goal?
 - a. 55
 - b. 140**
 - c. 165
 - d. 180
2. Nathan has a pain rating of 5 on a pain scale of 1 (no pain) to 10 (worst possible pain). One day later Nathan still has pain but not as much. Now, what pain rating might Nathan give?
 - a. 3**
 - b. 5
 - c. 7
 - d. 9
3. Natasha started taking a new medicine that may cause the side effects listed below. Which side effect is Natasha least likely to have?

a. Dizziness	1 in 5 people
b. Nausea	1 in 10 people
c. Stomach pain	1 in 100 people
d. Allergic reaction	1 in 200 people
4. Frank has a test done to look for blockages in the arteries of his heart. The doctor said that the greater the percent (%) blockage in the artery, the greater the risk of a heart attack. Which percent (%) blockage is most likely to cause a heart attack?
 - a. 33%
 - b. 50%
 - c. 75%
 - d. 98%**
5. The doctor told Maria not to take more than 3 grams (g) of Tylenol a day. Each Tylenol pill is 500 milligrams (mg). What is the greatest number of pills that Maria can take in one day?
 - a. 3 pills
 - b. 6 pills**
 - c. 8 pills
 - d. 12 pills
6. A medical study will randomly assign people so they are equally likely to get medicine A or medicine B. If there are 300 people in the study, about how many are expected to get medicine A?
 - a. 100 people
 - b. 150 people**
 - c. 200 people
 - d. 250 people
7. Older age and smoking both increase the risk of a heart attack over time. David is now 50 years old and smokes. His risk of a heart attack in the next 10 years is 10%. If he continues to smoke which of the following could be his risk of a heart attack over the next 20 years?
 - a. 5%
 - b. 10%
 - c. 30%**
 - d. 100%
8. James starts a new blood pressure medicine. The chance of a serious side effect is 0.5%. If 1000 people take this medicine, about how many would be expected to have a serious side effect?
 - a. 1 person
 - b. 5 people**
 - c. 50 people
 - d. 500 people
9. The PSA (Prostate Specific Antigen) is a blood test that can be used to screen for prostate cancer. However, 30% of men who have an abnormal test result will turn out not to have cancer. John has an abnormal test result. What is the chance that John has prostate cancer?
 - a. 0%
 - b. 30%
 - c. 70%**
 - d. 100%
10. Rebecca is treated for stage 2 breast cancer. The chance that the cancer will come back is 10% over 10 years. If Rebecca takes a new medicine, this chance will decrease by 30%. If 100 women like Rebecca take this medicine, how many are now expected to have breast cancer come back within 10 years?
 - a. 3 out of 100 women
 - b. 7 out of 100 women**
 - c. 10 out of 100 women
 - d. 30 out of 100 women

11. A study found that chemotherapy decreased the risk of dying from colon cancer by about 30%. The study was 95% sure that the actual benefit was between 10% and 50%. Which of the following is not in the expected range of benefit?
- 11% decrease in risk
 - 30% decrease in risk
 - 45% decrease in risk
 - 95% decrease in risk**
12. A study in arthritis patients found that Medicine A decreased arthritis pain 10% more often than Medicine B. The difference was not statistically significant. Which of the following best describes these results?
- Medicine A and Medicine B work equally well**
 - Medicine A is proven to be better than Medicine B
 - Medicine B is proven to be better than Medicine A
13. A study found that a new diabetes medicine controlled blood sugar 8% more often than the old medicine. This difference was statistically significant ($p < 0.05$). The probability that this finding is due to chance alone is less than:
- 1 in 5
 - 1 in 10
 - 1 in 15
 - 1 in 20**
14. In general, the results of randomized controlled trial will be more reliable if a larger number of people are in the study.
- True**
 - False
15. A study was done of health habits. A group of people took a survey every few years for 20 years. The study found that people who exercised 3 times a week or more lived an average of 2 years longer than those who did not. What does this study show?
- Exercise was the cause of living a longer life
 - There is a relationship between exercise and living a longer life**
16. According to the graph below, what percent (%) of adults in the 40-59 year old age group have diabetes?
- 5%
 - 10%**
 - 15%
 - 20%



17. John has a fever. The doctor tells him to come to the hospital if his temperature is above 102.5 °F. Otherwise, John should take Tylenol and rest. John's temperature is shown in the picture below. What should John do?
- Take Tylenol and rest**
 - Go to the hospital

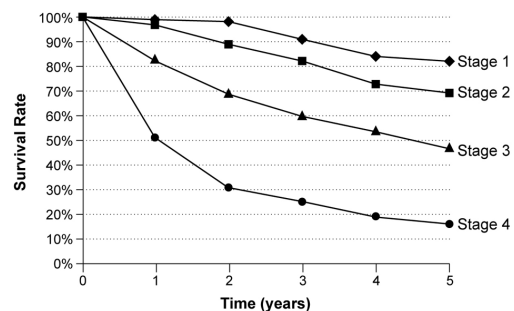


18. Mary has 2 cups of food whose nutrition label is below. How many calories are in the 2 cups of food?
- 140 calories
 - 280 calories
 - 560 calories**
 - 680 calories

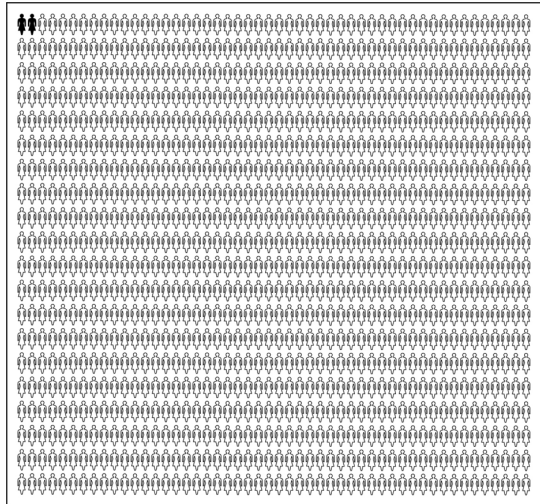
Nutrition Facts	
Serving Size 1 cup (228g)	
Servings per Container 2	
Amount Per Serving	
Calories 280	Calories from Fat 120
% Daily Value*	
Total Fat 13g	20%
Saturated Fat 5g	25%
Trans Fat 2g	
Cholesterol 2mg	10%
Sodium 660 mg	28%
Total Carbohydrate 31g	10%
Dietary Fiber 3g	
Sugars 5g	
Protein 5g	
Vitamin A 4%	Vitamin C 2%
Calcium 15%	Iron 4%

*Percent Daily Values are based on a 2,000-calorie diet. Your Daily values may be higher or lower depending on your calorie needs



19. Andrea has stage 2 breast cancer. According to the graph below, what is Andrea's chance of surviving 3 years after her diagnosis?
- 56%
 - 82%**
 - 92%
 - 100%



20. Carol is taking a new medicine. The risk of a side effect is very small. According to the picture below, what is her risk of having a side effect?
- a. 0.0002
 - b. 0.002**
 - c. 0.02
 - d. 0.20



Risk per 1000 women

 Side Effect  No Side Effect

Note to Appendix

The correct responses are in bold. The Numeracy Understanding in Medicine Instrument can be scored by determining the number correct out of 20. The percentage correct can provide a continuous measure of health numeracy ability, with higher numbers indicating a higher level of numeracy. A categorical scoring approach is presented below.

Level of Health Numeracy	Score
Low	0–7
Low-average	8–12
High-average	13–17
High	18–20

REFERENCES

1. Golbeck AL, Ahlers-Schmidt CR, Paaschal AM, Dismuke SE. A definition and operational framework for health numeracy. *Am J Prev Med.* 2005;29:375–6.

2. Anker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. *J Am Med Inform Assoc.* 2007;14:713–21.

3. Lipkus IM, Peters E. Understanding the role of numeracy in health: proposed theoretical framework and practical insights. *Health Educ Behav.* 2009;36:1065–81.

4. Schapira MM, Fletcher KE, Gilligan MA, et al. A framework for health numeracy: how patients use quantitative skills in health care. *J Health Commun.* 2008;13:501–17.

5. Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E. Clinical implications of numeracy: theory and practice. *Ann Behav Med.* 2008;35:261–74.

6. Apter AJ, Paasche-Orlow M, Remillard JT, et al. Numeracy and communication with patients: they are counting on us. *J Gen Intern Med.* 2008;23:2117–24.

7. Aggarwal A, Speckman JL, Paasche-Orlow MK, Roloff KS, Battaglia TA. The role of numeracy on cancer screening among urban women. *Am J Health Behav.* 2007;31:S57–68.

8. Schapira MM, Neuner J, Fletcher KE, Gilligan MA, Hayes E, Laud P. The relationship of health numeracy to cancer screening. *J Canc Educ.* 2011;26:103–10.

9. Janz NK, Champion VL, Strecher VJ. The health belief model. In: Glanz K, Rimer BK, Lewis FM, eds. *Health Behavior and Health Education: Theory, Research, and Practice.* 3rd ed. San Francisco: Jossey-Bass; 2002. p 45–66.

10. Hershey JC, Baron J. Clinical reasoning and cognitive processes. *Med Decis Making.* 1987;7:203–11.

11. Weinstein ND. Testing four competing theories of health-protective behavior. *Heath Psychol.* 1993;12:323–33.

12. Osborn CY, Cavanaugh K, Wallson KA, Rothman RL. Self-efficacy links health literacy and numeracy to glycemic control. *J Health Commun.* 2010;15:146–58.

13. Cavanaugh K, Huizinga M, Wallston KA, et al. Association of numeracy and diabetes control. *Ann Intern Med.* 2008;148:737–46.

14. Estrada CA, Martin-Hryniewicz M, Peek BT, Collins C, Byrd JC. Literacy and numeracy skills and anticoagulation control. *Am J Med Sci.* 2004;328:88–93.

15. Apter AJ, Cheng J, Small D, et al. Asthma numeracy skill and health literacy. *Asthma.* 2006;43:705–10.

16. Zikmund-Fisher BJ, Smith DM, Ubel PA, Fagerlin A. Validation of the subjective numeracy scale (SNS): effects of low numeracy on comprehension of risk communications and utility elicitation. *Med Dec Making.* 2007;27:663–71.

17. Schwartz SR, McDowell J, Yueh B. Numeracy and the shortcomings of utility assessments in head and neck cancer patients. *Head Neck.* 2004;26:401–7.

18. Woloshin S, Schwartz LM, Moncur M, Gabriel S, Tosteson ANA. Assessing values for health: numeracy matters. *Med Dec Making.* 2001;21:382–90.

19. Hamm RM, Bard DE, Hsieh E, Stein HF. Contingent or universal approaches to patient deficiencies in health numeracy. *Med Decis Making.* 2007;27:635–7.

20. Schwartz L, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med.* 1997;127:966–72.

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

22. Fagerlin A, Zikmund-Fisher BL, Ubel PA, et al. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Med Decis Making*. 2007;27:672–80.
23. Schwartz LM, Woloshin S, Welch HG. Can patients interpret health information? An assessment of the Medical Data Interpretation Test. *Med Decis Making*. 2005;25:290–300.
24. Weiss BD, Mays MZ, Castro KM, et al. Quick assessment of literacy in primary care: the newest vital sign. *Ann Fam Med*. 2005;3:514–22.
25. Huizinga MM, Elasy TA, Wallston KA, et al. Development and validation of the diabetes numeracy test (DNT). *BMC Health Serv Res*. 2008;8:96.
26. Woloshin S, Schwartz LM, Welch HG. Patients and medical statistics: interest, confidence, and ability. *J Gen Intern Med*. 2005;20:996–1000.
27. Parker RM, Baker DW, Williams MV, Nurss JR. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. *J Gen Intern Med*. 1995;10:537–41.
28. DeVellis RF. *Scale Development: Theory and Applications*. 2nd ed. Thousand Oaks (CA): Sage Publications; 2003.
29. Hambleton RK, Swaminathan H, Rogers HG. *Fundamentals of Item Response Theory*. Newbury Park (CA): Sage Publications; 1991.
30. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Norwell (MA): Kluwer Academic Publishers; 1985.
31. Gershon RC. Computer adaptive testing. *J Appl Meas*. 2005;6:109–27.
32. Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. *Meas Eval Counsel Dev*. 2004;37:70–84.
33. Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*. 1993;58:159–94.
34. Gierl MJ. Using dimensionality-based DIF Analyses to identify and interpret constructs that elicit group differences. *Educ Meas Issues Pract*. 2005;24:3–14.
35. Schapira MM, Fletcher KE, Ganschow PS, et al. The meaning of numbers in health: exploring health numeracy in a Mexican-American population. *J Gen Intern Med*. 2011;26:705–11.
36. Beatty PC, Willis JB. Research synthesis: the practice of cognitive interviewing. *Public Opin Q*. 2007;71:287–311.
37. Baker DW, Williams MV, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. *Patient Educ Couns*. 1999;38:33–42.
38. Matthews TD, Lassiter KS. What does the Wonderlic Personnel Test measure? *Psychol Rep*. 2007;100:707–12.
39. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–81.
40. Zimowski MF, Muraki E, Mislevy R, Bock RD. *BILOG-MG*. Chicago: Scientific Software International; 1996.
41. Jastak S, Wilkinson GS. *Wide-Range Achievement Test—Revised 3*. Wilmington (DE): Jastak Associates; 1993.
42. Institute of Medicine. *Health Literacy: A Prescription to End Confusion*. Washington, DC: National Academies Press; 2004.
43. Roussos LA, Stout WF. Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *J Educ Meas*. 1996;33:215–30.
44. Stout WF. A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika*. 1987;52:589–617.
45. Barnes DE, Tager IR, Satariano WA, Yaffe K. The relationship between literacy and cognition in well-educated elders. *J Gerontol*. 2004;4:390–95.
46. Federman AD, Sana M, Wolf S, Siu AL, Halm EA. Health literacy and cognitive performance in older adults. *J Am Geriatr Soc*. 2009;57:1475–80.
47. Abdel-Kader K, Dew MA, Bhatnagar M, et al. Numeracy skills in CKD: correlates and outcomes. *Clin J Am Soc Nephrol*. 2010;5:1566–73.
48. Kreuter MW, McClure SM. The role of culture in health communication. *Annu Rev Public Health*. 2004;25:439–55.
49. Wright GN, Phillips LD. Cultural variation in probabilistic thinking: alternative ways of dealing with uncertainty. *Int J Psychol*. 1980;15:239–57.
50. Warnecke RB, Johnson TP, Chavez N, et al. Improving question wording in surveys of culturally diverse populations. *Ann Epidemiol*. 1977;7:334–42.
51. Cokely ET, Galesic M, Schulz E, Garcia-Retamero R. Measuring risk literacy: the Berlin Numeracy Test. *Judgm Decis Mak*. 2012;7:25–47.