Published in final edited form as:
Med Decis Making. 2012 ; 32(6): 851–865. doi:10.1177/0272989X12447239.

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

The Numeracy Understanding in Medicine Instrument

Marilyn M. Schapira, MD, MPH,
Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Cindy M. Walker, Ph.D.,
Department of Educational Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI

Kevin J. Cappaert, MA,
Department of Educational Psychology, University of Wisconsin Milwaukee

Pamela S. Ganschow, MD,
Department of Medicine, John H. Stroger Jr Hospital of Cook County and Rush University Medical Center, Chicago, IL

Kathlyn E. Fletcher, MD, MA,
Department of Medicine, Clement J. Zablocki VA Medical Center and Medical College of Wisconsin, Milwaukee, WI

Emily L. McGinley, MS, MPH,
Center for Patient Care and Outcomes Research, Medical College of Wisconsin, Milwaukee, WI

Sam Del Pozo, MA,
Department of Medicine, John H. Stroger Jr Hospital of Cook County and Rush University Medical Center, Chicago, IL

Carrie Schauer, BSW,
Center for Patient Care and Outcomes Research, Medical College of Wisconsin, Milwaukee, WI

Sergey Tarima, Ph.D., and
Institute of Health and Society, Medical College of Wisconsin, Milwaukee, WI

Elizabeth A. Jacobs, MD, MAPP
Department of Medicine, University of Wisconsin, Madison, WI

Abstract

Background—Health numeracy can be defined as the ability to understand medical information presented with numbers, tables and graphs, probability, and statistics and to use that information to

Marilyn M. Schapira, MD, MPH, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA, 19104, 215-898-2022, mschap@upenn.edu.

communicate with one's health care provider, 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 (IRT) 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% Non-Hispanic white, and 42% Non-Hispanic black. Forty-one percent (41%) had no more than a high school education. The mean score on the NUMi was 13.2 (SD 4.6) with a Cronbach's alpha of 0.86. Difficulty and discrimination IRT 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 WRAT-arithmetic test (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.

INTRODUCTION

Health numeracy can be defined as the ability to understand medical information presented with numbers, tables and graphs, probability, and statistics and to apply numerical information for the purpose of communicating with health care providers, taking care of one's health, and participating 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 numeracy (6–8). Further, 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 aspects of the full construct of health numeracy, be it number sense, risk communication and probability, or the interpretation of medical study results. For some purposes, a measure that reflects the full spectrum of health numeracy skills may be optimal. Further, existing measures have not been developed using cross-cultural approaches, making it unclear if these measures are valid for certain groups in the population that we serve, such as Hispanics. Finally, existing measures have not been developed using Item Response Theory (IRT). 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 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) that 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 Computer Adaptive Test 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 upon 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 medical information presented with numbers, tables and graphs, probability, and statistics and to use that information to communicate with your health provider, to take care of your health, and to 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 four domains of numeric skills that are widely applied in health; 1) number sense, 2) tables & graphs, 3) probability, and 4) 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 (TST) was developed to represent the set of skills comprising the health numeracy construct (Table 1) and the health care context in which the skills are applied. An expert panel was convened to review the health numeracy framework, the TST, and the initial items generated. The expert panel consisted of five members: one clinician who was bilingual and practiced in the Chicago community, two clinician-investigators (one with research expertise in the area patient physician communication and one 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. 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. The interviews 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). 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 comprising the item bank (n=110) were divided into two parallel forms (A and B) with 64 items per form to reduce the respondent burden (Figure 1). Each form contained 18 common linking items 46 unique items. Unique items were selected by ensuring that the content domain and perceived level of difficulty was similar on both forms to create two 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 two 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, recruitment booths in community centers, and recruitment booths in clinical settings. Recruitment booths were staffed by bilingual study personnel. Inclusion criteria included age 21 years or older and 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 socio-demographic information and the print literacy version of the Short Test of Functional Health Literacy in Adults (S-TOFHLA), and the Wonderlic Aptitude Test (Wonderlic). The S-TOFHLA is a 17 items test that has a potential score of 0–36 and classifies respondents as 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 a respondent 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 an item blank rather than to 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 Cognitive Aptitude test.

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 for analysis (39). The IRT software BILOG was used to calibrate items using a unidimensional dichotomous 2-parameter IRT model (40). The 2-parameter IRT model 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 ability (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 the Equation below where θ represents that level of ability of the respondent, alpha represents the discrimination of the item, and beta represents the difficulty of the item.

$$P(\theta)=1/1+\exp[-1.7a(\theta-b)]$$

Estimated A Priori (EAP) latent trait scoring was used to obtain parameters. In addition to the IRT parameters, classical test theory (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) as well as Cronbach's alpha for the total scale were calculated. 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 than form A. Choosing all items from one form was necessary in order to use original response data to obtain a total score in the validation analyses. Difficulty was assessed with the IRT beta parameter that typically ranges from −3.0 to 3.0. Discrimination was assessed with the IRT alpha parameter that typically ranges from zero to 3. 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 test includes 5 items from each of the four content areas: number sense, tables & 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 comprised 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: 1) the Wide Range Achievement Test-Arithmetic (WRAT-A) consisting of 40 math problems (41) the Lipkus Expanded Numeracy Scale consisting of 11 items (Lipkus) (21), 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. Further, 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 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). Subjects 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 analysis were conducted to evaluate for differential item functioning across groups, evaluate for unidimensionality of the measure, and to test whether the model fit was improved using the 2 Parameter compared to 1 Parameter IRT models. Standard IRT methods were used for these additional analyses with details presented with the results in sections below (29, 30, 33, 34, 43, 44).

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.

RESULTS

Study Population

One thousand (n=1000) subjects 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 (29%) were Hispanic. Sixty percent (60%) were female. Eight percent (8%) had inadequate or marginal health literacy as measured by the S-TOFHLA. Forty-one percent (41%) had no more than a high school level education. The Wonderlic Aptitude Test score (with a potential range of 0 to 50) had a lower mean score for the study population (Mean 17.5, SD 8.7) than the published norms of working adults in the United States of (Mean 21.7, SD 7.6, $p<0.01$) (38). The items were read aloud to 46 (4.6%) of the 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. Item #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 #7 (probability domain, understanding the relationship of short and long term risk of mortality) and #10 as described above, with IRT discrimination parameters of .42 and .39, respectively.

The Test Information Function (TIF) is a function of the item parameters and provides a summary of information provided by the full test (29, 30). 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).

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 4a). The construct validity of the NUMi also is also supported by the association observed between socio-economic characteristics and performance of on the NUMi with increasing levels of education associated with greater ability on the NUMi (Table 4b).

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 two items in the statistics sub-domain were noted to have some redundancy in content. Therefore, a replacement item was identified based upon content and discrimination and difficulty parameters. Minor modifications were also made to the wording of some items on the final version based upon feedback from the expert panel. The substitution did not impact the shape of the Test Information Function.

Differential Item Functioning Analysis

Two sets of exploratory DIF analyses were conducted using SIBTEST (Simultaneous Item Bias), a non-parametric statistical procedure, to test for item bias on the twenty 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=302) 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 two-step process was undertaken as is recommended in exploratory DIF analyses. In both analyses, DIF 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 & 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 & graphs). Neither of these hypotheses yielded significant findings (t=0.00, p=0.500; and 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-parameter and 2-parameter IRT models for our data (29, 30). The chi-square statistic was large with a rejection of the null hypothesis of no difference between the 1 parameter 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 done using IRT software to estimate an examinee's latent trait, θ. However, as this may not be feasible in practice, another option is to calculate the number correct. As total score is an examinee's estimate of ability using CTT, it is strongly correlated with θ. As Table 4a illustrates, the correlation between total score and θ on the NUMi is 0.98. The construct validity data (Table 4b) indicates that the correlations of the NUMi total correct with the external validation measures are comparable to those obtained when correlating the external measures with θ. Both measures 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. 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 (SD) of scores in the validation sample was 13.2 (SD 4.6) with a range of 2 to 20.

We propose using categories of the number correct to score that correspond to cut-off values determined by being more or less than 1 SD from the mean score in our study population. Given a mean score of 13.2 and a SD 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 would be defined as a score of 8 to 12, high average numeracy by a score of 13 to 17, and high numeracy by a score of 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 & graphs, probability, and statistics.

DISCUSSION

We report on the development and evaluation of a new measure of health numeracy called the Numeracy Understanding in Medicine Instrument (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 & 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 given 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 of the results of medical studies as may be communicated by health professionals or through other communication channels (23) 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 Medical Data Interpretation Test. 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 & 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 of 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 relating 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 a 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). The 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 future development of a Computer Adaptive Test (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 of the 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 computer adaptive test of statistical and risk literacy designed for highly educated samples has been developed and demonstrates how ability assessments can be obtained with limited 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 four 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 these scores 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. Further, 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, forty-one percent (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 U.S. population. Our study population, therefore, was not only diverse in race and ethnicity but in the level of education and cognitive aptitude. In many ways, they are representative of an urban primary care population. Finally, the Test Information Function 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. Further, difficult items were generally found to have poorer discrimination than easier items and therefore were 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 distinguish between 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 Computer Adaptive Test modality will help to address this limitation. Finally, the NUMi is currently only available in English. However, the cross-cultural methods used and the anticipated translation and development of a Computerized Adaptive Test modality in the future 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 Numeracy Understanding in Medicine Instrument, the first health numeracy test that we are aware of that was developed using principles of item response theory 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 classical test theory. 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 differential item functioning analyses and the ability to develop a computer adaptive test modality (CAT) 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.

Acknowledgments

This study was funded by the National Cancer Institute of the NIH, #NCIR01CA115954. The work was presented at the 2010 Annual Society for Medical Decision Making meeting, Toronto, CA. The REDCap database use in this work was supported by a Clinical and Translational Science Institute grant 1UL1-RR031973(-01)

References

1. Golbeck AL, Ahlers-Schmidt CR, Paaschal AM, Dismuke SE. A definition and operational framework for health numeracy. Am J Prev Med. 2005; 29:375–376. [PubMed: 16242604]
2. Anker JS, Kaufman D. Rethinking health numeracy: A multidisciplinary literature review. J Am Med Inform Assoc. 2007; 14:713–721. [PubMed: 17712082]
3. Lipkus IM, Peters E. Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. Health Education & Behavior. 2009; 36:1065–1081. [PubMed: 19834054]
4. Schapira MM, Fletcher KE, Gilligan MA, et al. A framework for health numeracy: How patients use quantitative skills in health care. Journal of Health Communication. 2008; 13:501–517. [PubMed: 18661390]
5. Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E. Clinical implications of numeracy: Theory and practice. Ann Behav Med. 2008; 35:261–274. [PubMed: 18677452]
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. [PubMed: 18830764]
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–S68. [PubMed: 17931137]
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; 26:103–110.
9. Janz, NK.; Champion, VL.; Strecher, VJ. The health belief model. In: Glanz, K.; Rimer, BK.; Lewis, FM., editors. Health behavior and health education: theory, research, and practice. 3. Jossey-Bass; San Francisco: p. 45-66.

10. Hershey JC, Baron J. Clinical reasoning and cognitive processes. Med Decis Making. 1987; 7:203–11. [PubMed: 3683107]
11. Weinstein ND. Testing four competing theories of health-protective behavior. Heath Psychology. 1993; 12:323–333.
12. Osborn CY, Cavanaugh K, Wallson KA, Rothman RL. Self-efficacy links health literacy and numeracy to glycemic control. J of Health Communication. 2010; 15:146–158.
13. Cavanaugh K, Huizinga M, Wallston KA, et al. Association of numeracy and diabetes control. Ann Intern Med. 2008; 148:737–746. [PubMed: 18490687]
14. Estrada CA, Martin-Hryniewicz M, et al. Literacy and numeracy skills and anticoagulation control. Am J Med Sci. 2004; 328:88–93. [PubMed: 15311167]
15. Apter AJ, Cheng J, Small D, et al. Asthma numeracy skill and health literacy. Asthma. 2006; 43:705–710.
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; 27:663–671.
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. [PubMed: 15122656]
18. Woloshin S, Schwartz LM, Moncur M, Gabrial S, Tosteson ANA. Assessing values for health: Numeracy matters. Med Dec Making. 2001; 21:382–390.
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. [PubMed: 17921452]
20. Schwartz L, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med. 1997; 127:966–972. [PubMed: 9412301]
21. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001; 21:37–44. [PubMed: 11206945]
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; 27:672–680. [PubMed: 17641137]
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. [PubMed: 15951456]
24. Weiss BD, Mays MZ, Castro KM, et al. Quick assessment of literacy in primary care: The newest vital sign. Ann Fam Med. 2005; 3:514–522. [PubMed: 16338915]
25. Huizinga MM, Elasy TA, Wallston KA, et al. Development and validation of the diabetes numeracy test (DNT). BMC Health Services Research. 2008; 8:96. [PubMed: 18452617]
26. Woloshin S, Schwartz LM, Welch HG. Patients and Medical Statistics: Interest, Confidence, and Ability. J Gen Intern Med. 2005; 20:996–1000. [PubMed: 16307623]
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; 10:537–541. [PubMed: 8576769]
28. DeVellis, RF. Applied Social Research Methods Series. 2. Sage Publications; Scale Development: Theory and Applications; p. 26
29. Hambleton, RK.; Swaminathan, H.; Rogers, HG. Fundamentals of Item Response Theory. Sage Publications; Newbury Park, CA: 1991.
30. Hambleton, RK.; Swaminathan, H. Item Response Theory: Principles and Applications. Kluwer Academic Publishers; Norwell, MA: 1985.
31. Gershon RC. Computer adaptive testing. Journal of applied measurement. 2005; 6:109–127. [PubMed: 15701948]
32. Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development. 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. Educational Measurement: Issues and Practice. 2005; 24(1):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. Feb.2011 Epub ahead of print.

36. Beatty PC, Willis JB. Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly. 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. [PubMed: 14528569]

38. Matthews TD, Lassiter KS. What does the Wonderlic Personnel Test measure? Psychological Reports. 2007; 100:707–12. [PubMed: 17688083]

39. Harris, Paul A.; Taylor, Robert; Thielke, Robert; Payne, Jonathon; Gonzalez, Nathaniel; Conde, Jose G. Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr; 42(2):377–81. [PubMed: 18929686]

40. Zimowski, MF.; Muraki, E.; Mislevy, R.; Bock, RD. BILOG - MG. Scientific Software International; Chicago, IL: 1996.

41. Jastak, S.; Wilkinson, GS. Wide-Range Achievement Test-Revised. Vol. 3. Wilmington, Del: 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. Journal of Educational Measurement. 1996; 33:215–230.

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. Journal of Gerontology. 2004; 4:390–395.

46. Federman AD, Sana M, Wolf S, Siu AL, Halm EA. Health literacy and cognitive performance in older adults. J Am Geriatr Soc. 2009; 57:1475–1480. [PubMed: 19515101]

47. Abdel-Kader K, Dew MA, Bhatnagar M, et al. Numeracy skills in CKD: Correlates and outcomes. Clin J Am Soc Nephrol. 2010; 5:1566–1573. [PubMed: 20507954]

48. Kreuter MW, McClure SM. The role of culture in health communication. Annu Rev public Health. 2004; 25:439–55. [PubMed: 15015929]

49. Wright GN, Phillips LD. Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. International Journal of Psychology. 1980; 15:239–257.

50. Warnecke RB, Johnson TP, Chavez N, et al. Improving question wording in surveys of culturally diverse populations. Ann Epidemiol. 1977; 7:334–342. [PubMed: 9250628]

51. Cokely ET, Galesic M, Schulz E, Garcia-Retamero R. The Berlin Advanced Numeracy Test for highly educated samples (ANT-E): A three minute test of statistical and risk literacy. Judgment and Decision Making. (in press).

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 it is better. Now, what pain rating might Nathan give?

    a. **3**

    b. 5

    c. 7

    d. 9

3. Natasha started a new medicine and was given a handout showing the chance that side effects will occur as in the table below. Which side effect is Natasha least likely to get?

Side Effect	Chance of Occurring
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 to look for blockages in the arteries of his heart. The doctor said that a person with a higher percent (%) blockage has a high chance of having a heart attack. Which percent (%) blockage has the highest chance of 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 highest 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 that people 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. David is 50 years old and smokes cigarettes. His doctor tells him that the chance of having a heart attack increases as people age and if they smoke. His current chance of a heart attack is 10% over the next 10 years. Which of the following is the best guess of David's chance of a heart attack in 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 looks for prostate cancer. The test has false alarms so about 30% of men who have an abnormal test turn out not to have prostate cancer. John had an abnormal test. What is the chance that John has prostate cancer?

a. 0%

b. 30%

c. **70%**

d. 100%

10. Rebecca was treated for stage 2 breast cancer. The chance that the breast cancer will come back is 10% over the next 10 years. If Rebecca takes a new medicine, this chance will decrease by about 30%. Out of 100 women like Rebecca who take the medicine, how many will 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 real benefit was between 10% and 50%. Which of the following is not in the expected range of benefit?

a. 11% decrease in risk

b. 30% decrease in risk

c. 45% decrease in risk

d. **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?

a. **Medicine A and medicine B work equally well**

b. Medicine A is proven to be better than medicine B

c. Medicine B is proven to be better than medicine A

13. A study found that a new diabetes medicine led to control of blood sugar in 8% more patients than the old medicine. This difference was statistically significant (p=0.05). The likelihood that this finding was due to chance alone is:

a. 1 in 5

b. 1 in 10

c. 1 in 15

d. **1 in 20**

14. In general, the results of a randomized controlled trial will be more reliable if a larger number of people are in the study.

a. **True**

b. False

15. A survey asked a group of people about their exercise habits and followed them; over time. The study found that those who exercised 3 times a week or more lived an average of 2 years longer than those who did not What did this study show?

a. Exercising causes people to live longer

b. **There is a relationship between exercising and living longer**

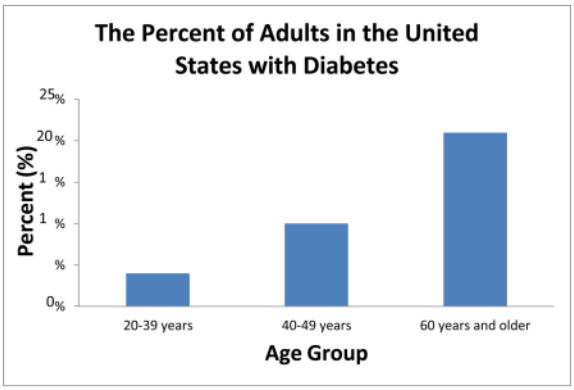
16. According to the graph below, what percent (%) of adults in the 40–49 year old age group have diabetes?

a. 5%

b. **10%**

c. 15%

d. 20%



17. John had a fever. The doctor told him to come to the hospital if his temperature was above 102.5 F. Otherwise, John should take Tylenol and rest. If John's temperature is as shown in the picture below, what should John do?

a. **Take Tylenol and rest**

b. Go to the hospital

18. A nutrition label is shown below. How many calories did Mary eat if she had 2 cups of food?

a. 140 calories

b. 280 calories

c. **560 calories**

d. 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.

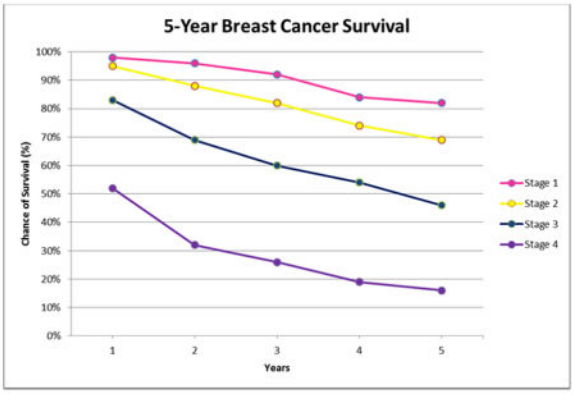
19. The graph below shows the outcomes of a group of women diagnosed with breast cancer. Andrea has stage 2 breast cancer. According to the graph, what is her chance of surviving 3 years after diagnosis?

a. 56%

b. **82%**

c. 92%

d. 100%



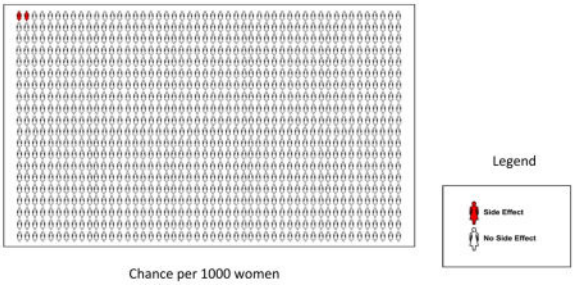
20. Carol is taking a new medicine. The chance of a side effect is very small as shown in the graph below. What number best shows her chance of having a side effect?

a. 0.0002

b. **0.002**

c. 0.02

d. 0.20



Note to Appendix

The correct responses are in bold. The NUMi can be scored by determining the number correct out of 20. The percent correct can provide a continuous measure of health numeracy ability with higher numbers indicating a higher level of numeracy. The following categorical scoring can also be used:

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

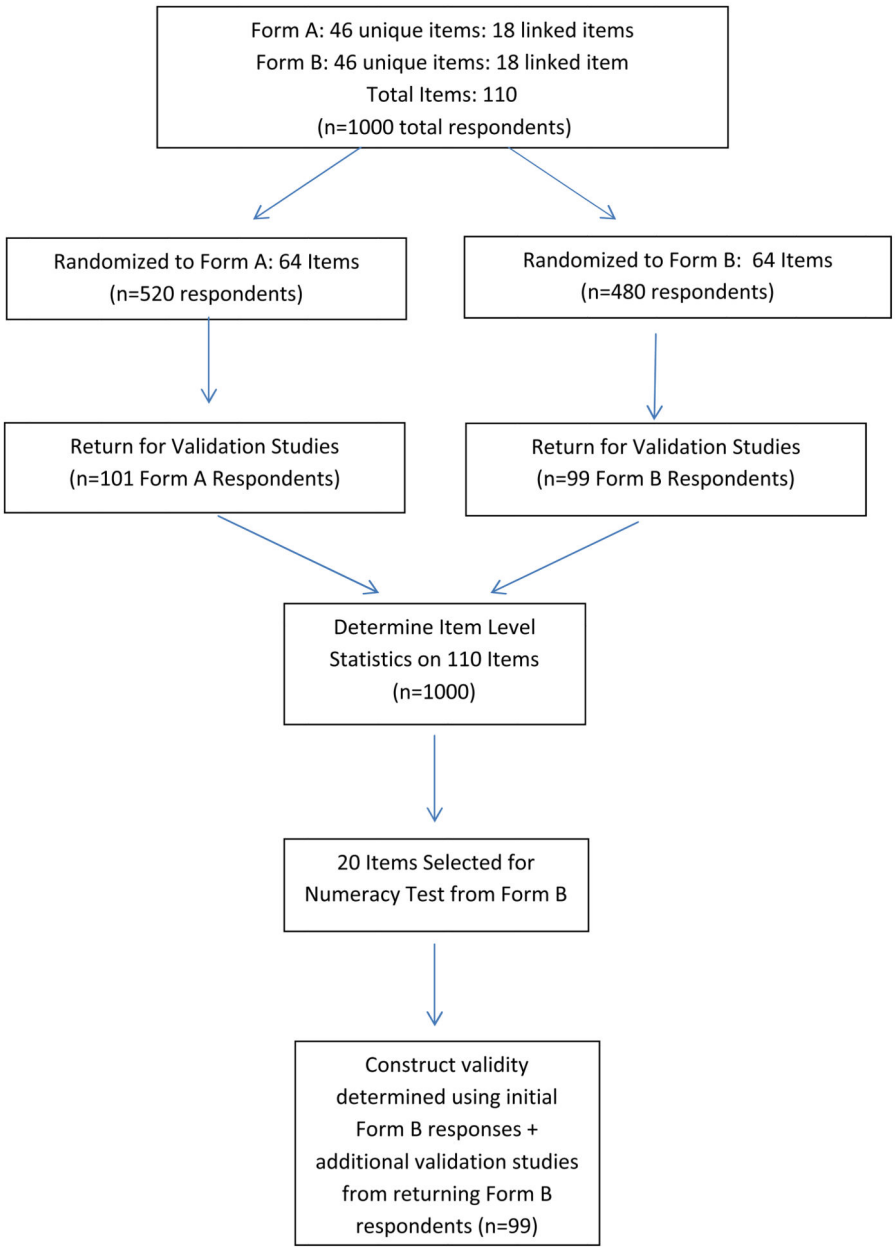


Figure 1.
Flow Chart of Study Population Used to Obtain Psychometric Data

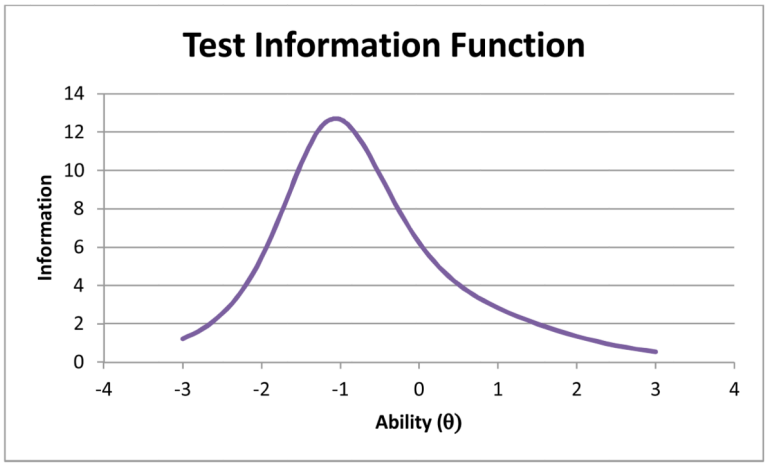


Figure 2.

Test Information Function of the Numeracy Understanding in Medicine Instrument. The x-axis represents the degree of ability level and 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 1

Definition of Numeracy Construct and Skills outlined in the Test Specification Table

Construct	Definition
Health Numeracy	The ability to understand medical information presented with numbers, tables and graphs, probability, and statistics and to use that information to communicate with your health care provider, take care of your health, and participate in medical decisions.
Number Sense	The ability to represent, order, compute and estimate numbers; to understand how fractions and decimals relate to each other and how each can best be used to describe a particular health related situation. • 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 numbers 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 & Graphs	The ability to read and use one and two-dimensional graphic forms such as tables, charts, and graphs; and to apply these skills in a health situation. • Use a table to identify the appropriate information given two 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; to compute the probability of an event occurring. • Convert a risk from a probability statement to a frequency statement • Understand the probability of two independent events • Understand the probability of two non-independent events • Understand the importance of pre-test 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; to interpret such data to make informed decisions about health. • 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; Concept that a reported finding may be due to chance alone • Understand the concept of uncertainty in estimates; Uncertainty as part of the scientific process, general understanding of confidence intervals

These items define the overall construct of health numeracy and the 4 topics that comprise the health numeracy framework used to develop the Numeracy Understanding in Medicine Instrument.

Table 2

Characteristics of the Study Population

NUMi Demographics	Total (n=1000)	Form A (n=520)	Form B (n=480)	Validation Sample (n=99)
	n (%)	n (%)	n (%)	n (%)
GENDER				
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				
< 45 years	536 (53)	277 (53.3)	259 (54)	51 (51.5)
45–59 years	329 (33)	178 (34.2)	151 (31.4)	36 (36.4)
60–74 years	119 (12)	58 (11.2)	61 (12.7)	11 (11.1)
≥ 75 years	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)
TOFHLA Score				
0–16 (Inadequate functional health literacy)	43 (4)	19 (3.7)	24 (5)	2 (2.0)
17–22 (Marginal functional health literacy)	43 (4)	25 (4.8)	18 (3.8)	2 (2.0)
23–36 (Adequate functional health 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)
Four 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 Aptitude 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 (--)

NUMi Demographics	Total (n=1000)	Form A (n=520)	Form B (n=480)	Validation Sample (n=99)
	n (%)	n (%)	n (%)	n (%)
Item Administration				
Read aloud to respondent	46 (5)	16 (3.1)	30 (6.3)	98 (99.0)
Self-administered by respondent	954 (95)	504 (96.9)	450 (93.7)	1 (1.0)

Table 3

Item Level Analysis of the Numeracy Understanding in Medicine Instrument (NUMi) Questions

		Classical Test Theory Item Statistics		Item Response Theory Item Statistics	
		Difficulty 0–1	**Discrimination 0–1**	**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
4*	Frequency format/Side effects*	.76	.51	−.85	.87
3	Ordering numbers/Test results	.76	.42	−.71	1.32
5	Measurement/Dosing medication	.52	.40	−.18	.73
Probability					
6	Randomization/Study participation	.86	.39	−1.27	1.27
7≠*	Class inclusion judgments/Life expectancy	.57	.33	−.59	.42
8	Small risks/Side effects	.49	.53	.20	.62
9*	Calculating probability/Screening tests	.46	.47	.002	.81
10	Relative risk reduction/Cancer recurrence	.30	.33	1.45	.39
Statistics					
11	Uncertainty/95% CI of treatment efficacy	.64	.68	−.84	1.40
12*	Statistical significance/Treatment efficacy	.58	.52	.06	.63
13*	P-value/Interpretation of study results	.27	.51	1.19	.85
14	Sample size implications/Interpretation of study results	.77	.33	−1.70	.53
15	Causation vs. association/Interpretation of study results	.68	.50	−.71	.57
Tables & 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≠*	Reading a table/Interpretation a nutrition label	.82	.29	−1.31	.73
19	Interpreting survival curve/Survival estimates	.77	.55	−.80	1.12
20	Small risk formats/Pictogram	.48	.48	0.55	.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 percent 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 difficult 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 ability level that equals the difficulty of the given item. In a two-parameter model such as 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.

* Linked items. The linked items were administered to respondents form A and form B and therefore had a larger sample size from which to obtain item statistics.

≠ Revised items. These items were revised and retested in the validation sample of 200 respondents.

Table 4

Table 4a. Construct Validity of the Numeracy Understanding in Medicine Instrument (NUMi)

	NUMi Score Number of Items Correct	Ability 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

Table 4b. Construct Validity of the Numeracy Understanding in Medicine Instrument (NUMi)

Socio-demographic Factor	NUMi Score Mean (SD)	Ability Score (θ) Mean (SD)
Race/Ethnicity*		
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*		
Up to 12 years	9.8 (3.7)	−0.56 (0.72)
Some college	11.3 (4.0)	−0.25(0.77)
Four 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. Table 4a presents Pearson correlation coefficients between the following measures: NUMi: Numeracy Understanding in Medicine Instrument; Lipkus: Lipkus expanded numeracy scale (21); MDIT: Medical Data Interpretation Test (22); WRAT-A: Wide Range Achievement Test in Arithmetic (42); Wonderlic: Wonderlic Cognitive Aptitude Test (39), S-TOFHLA; Short Test of Functional Health Literacy in Adults (Ref)

Table 4b demonstrates association of socio-demographic factors and performance on the NUMi.

*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.