Numeracy
Advancing Education in Quantitative Literacy

Volume 13 | Issue 1 | Article 3

2020

Effects of Quantitative Literacy on Healthcare Decision-Making: An Aural Context

Robert G. Root
Lafayette College, robroot@lafayette.edu
Sonia Bhala
National Cancer Institute, Division of Cancer Epidemiology and Genetics, sonia.bhala@gmail.com

Recommended Citation
Root, Robert G., and Sonia Bhala. "Effects of Quantitative Literacy on Healthcare Decision-Making: An Aural Context." *Numeracy* 13, Iss. 1 (2020): Article 3. DOI: https://doi.org/10.5038/1936-4660.13.1.3

Effects of Quantitative Literacy on Healthcare Decision-Making: An Aural Context

Abstract

We propose a relationship between sensory modality, numerical formatting, and performance on a survey simulating healthcare decision-making. We examine the current literature on aural health literacy, and specifically aural literacy coupled with health numeracy. We then create a survey instrument called the Bhala test for this purpose and demonstrate that it is moderately internally consistent and provides results that correlate with the NUMi assessment, a widely accepted measure of health numeracy. The quantitative information provided in the Bhala test has two treatments, percentage and natural frequency formats, in an effort to determine which format is easier for subjects to use in decision-making. The Bhala test is administered to a convenience sample of Mechanical Turk workers in a randomized comparative structure. The results do not support the hypothesis that numerical formatting affects subjects' ability to make healthcare decisions. By comparing these results to previous studies on numerical formatting, we provide evidence to support the notion that sensory modality is an essential component of numeracy, and that aural numeracy should be considered separately from print numeracy.

Keywords

numerical formatting, health literacy, subjective numeracy, percent, natural frequency

Creative Commons License

Cover Page Footnote

Robert G. Root is professor and former head of the Department of Mathematics at Lafayette College in Easton, PA. He is currently serving as the Clerk of the Faculty at Lafayette. He enjoys reading Numeracy; this is his third contribution to the journal.

Sonia Bhala is currently a first-year medical student who is passionate about gaining a better understanding of health numeracy and one day applying that knowledge to patient care. She completed this interdisciplinary scholarly work under the mentorship of Dr. Root while obtaining her B.S. in Neuroscience at Lafayette College.

Introduction

The primary goal of this paper is to describe an experiment assessing an intersection between health literacy and numeracy, particularly differentiating between whether quantitative information is presented orally or in writing. Health numeracy, a component of the broader health literacy, is associated with quantitative health information. Health literacy has been associated with positive health outcomes. Poor health literacy can lead to "more hospitalizations; greater use of emergency care; lower receipt of mammography screening and influenza vaccine; poorer ability to demonstrate taking medications appropriately; [and] poorer ability to interpret labels and health messages" (Berkman et al. 2011). Approximately 80 million US adults have inadequate health literacy, which is especially high in vulnerable populations such as those in poverty, racial minorities, and the elderly (Kutner et al. 2006). Oral and aural health literacy are components of health literacy that refer to speaking and listening skills, respectively (Nouri and Rudd 2015). Poor aural literacy is associated with negative patient outcomes (i.e., poor asthma management and higher risk of cardiovascular disease) (Martin et al. 2011a; Rosenfeld et al. 2011; Nouri and Rudd 2015). Furthermore, oral/aural health literacy is significantly associated with increased patient self-advocacy (Martin et al. 2011b; Nouri and Rudd 2015). Low health numeracy biases patient choice and interferes with effective physician-patient risk communication, which is an essential component of quality care. Individuals with low numeracy have lower health outcomes, are less likely to comply with their medication, and have less access to treatment (Bowling 2001; Reyna et al. 2009). Thus, improving health numeracy is critical in efforts to reduce healthcare disparities.

One example of using health numeracy in practical decision making is interpreting numbers to formulate informed decisions about options for treatment incorporating known risks and benefits. Including this type of numeracy as a critical component to health literacy is a relatively new and understudied concept. Since health numeracy research has become a growing field in the last couple decades, reviews of health numeracy have only recently appeared. Defining health numeracy has been a difficult task, since it is a broad concept which requires several different skills (Ancker and Kaufman 2007). In contemporary healthcare, data is complex, and proficiency requires more than just basic calculations, extending to comfort making decisions that compare magnitudes of numbers represented in a variety of ways and making cost-benefit analyses. For the purposes of our study, we consider health numeracy to be applying numeracy to weigh costs, benefits, and health risks of available drugs and treatment options. This concept is a subset of the skills associated with "expressing" quantitative literacy in support of personal health in *Mathematics and Democracy*: *The Case for Quantitative Literacy* (Steen 2001).

Health numeracy in this context does not involve identifying the relevance of quantitative information, nor does it include unprompted application of quantitative reasoning. Limiting the definition like this is adequate for this paper, as it includes all contexts that the study described here attempts to measure.

In this paper we hypothesize that health numeracy is dependent on the sensory modality used to present information, and health numeracy builds upon the skill set associated with the sensory modality used. That is, without mastering the appropriate skill set for processing information associated with the sensory modality used, one cannot hope to achieve adequate health numeracy, as evidenced by the work of Boersma and Klyve (2013).

It seems obvious that health numeracy requires various skill sets because healthcare information is presented in a variety of formats and contexts (i.e., food labels, TV ads, listening to a physician during an examination). However, there is nuance even in this common-sense observation. Previous work has shown that adults with low health literacy obtain healthcare information differently from those proficient in health literacy. According to Kutner et al.'s (2006, 18) assessment of health literacy in U.S. adults, "lower percentages of adults with Below Basic health literacy than adults with Basic, Intermediate, or Proficient health literacy reported that they got information about health issues from any written sources, including newspapers, magazines, books or brochures, and the Internet … higher percentages of adults with Below Basic or Basic health literacy than adults with Intermediate health literacy received a lot of information about health issues from radio and television" (. This evidence suggests that those with low health literacy are more likely to obtain information in a way that is delivered aurally than in print, highlighting the need for aural health literacy assessment tools. Further highlighting this need, critical healthcare information is frequently given in a physician's consultation, where it is provided almost exclusively aurally (Nouri and Rudd 2015).

So, in this paper we focus on the least studied and arguably most critical sensory modality: aural health numeracy. This focus addresses a gap in the literature: to the best of our knowledge, there are no previous studies on aural health numeracy. The evidence that sensory modality matters when assessing health literacy is extensive (Baker 2006; Schonlau et al. 2011). The effect of different sensory modalities on the numeric subset of health numeracy remains an unexplored area of study, to the authors' knowledge. Exploring this effect motivated the decision to construct a survey instrument parallel to that of Woloshin and Schwartz (2011).

Prior studies on the effect of number formatting can be interpreted to support the claim that the sensory modality used to convey quantitative information in a particular format affects the subjects' success in interpreting the formatted information. In attempts to compare cognitive load, previous studies have examined whether presenting numbers in percentage (i.e., 2%) vs. frequency formats (i.e., 2 in 100) are more efficient in a healthcare context, but results of these studies have appeared contradictory (Galesic et al. 2009; Woloshin and Schwartz 2011). The 2009 study by Galesic et al. demonstrates that natural frequencies improve comprehension compared to percentages for older adults and people with low numerical literacy in a health literacy context. Gigerenzer believes that natural frequencies are more efficient at facilitating insight because understanding percentages requires calculating conditional probabilities, a complex formula that entails multiple multiplications and involves normalizing the numbers to add up to 100%, whereas natural frequencies do not (2011). However, the 2011 article by Woloshin and Schwartz offers contradictory evidence, namely that using percentages improved numerical comprehension of healthcare data among their subjects. In their study, Woloshin and Schwartz (2011, 95) offer justification for why percentages may be more efficient based on the concept of denominator neglect, stating "people tend to focus on the numerator of a frequency and ignore the denominator …. [V]ariable frequency formats may be confusing because the larger number in the denominator means a smaller probability." They also state that the "denominator effect may cause problems even when the denominator is held constant … because it magnifies numerically small effects." In response, Gigerenzer (2011, 2) claims that the discrepancy lies in the definition of natural frequency used; he states that ""Natural frequencies are joint frequencies, such as the number of women who test positive and who have breast cancer. These differ from simple frequencies, such as two in 10 people who test positive.... What the *Annals of Medicine* article [Woloshin and Schwartz (2011)] did was compare simple percentages (such as 2% of people who took a drug and had diarrhea) and other formats against simple frequencies (20 in every 1000 people who took the drug had diarrhea), which it called natural frequencies. However, the computational advantage does not apply to simple frequencies."

We believe that the discrepancy between Woloshin and Schwartz's (2011) results and those of Galesic et al. (2009) are a consequence of more than just different definitions and settings; these two studies also use distinct methodologies which call for different sets of information processing skills, comparable to the different skill sets associated with different sensory modalities. A consistent interpretation of these different results would be that percentages appear to be a more efficient format when data is presented in tables (Woloshin and Schwartz 2011), whereas frequency appears to be the more efficient format when the data is presented in a written narrative (Galesic et al. 2009). Gigerenzer's narrative relies on locating information from prose, whereas Woloshin and Schwartz's tables could be considered a graphical representation of numerical information.[1] Just as

[1] This paper does not test this specific hypothesis concerning the difference between reading quantitative information in text vs. in a table. We leave that to later work.

presentation in text vs. table can alter health numeracy, we hypothesize that sensory modality must be taken into account; one of the aims of this paper is to stimulate dialogue about the role of sensory modality on health numeracy in order to better understand the reason for these types of discrepancies (Galesic et. al 2009; Gigerenzer 2011; Woloshin and Schwartz 2011).

This paper is a first step in addressing these conflicting results in the literature. Our study seeks to explore how people process different forms of numerical information when it is spoken, and whether numeric formatting affects performance on a survey simulating healthcare decision-making.

Methods

To assess respondents' numeric competencies when presented orally with data in natural frequency (NF) or percentage (P) format, we adapted a 10-question assessment from the works of Galesic et al. (2009), Gigerenzer (2011), and Woloshin and Schwartz (2011); we call it the Bhala test. Each question has 1–3 parts. In total, our survey takes less than 10 minutes to complete, with some respondents finishing it in less than 3 minutes. This design allows us to observe differences in response due to numerical formatting after presenting relevant information aurally and only once, which are prevalent features when information is presented in a doctor's office or in an advertisement. The full test and scoring guidelines are included in the Appendix.

Choice of Instrument

Some established health numeracy measurement tools include the Test of Functional Health Literacy in Adults (TOFHLA) and its modified shorter version, the S-TOFHLA (Parker et al. 1995; Baker et al. 1999). Others include the Newest Vital Sign (NVS) test and the Medical Data Interpretation Test (MDIT). No studies have been conducted to date where the information provided in these tests is given aurally. In administration of the TOFHLA/S-TOFHLA and NVS, for example, patients are given information to read (print-format), and then orally asked questions about the information (Baker et al. 1999; Weiss et al. 2005). We believe that although the oral (speaking) component is essential to physician-provider dialogue, the aural (listening) component needs to be studied as well.

In determining a tool for our study, we could not use the S-TOFHLA because its questions do not lend itself to altering numerical formatting since it does not have questions that use either percentage or frequency format (Baker et al. 1999). Additionally, S-TOFHLA is not comprehensive enough for our study; it does not lend itself to assessing multiple numeric skills and thus may not adequately assess the broad spectrum of skills required for health numeracy. According to a study by Housten et al. (2018), by adding two additional numeric measures to the S-

TOFHLA, results of performing with “adequate” health literacy were altered 40 to50% of the time among 187 English-speaking adults. The NVS tests numeracy using a nutrition label given in print format and was subject to many of the same limitations as the S-TOFHLA. The NVS does not have questions that utilize either percentages or natural frequencies (Rowlands et al. 2013), and the tabular presentation of the nutrition label does not lend itself to aural formatting. Based on our aims, we found the MDIT (Schwartz et al. 2005) and Woloshin and Schwartz’s 2011 survey to have questions that were closest to our aims, and also include formatting that allows for comparison with previous findings on the effect of numerical formatting on health numeracy. However, MDIT does not test one’s perceptions of their tendency to receive treatment. Consequently, we loosely adapted the fundamental concepts and format of MDIT in the construction of our survey questions.

To the best of our knowledge, a comprehensive aural health numeracy measurement tool assessing multiple numeric competencies, likelihood of requesting treatment, and the ability to make cost-benefit analyses does not currently exist. This study reports on a survey, the first of its kind, in which respondents listened to descriptions of medications, with the formatting of risks taking one of two forms: percentages or natural frequencies. This survey is based on concepts from existing surveys by the Schwartz et al. (2005) and Woloshin and Schwartz (2011), but the original tabular format of presenting healthcare data in the survey was modified into narrative questions presented to subjects aurally.

Development of the Bhala test

Many of our questions are similar in nature to those from the previous study of Woloshin and Schwartz (2011), but the Bhala test uses a narrative format similar to Galesic et al. (2009), although the information is administered orally. The fictional drug names (PAXCID and QUESTOR), the ailment they are to treat (heartburn), and other features of these fictional medications mentioned in the questions were all adopted from the 2011 Woloshin and Schwartz study. Similar to the Woloshin and Schwartz study, our questions ask respondents to assess risks that vary from prevalent (i.e., 89 out of 100) to extremely rare (i.e., 1 out of 20,000) since rarity should be taken into account as a potential factor that may affect risk perception.

Each question is based on a short voice recording (ranging from 2 to10 seconds in length), played only once. After hearing the drug information, participants read the associated question and either typed or clicked on their answer. All questions requiring objective calculations are open-ended, but questions with a subjective component are multiple choice.

Here is the temporal structure of the test: During the test, a sound clip at the beginning of the survey made sure that the participant’s audio was functioning properly. The first question contains verbal information only with no numeric information; this question served as a screening tool for basic aural literacy prior to beginning the aural numeracy assessment. The main body of the test, questions 2–8, provide brief descriptions delivered in an audio recording followed by a (usually multipart) question assessing the subject's understanding of the information provided by the clip and the ability to reason using that information. Upon post-experimental peer review of the survey, Question 5 was considered inappropriate for our study aims, and later excluded from our analyses. Question 9 asks the participant to self-report his or her comfort with numbers on a Likert scale; we refer to this as a subjective assessment of numeracy (SubjResp). The last question (Question 10) is a demographic question asking how many medications the participant takes and was included for research purposes in order to observe whether there was any association between the number of medications a subject used and the subject's health numeracy levels. Further information, including a copy of the test and scoring guidelines, is included in the Appendix.

Two versions of the Bhala test were administered, one with all numerical data in percentages (P) format and one with the same numerical data in natural frequency (NF) format; both versions are indicated in the Appendix. The survey interface included a component that allowed for randomization of which format of the test was given to the participant. Scoring guidelines and a short description of what each question intended to measure is also included in the Appendix.

It is important to note that even though all numerical healthcare data was presented aurally, the questions were in print format. This test design focused on altering the modality in which data was presented relative to previous health literacy assessments. This design allowed for a closer simulation of the sensory modality in which healthcare information is often presented in a physician's office or pharmaceutical advertisement. Requiring the participant to store the question in short-term memory was seen as an unnecessary exaggeration of the increase in cognitive load associated with aural healthcare information.

All statistical analysis was performed using the R programming language, and occasionally Microsoft Excel for data manipulation and basic calculations.

Pilot Testing

Content validity was established by careful selection of test items and by pilot testing. Pilot testing involved 27 students enrolled in two classes at Lafayette College. After analyzing Haun et al.'s (2014) systematic review of 51 health literacy assessments, NUMi was chosen due to its open-source accessibility, short length, and exclusive focus on quantitative literacy in a health-care context. We believe Schapira's 2012 NUMi to be currently the most comprehensive and practical numerical literacy assessment due to its inclusion of tables and graphs, probabilities, and statistical questions in a 20-item multiple choice format. NUMi is widely accepted as a reliable measure of numerical literacy in clinical settings (Haun et al. 2014; Duell et al. 2015).

The pilot testing established that NUMi and Bhala test results correlate, supporting the validity of the Bhala test. The Bhala test is not a duplicate of NUMi and is differentiated from all currently available numerical health literacy tests. The Bhala test is completed more quickly than NUMi and presents all numerical data in an oral format. By presenting information orally and only once, the Bhala test takes short term memory into account and is better able to assess consequences of the subject's cognitive load. Thus, the Bhala test attempts to replicate an important aspect of the process of healthcare decision-making. Since NUMi and most other health numeracy assessments provide all data in readily available print that can be re-read as needed, they have not tested for the cognitive skills that are required to complete the Bhala test.

In the pilot test, first NUMi and then the Bhala test were administered to two groups, one for the natural frequency (NF) version and a second for the percentage (P) version of the Bhala test. NUMi was administered in its typical print form, and the Bhala test was administered aurally. In both NF and P groups, NUMi was given first followed by the Bhala test. Responses for both surveys were recorded on paper. The audio for the Bhala test was provided one time only to all the subjects in the classroom at once. For the percentage version group, the audio was playback of recordings; however, for the natural frequency group the audio was read aloud from the script to all subjects in the classroom.

Participants

In the main study, the Bhala test was administered over the Internet using the Qualtrics platform (https://www.qualtrics.com/) to subjects contacted and recompensed through Amazon's Mechanical Turk (AMT) online marketplace (https://www.mturk.com/mturk/welcome). AMT is a new source of inexpensive, yet high-quality data. Research shows that AMT participants (MTurkers) are "slightly more demographically diverse than standard Internet sample and significantly more diverse than typical American college samples" (Buhrmester et al. 2011). Furthermore, compensation rates do not affect data quality (Buhrmester et al. 2011), and the data obtained through AMT are at least as reliable as traditional data collection methods in cognitive psychology (Buhrmester et al. 2011; Gardner et al. 2012; Bartneck et al. 2015). MTurkers have also been shown to be "more attentive to instructions than were college students," and "MTurkers showed larger effects in response to a minute text manipulation" (Hauser and Schwarz 2016, 400). This finding is especially important since our study is focused on minute manipulation of numerical text. Furthermore, data collection through MTurk has been shown to have "strong test-retest reliability, indicating that [the results do] not significantly change between administration dates" (Holden et al. 2013, 1749).

Consistent with Paolacci et al. (2010), Shapiro et al. (2013,3) report that MTurkers are "younger and more educated than the general US population and are predominantly Caucasian and middle class," which is also true of most college populations. However, based on our review of the literature, we concluded that AMT would provide more accurate results for our study than the traditional college population convenience sample.

The main study subjects were 408 AMT workers, US adults over the age of 18 who completed the survey. We selected for workers who had completed at least 50 Mechanical Turk surveys prior to ours with at least a 95% quality worker rating in order to make sure that we were receiving a high-quality sample that was familiar with the online survey format. The workers were compensated $0.70 and were given a maximum time of 30 minutes to complete the survey, but the average time to complete was approximately 5 minutes.

The survey, administered in December 2016, was titled "Healthcare Decision Making Survey" and was tagged with the keywords numeracy, health, risk, decisions, and math. The description for the survey presented to participants stated, "This is a research study that REQUIRES AUDIO on how numbers affect healthcare decisions. It will require you to do BASIC MATH in your head. It will not work on phones or tablets." After reading this description, the participant could choose whether or not to complete the survey.

Standard exclusion criteria included participants who scored less than 2 out of 6 (equivalent to 1 question) on the verbal literacy screening portion of the assessment. Also, no participants were permitted to take the assessment twice in order to ensure that participants in the percentages group and the natural frequencies group were mutually exclusive. Two participants were rejected using this exclusion criteria, and the remaining 406 participants' data was used for analysis.

Scoring

Detailed scoring guidelines for the Bhala test are included in the Appendix. In assessing the Bhala test, we calculated 3 sum scores: a verbal sum score (*VerbalSum*), an objective sum score (*ObjSum*), and a perception sum score (*PerceptionSum*).

VerbalSum indicates non-numeric oral literacy and ranges from 0–6, with a higher score indicating higher oral literacy.

ObjSum ranges from 0–11, with a higher score indicating higher health numeracy, and assesses the participant's ability to make comparisons, assess ratios, perform common mathematical calculations, and make logical decisions. ObjSum includes all questions that have a correct answer, excluding the non-numeric verbal screening assessment at the beginning of the test.

PerceptionSum ranges from 0–3; a higher score indicates the subject was more likely to request healthcare services. The questions included in the PerceptionSum do not have an objectively right answer, but rather provide information on the participants' preferences. It is important to note that PerceptionSum does not measure a subject's likelihood to take risks, but rather willingness to consume medical care which may be due to risk aversion, or to difference in respondents' belief framework of the medical field (i.e., how they perceive a particular side effect, their trust in the medical field's ability to cure cancer, or their willingness to undergo well-known cancer treatments like chemotherapy). Despite these variables, which should theoretically contribute to within-group variation but not significant between-group variation, we found PerceptionSum a useful measure to explore whether differing numerical formatting significantly altered respondents' likelihood to receive medical treatment.

Statistical Analysis

Statistical analysis was carried out using Microsoft Excel and RStudio. We performed a chi-squared test between the natural frequency (NF) and percentage (P) groups on each part of each question and visualized the data using a mosaic plot or probability histogram. We also performed a regression of objective sum against subjective numeracy assessment (SubjResp) and test type, NF or P. On the verbal sum, objective sum, and subjective score we performed additional tests such as a principal component analysis and the Shapiro-Wilks normality test, and if not normal, a nonparametric bootstrap probability histogram.

Testing the Validity of our Results. In order to test the validity of our results, we used McDonald's omega calculated via the MBESS package in R (Kelley 2017). Although it is common to use Cronbach's alpha in health literacy assessment testing (e.g., Taylor & Byrne-Davis 2016), we support the ideas presented in Dunn et al.'s (2014, 404) review of the advantages of using omega as a measure of internal consistency namely that: "1) Omega makes fewer and more realistic assumptions than alpha; 2) problems associated with inflation and attenuation of internal consistency estimation are far less likely; 3) employing 'omega if item deleted' in a sample is more likely to reflect the true population estimates of reliability through the removal of a certain scale item, and; 4) the calculation of omega alongside a confidence interval reflects more closely the variability in the estimation process, providing a more accurate degree of confidence in the consistency of the administration of a scale." Because the Bhala test is short, alpha is likely to be artificially low, making omega particularly appropriate in this instance.

Since our test was multifactorial, we believe it most appropriate to apply omega to assess our test. Omega for the objective sum included all questions used in the scoring of ObjSum, including the consistency of results measure. Similarly, omega for the perception sum component included all questions used for the

PerceptionSum score. We also calculated an "if item deleted" omega for the objective and perception sum components of our survey if any one question was excluded. It was inappropriate to compute an omega for VerbalSum due to the fact that only three questions were included in this component and there was not enough variance among VerbalSum scores, since most participants received a perfect score on the VerbalSum section and those who did not get at least two questions correct were excluded from the study.

Omega for the objective sum score was computed to be 0.52 (with 95% confidence omega is between 0.44 and 0.60) and omega for the perception sum score was 0.72 (0.65, 0.78), indicating that the Bhala test has a moderate level of internal consistency. Excluding any particular question from our scoring did not considerably improve omega values (Table 1).

Table 1a
Objective Sum Omega

Item Deleted	Omega	95% CI
None (Objective Sum Total Omega)	0.52	[0.44, 0.60]
2A	0.52	[0.44, 0.58]
2B	0.50	[0.41, 0.60]
3A	0.50	[0.42, 0.57]
3B	0.36	[0.25, 0.47]
4A	0.51	[0.44, 0.58]
4B	0.40	[0.30, 0.50]
6A/7A/8A Consistency	0.56	[0.50, 0.62]

Table 1b
Perception Sum Omega

Item Deleted	Omega	95% CI
None (Perception Sum Total Omega)	0.72	[0.65, 0.78]
Q6A	0.73	[0.67, 0.78]
Q7A	0.50	[0.45, 0.55]
Q8A	0.61	[0.55, 0.67]

To demonstrate the independence of the three measures, we performed a principal component analysis on the correlation matrix of the ObjSum, PerceptionSum, and VerbalSum variables. The analysis returned three eigenvalues that were all of roughly equal magnitude (1.07, 1.02, 0.90), indicating that there is no linear dependence among the variables; they are far from multicollinear.

Results

Pilot Testing

Pilot testing done with 27 participants, comprised of two undergraduate classes at the authors' home institution, has shown that the objective sum score of the Bhala test correlates with NUMi. Only the objective sum score—which sums the scores (given in the scoring guidelines in the Appendix) of all the numeric questions that have correct answers—was appropriate for comparison against the NUMi total score because NUMi does not provide information about the participant's verbal literacy or assess the participant's likelihood to receive treatment. This result indicates the validity of the Bhala test in assessing numerical health literacy. In our pilot test, we gave the percentage version of the Bhala test to 14 participants (Group P) and the natural frequency version of the test to 13 participants (Group NF). Group P appeared to have a somewhat significantly higher mean level of quantitative literacy, as indicated by the fact that the mean NUMi score for Group P was 18.36 and the mean NUMi score for Group NF was 17.07 (*p*=0.042). Another indication that Group P was more quantitatively literate was that 50% of participants in Group P scored in the 19–20 range on NUMi (out of a maximum score of 20), whereas only 23% of Group NF participants scored in the 19–20 range. Yet, interestingly, the Bhala test scores were higher for Group NF. Out of a maximum Bhala objective sum score of 10, Group NF scored a mean of 8.85 on the Bhala test and Group P scored a mean of 7.64.[2] Based on a 2-sample *t*-test, we acquired a *p*-value of 0.08, indicating that Group P and Group NF scored similarly on the Bhala test. Although this difference is only marginally significant, the size of the difference appears meaningful.

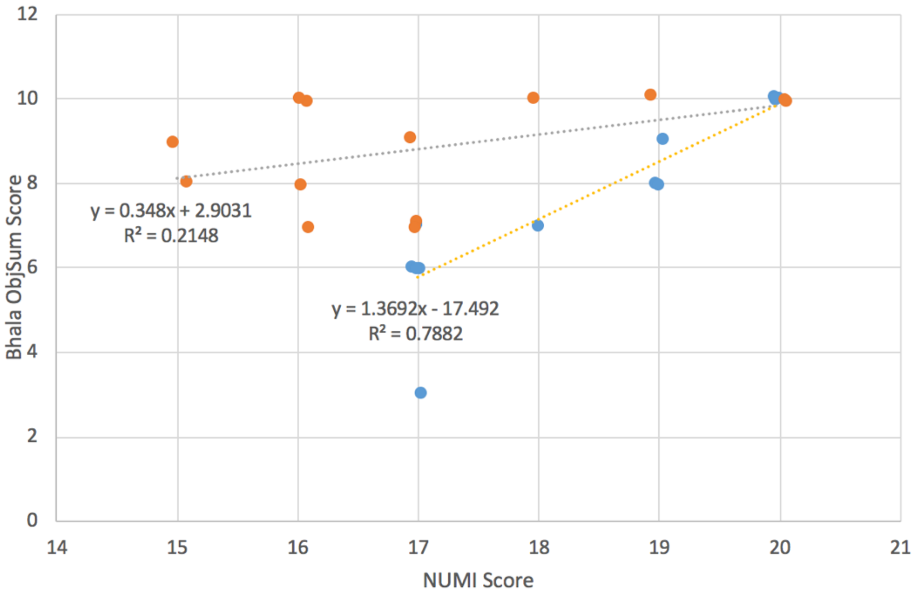

Figure 1. Pilot testing with 27 Lafayette students. Blue circles represent the scores of those given the percentages version of the Bhala test (Group 1, *n*=14) and orange circles represent those given the natural frequencies version (Group 2, *n*=13). Orange circles had lower NUMi scores but higher Bhala test scores; this may be because this group's survey audio was administered by a person in the room rather than a recording.

[2] The scoring in the pilot test was slightly different, resulting in this higher maximum score.

As illustrated in Figure 1, the group with lower mean literacy, as indicated by lower NUMi score, ended up scoring better on the natural frequency test (Group NF) than the group with higher NUMi scores that was given the percentages test (Group P). This result suggests natural frequencies may be a more effective form of communication than percentages when orally presenting healthcare information. However, it is important to note that Group P heard recordings of the information on which the questions are based, and for Group NF the information was read aloud. We believe that if a person in the room is reading the quantitative information, the participant will have a heightened sense of attention and focus on the information being presented, which could contribute to higher Objective Sum scores on the Bhala test and could render the comparison of these two groups suspect. More important is that the two groups each demonstrate that the Bhala test objective sum score and the NUMi score are positively and significantly correlated (r=0.89; p=2.8 x 10^{-9} and r=0.46; p=0.036, for Groups P and NF, respectively). We calculated the one-sided p-values based on the whole model t-statistic. The stronger correlation is associated with the more realistic administration of the Bhala test, with audio coming from a recording.

We considered the possibility that the effect of numeracy skill on health numeracy depends on numerical formatting and/or how the Bhala test is administered. To test this, we performed a regression of ObjSum against two explanatory variables, namely the type of test administered (Type) and NUMi score. The regression included an interaction between the two variables. We found that NUMi score was significantly more predictive of the Bhala test objective sum score in the P group than in the NF group (p=0.002; see Table 2 for a complete summary). This result suggests that the tested dependency does exist, and more specifically, one of two possibilities: either the Bhala test offers increased validity when given in the format of an audio recording as opposed to presented orally in real-time by an individual, or the percentage format of the test has greater validity than the natural frequencies format. The main experimental results will demonstrate that the second of these hypotheses seems less likely. (See Table 4 and the associated description.)

Table 2
Regression of Objective Sum against NUMi and the format of the survey

Coefficients	Estimate	Standard Error	t-value	p-value
Intercept	2.9031	3.1543	0.920	0.3669
NUMi	0.3480^{+}	0.1839	1.893	0.0710
Format P	-20.3954***	5.2127	3.913	0.0007
NUMi-Format P interaction	1.0212**	0.2910	3.510	0.0019
R^2=0.6793				
F-statistic: 16.24 on 3 and 23 DF; p-value: 0.0000				

Note: Significance codes: 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '$^{+}$'

Main Experiment

As described in the Methods section, data were obtained from 406 participants after standard exclusion criteria applied (406/408, 99.5%). 208 of the participants received the natural frequency (NF) version of the test (208/406, 51.2%), and 198 participants received the percentages (P) version of the test (198/406, 48.8%).

Non-Numeric Verbal Literacy. We performed a Shapiro-Wilk normality test and found verbal sum (VerbalSum) to be not normal (*W*=0.31051, *p*<0.05). We subsequently conducted a bootstrap probability test on the verbal sum. As shown in Figure 2, we observed high levels of non-numeric verbal literacy on the verbal screening portion of the Bhala test. 369 out of 406 participants (90.9%) received full points on all three portions of the verbal literacy screening item. Due to the fact that the verbal literacy in our participant population was skewed, it was not used to approximate any other measures or in association with other measures; it merely describes the sample population. Yet, despite this skew, we found that the VerbalSum was significantly correlated with the objective portion of the test (ObjSum), as shown in Table 3 using both the pilot and main experimental data.

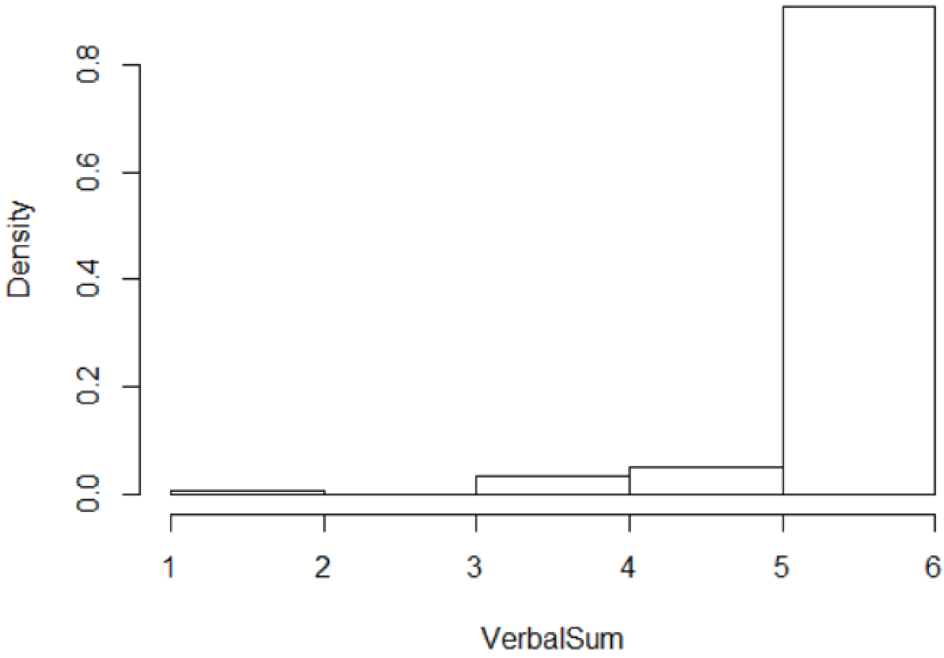

Figure 2. Distribution of verbal responses. Non-normal distribution of non-numeric verbal literacy was indicated by the verbal literacy screening question, which was scored on a scale from 0 to 6. Almost all the participants got full points (6/6) on the 3-part verbal literacy question, indicating high levels of non-numeric oral health literacy within the sample.

Table 3a
Regression of Objective Sum against Verbal Sum in the Pilot Test

Coefficients	Estimate	Standard Error	*t*-value	*p*-value
Intercept	-3.540	3.194	-1.108	0.27831
VerbalSum	4.020**	1.087	3.697	0.00107
R^2=0.3535				
F-statistic: 13.67 on 1 and 25 DF; *p*-value: 0.001074				

Note: Significance codes: 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘+’

Table 3b
Regression of Objective Sum against Verbal Sum in the Experimental Results

Coefficients	Estimate	Standard Error	*t*-value	*p*-value
Intercept	2.0081	1.5744	1.275	0.20287
VerbalSum	0.7620**	0.2677	2.846	0.00465
R^2=0.01966				
F-statistic: 8.1 on 1 and 404 DF; *p*-value: 0.004652				

Note: Significance codes: 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘+’

Health Numeracy. Our results support the assertion that numerical formatting does not affect health numeracy when information is presented aurally. Chi-squared tests were conducted on each question and no question was shown to have significant difference between groups, nor did aggregating questions demonstrate significance. Specifically, numeric format did not lead to significant differences (p<0.05) between the NF or the P groups in either ObjSum or PerceptionSum. The lack of significant differences between the mean objective sum (ObjSum) scores of NF and P are highlighted in Figure 3. We also found no significant association between the portions of the test that assessed for the objective portion of the test (ObjSum) and likelihood of requesting treatment (PerceptionSum) (r=-0.051, p=0.85). There was also no significant correlation between self-reported confidence with numbers (SubjResp) and PerceptionSum (r=0.005, p=0.92); however, there was a modest but highly significant correlation between one’s self-reported confidence in their numeracy and the objective portion of the test (ObjSum) (r=0.281, $p=8 \times 10^{-9}$).

We considered the possibility that, as was observed in the pilot study, the effect of numeracy skill on health numeracy depends on numerical formatting. To test this, we performed a regression of ObjSum against two explanatory variables. Without the NUMi result as an independent measure of numeracy, we used the subjective numeracy response (SubjResp) as a proxy for numeracy skill, and the numerical formatting of the test was an indicator variable (Type). The regression included the interaction term. See Table 4 for a summary of the regression. Although the fit of the model to the data was weak (r^2=0.082) the subjective numeracy response was strongly predictive of ObjSum score (p-value 4.6 x 10^{-5}). However, the numerical formatting made no practical or statistically significant difference in either the intercept or the slope of the regression line, with p-values 0.77 and 0.92, respectively). Thus, we conclude that even for subjects with subjectively low numeracy, the numerical formatting had no effect on their performance.

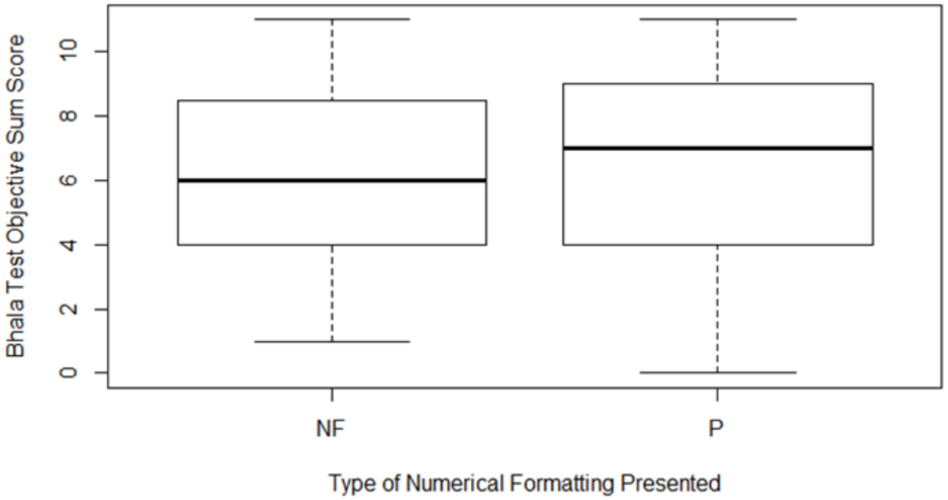

Figure 3. The NF group and P groups of Mechanical Turk participants had very similar mean Bhala test objective sum scores (NF=6.30, P=6.65). Thus, the difference in means was less than 1 point. Note that the objective scale ranged from 0–11 and scores were given as whole numbers only. This means that the difference in Bhala test score between the NF and P group differed by less than 1 question. Differences at least this large occur 32% of the time simply by random variation, so this is unlikely to be due to a meaningful difference between the two forms of the test.

Table 4
Regression of Objective Sum against Subjective Response and Survey Type

Coefficients	Estimate	Standard Error	*t*-value	*p*-value
Intercept	4.33939***	0.50964	8.515	3.37e-16
SubjResp	0.90355***	0.21924	4.121	4.58e-05
Type	0.21346	0.74376	0.287	0.774
SubjResp:Type	0.03023	0.31519	0.096	0.924
R^2=0.08177				
F-statistic: 11.93 on 3 and 402 DF; *p*-value: 1.683e-07				

Note: Significance codes: 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘+’

Discussion

We developed a 10-question aural health numeracy survey. All items were set in a healthcare context and all numerical data was presented orally once and only once. Two versions of our assessment were given; one with all numerical data in percentages and the other with all numerical data in natural frequency format. We found no significant differences between the scores, which differs from the results of previous studies, as illustrated by Table 5.

Table 5
Comparison of Studies of Effect of Presentation on Numeracy

Experimental Study	Sensory Modality Used to Present Numerical Data	Results
Galeic et al. (2009)	Narrative, print format	Participants scored significantly better when information was given in natural frequencies.
Woloshin and Schwartz (2011)	Tabular, print format	Participants scored significantly better when information was given in percentages.
This study	Narrative, oral format	There was no significant difference in scores when information was given in percentages vs. frequencies.

Note: Comparative review of studies assessing whether percentages or frequency format is a more efficient way of presenting data. The results provided in this study is not consistent with those seen by previous studies.

Although our results were never meant to resolve an apparent inconsistency between the previous studies (Galesic et al. 2009; Woloshin and Schwartz 2011), collectively the results of the three studies offer at least one insight. The evidence offered here suggests no difference in comprehension based on numerical format, which is to say that the format seems to be immaterial in an aural modality. The other studies suggest that the different formats do matter when the information is provided in tabular or textual written modality. Collectively, these results support the hypothesis that the sensory modality used to present data might have a significant effect on a participant’s health numeracy. Thus, textual, tabular, and aural numeracy might be hypothesized to be distinct modalities when assessing health literacy. Thus the notion that sensory modality is a key variable affecting health numeracy assessments may shed light on the current inconsistency in results assessing whether natural frequencies or percentages are more effective format. Of course, rigorous testing of this hypothesis is a topic for future work, as we mention in the final paragraph of this paper.

Since there were no significant differences between the frequency and percentage versions of the Bhala test, we conclude that when data is presented aurally, numerical formatting has no significant effect, at least in a population reflected by the sampling frame of Amazon Mechanical Turk workers. It is possible that this is explained by a lack of genuinely low numeracy subjects within the sampling frame, a possible shortcoming of the use of Amazon Mechanical Turk. It is also possible that this is due to the use of SubjResp as a proxy for numeracy skill; it is less correlated with ObjSum than the NUMi score was in the pilot study. As a result, meaningful differences between survey types might be lost in random variation despite the relatively large sample size.

The effect that numerical formatting has on health numeracy may be related to whether the number is being heard rather than being viewed. When data is presented orally in a physician’s office or on a pharmaceutical TV/radio advertisement, verbal context may affect health numeracy more than numerical formatting. Unsurprisingly, higher verbal literacy was correlated with higher performance on the objective portion of the Bhala test. This finding further supports previously published evidence that verbal literacy is needed to accurately assess quantitative information within context (Boersma and Klyve 2013). While this study offers no evidence that numerical formatting alters respondents' ability to process information, this lack of difference is interesting in light of existing work.

Limitations

A major limitation of this study derives from the lack of a probability sample from a representative population: non-numerical aural literacy (also known as verbal literacy) was high in almost all participants. This does not adequately represent the verbal literacy of the general population. Since literacy must exceed a threshold of competency before numeracy can be expected (Boersma and Klyve 2013), we expect that the high verbal literacy of our convenience sample makes it dangerous to generalize our results. Additionally, we disclosed that the survey includes basic math and subject response was voluntary. Those who were more comfortable with math may be more likely to volunteer for a math-related task. Furthermore, like Paolacci et al. (2010), Shapiro et al. (2013, 3) find that MTurkers are "younger and more educated than the general US population and are predominantly Caucasian and middle class," which is also true of most college populations. Using a sample that most likely had higher verbal literacy and numeracy compared to the general population is likely to have biased our results.

Another major limitation of this study was lack of control for whether the participant was taking notes while the healthcare data was being presented. If the participant wrote down the numbers, the cognitive load imposed by short-term memory would be reduced and the participant would be using multimodal processing (motor aspect of writing the number down, orally listening to the numbers, and visually processing what was written down) when interacting with the information, which is distinct from exclusively aural processing. However, since both groups could have potentially jotted down notes, this would account for within-group variation rather than between-group variation, allowing us to still test for differences between the NF and P groups. Additionally, participants who took the initiative to write notes may also demonstrate this behavior while in a physician's office. Thus, including the potential to write down notes into the survey design provided a more generalizable testing environment, at the cost of more accurately capturing the effects of exclusive aural processing in a healthcare setting.

We are also limited by the fact that our participants could not be given the exact stimuli from Woloshin and Schwartz (2011), because the data in that study was presented in a tabular format. Although having exactly the same questions would make our comparison stronger by controlling for more variables, the original instrument did not allow for perfect audible adaption. Additionally, it was not feasible to use AMT to create an assessment that measured both oral and aural numeracy (speaking and listening skills, respectively). There is a need for future studies that simultaneously measure oral and aural numeracy, providing a better representation of the dynamic, two-sided nature of the patient-provider dialogue. It is also important to note that the use of hypothetical scenarios, especially given via an online interface such as AMT, will likely underestimate the barriers of aural numeracy in the healthcare environment, since it removes many stressors (anxiety, critical illness, time pressure, etc.) that influence quantitative literacy.

We further acknowledge that we are limited by our operational definitions. For example, measuring risk perception and risk comprehension is a whole field of study and there are many measures for this. It would be impractical to provide all of the measures needed to capture risk perception and comprehension as a whole. Therefore, in our study, we used how likely the participant is to choose a drug treatment option and whether these decisions were logical given known risk and benefit information as a proxy. This decision provides some insight into the participant's risk perception and comprehension but is unable to capture all components involved when assessing risk, such as participant attitudes and motivation. In a broader sense, the measures of aural health numeracy need to be further validated and used in future work—increased use of consistent measures of aural numeracy in published literature will allow for better cross-comparison studies on numerical formatting. We were limited by a paucity of literature on the subject and the fact that there is no existing aural health numeracy instrument.

Future Directions

Developing a measure of aural health numeracy is a crucial lacuna in the study of health literacy, since healthcare information is often given exclusively aurally within a physician's office (Nouri and Rudd 2015). As suggested by Roter et al.'s (2009, 396) study on aural health literacy, "the medical dialogue may also be made more effective and 'patient-friendly' by attending to language characteristics and dialogue structure"; we consider spoken numerical formatting to be an essential component of this dialogue structure. Thus, we believe that aural health literacy and aural numeracy are important issues that require further investigation. Although future validation and refinement is needed, the Bhala test is a useful and novel measure of aural health numeracy that will be helpful in future research on numerical formatting, especially because there are no other aural health numeracy assessments currently available.

Future research in this field may aid community programs that serve populations with low health literacy; these organizations will be able to create more targeted and more effective programs with increased knowledge about the subtypes of health literacy, including aural health numeracy. Better understanding the role of numerical formatting may also influence health policy decisions and have the potential to influence health communication training for providers.

It is important to note sensory modality types (visual, aural, etc.) are not mutually exclusive; healthcare information is sometimes given in multi-modal formats. Future research would benefit from controlled and randomized studies on health numeracy and numerical formatting from a multi-modal perspective.

Acknowledgements

The authors would like to thank Dr. Luis Schettino for his constructive feedback, which was essential to developing this manuscript. We would like to thank Dr. Alan Childs, Dr. Michael Nees, Ms. Terese Heidenwolf, Mr. Jon Macasevich, Ms. Greta Brubaker, Dr. Andrew Vinchur and the Lafayette College IRB, the Lafayette College Academic Research Committee and Dean Jamila Bookwala, Dr. William Klein, Dr. Rebecca Ferrer, Dr. Sharon A. Savage, and the editor and anonymous reviewers of *Numeracy*, for their invaluable support and advice during the course of the study. We also would like to thank the Lafayette College for financially supporting this project. Finally, we would like to thank Dr. Gary Gordon and the Lafayette Mathematics Department for encouraging interdisciplinary independent research courses, without which this research would not have been possible.

References

Ancker, Jessica S., and David Kaufman. 2007. “Rethinking Health Numeracy: A Multidisciplinary Literature Review.” *Journal of the American Medical Informatics Association* 14.6: 713–721. https://doi.org/10.1197/jamia.M2464

Baker, David W. 2006. “The Meaning and the Measure of Health Literacy.” *Journal of General Internal Medicine* 21.8: 878–883. https://doi.org/10.1111/j.1525-1497.2006.00540.x

Baker, David W., Mark V. Williams, Ruth M. Parker, Julie A. Gazmararian, and Joanne Nurss. 1999. “Development of a Brief Test to Measure Functional Health Literacy.” *Patient Education and Counseling* 38.1: 33–42. https://doi.org/10.1016/S0738-3991(98)00116-5

Bartneck, Christoph, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. “Comparing the Similarity of Responses Received from Studies in Amazon’s Mechanical Turk to Studies Conducted Online and with Direct Recruitment.” *PloS one* 10.4: e0121595. https://doi.org/10.1371/journal.pone.0121595

Berkman, Nancy D., Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karen Crotty. 2011. “Low Health Literacy and Health Outcomes: An Updated Systematic Review.” *Annals of Internal Medicine* 155.2: 97–107. https://doi.org/10.7326/0003-4819-155-2-201107190-00005

Boersma, Stuart, and Dominic Klyve. 2013. “Measuring Habits of Mind: Toward a Prompt-less Instrument for Assessing Quantitative Literacy.” *Numeracy* 6.1: 6. https://doi.org/10.5038/1936-4660.6.1.6

Bowling, Anne, and Shah Ebrahim. 2001. “Measuring Patients’ Preferences for Treatment and Perceptions of Risk.” *Quality in Health Care* 10.S1: i2–i8. https://doi.org/10.1136/qhc.0100002

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. “Amazon’s Mechanical Turk: A New Source of Inexpensive, yet High-quality, Data?” *Perspectives on Psychological Science* 6.1: 3–5. https://doi.org/10.1177/1745691610393980

Duell, Paul, David Wright, Andre M. N. Renzaho, and Debi Bhattacharya. 2015. “Optimal Health Literacy Measurement for the Clinical Setting: A Systematic Review.” *Patient Education and Counseling* 98.11: 1295–1307. https://doi.org/10.1016/j.pec.2015.04.003

Dunn, Thomas J., Thom Baguley, and Vivienne Brunsden. 2014. “From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation.” *British Journal of Psychology* 105.3: 399–412. https://doi.org/10.1111/bjop.12046

Galesic, Mirta, Gerd Gigerenzer, and Nils Straubinger. 2009. “Natural Frequencies Help Older Adults and People with Low Numeracy to Evaluate Medical Screening Tests.” *Medical Decision Making* 29.3: 368–371. https://doi.org/10.1177/0272989X08329463

Gardner, Rick M., Dana L. Brown, and Russell Boice. 2012. “Using Amazon’s Mechanical Turk Website to Measure Accuracy of Body Size Estimation and Body Dissatisfaction.” *Body image* 9.4: 532–534. https://doi.org/10.1016/j.bodyim.2012.06.006

Gigerenzer, Gerd. 2011. “What Are Natural Frequencies? Doctors Need to Find Better Ways to Communicate Risk to Patients.” *BMJ* 343.7828. https://doi.org/10.1136/bmj.d6386

Haun, Jolie N., Melissa A. Valerio, Lauren A. McCormack, Kristine Sørensen, and Michael K. Paasche-Orlow. 2014. “Health Literacy Measurement: An Inventory and Descriptive Summary of 51 Instruments.” *Journal of health communication* 19.sup2: 302–333. https://doi.org/10.1080/10810730.2014.936571

Hauser, David J., and Norbert Schwarz. 2016. “Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants.” *Behavior research methods* 48.1: 400–407. https://doi.org/10.3758/s13428-015-0578-z

Holden, Christopher J., Trevor Dennie, and Adam D. Hicks. 2013. “Assessing the Reliability of the M5-120 on Amazon’s Mechanical Turk.” *Computers in Human Behavior* 29.4: 1749–1754. https://doi.org/10.1016/j.chb.2013.02.020

Housten, Ashley J., Lisa M. Lowenstein, Diana S. Hoover, Viola B. Leal, Geetanjali R. Kamath, and Robert J. Volk. 2018. "Limitations of the S-TOFHLA in Measuring Poor Numeracy: A Cross-Sectional Study." *BMC Public Health* 18.1: 405. https://doi.org/10.1186/s12889-018-5333-9

Kelley, K. 2017. MBESS Version 4.4.0 [computer software and manual]. Accessible from http://cran.r-project.org.

Kutner, Mark, Elizabeth Greenberg, Ying Jin, and Christine Paulsen. 2006. "The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483." *National Center for Education Statistics*.

Martin, Laurie T., Matthias Schonlau, Ann Haas, Kathryn Pitkin Derose, Rima Rudd, Eric B. Loucks, Lindsay Rosenfeld, and Stephen L. Buka. 2011a. "Literacy Skills and Calculated 10-year Risk of Coronary Heart Disease." *Journal of General Internal Medicine* 26.1: 45–50. https://doi.org/10.1007/s11606-010-1488-5

Martin, Laurie T., Matthias Schonlau, Ann Haas, Kathryn Pitkin Derose, Lindsay Rosenfeld, Stephen L. Buka, and Rima Rudd. 2011b. "Patient Activation and Advocacy: Which Literacy Skills Matter Most?" *Journal of Health Communication* 16.S3: 177–190. https://doi.org/10.1080/10810730.2011.604705

Nouri, Sarah S., and Rima E. Rudd. 2015. "Health Literacy in the 'Oral Exchange': An Important Element of Patient–Provider Communication." *Patient Education and Counseling* 98.5: 565–571. https://doi.org/10.1016/j.pec.2014.12.002

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010."Running Experiments on Amazon Mechanical Turk." *Judgement and Decision Making* 5.5: 411–419.

Parker, Ruth M., David W. Baker, Mark V. Williams, and Joanne R. Nurss. 1995. "The Test of Functional Health Literacy in Adults." *Journal of General Internal Medicine* 10.10: 537–541. https://doi.org/10.1007/BF02640361

Peters, Ellen, and Par Bjalkebring. 2014. "Multiple Numeric Competencies: When a Number is Not Just a Number." *Journal of Personality and Social Psychology* 108.5: 802. https://doi.org/10.1037/pspp0000019

Reyna, Valerie F., Wendy L. Nelson, Paul K. Han, Nathan F. Dieckmann. 2009. "How Numeracy Influences Risk Comprehension and Medical Decision Making." *Psychological Bulletin*, 135.6: 943–973. https://doi.org/10.1037/a0017327

Roter, Debra L., Lori Erby, Susan Larson, and Lee Ellington. 2009. "Oral Literacy Demand of Prenatal Genetic Counseling Dialogue: Predictors of Learning." *Patient Education and Counseling* 75.3: 392–397. https://doi.org/10.1016/j.pec.2009.01.005

Rosenfeld, Lindsay, Rima Rudd, Karen M. Emmons, Dolores Acevedo-García, Laurie Martin, and Stephen Buka. 2011. "Beyond Reading Alone: The Relationship between Aural Literacy and Asthma Management." *Patient Education and Counseling* 82.1: 110–116. https://doi.org/10.1016/j.pec.2010.02.023

Rowlands, Gill, Nina Khazaezadeh, Eugene Oteng-Ntim, Paul Seed, Suzanne Barr, and Barry D. Weiss. 2013. "Development and Validation of a Measure of Health Literacy in the UK: The Newest Vital Sign." *BMC Public Health* 13.1: 116. https://doi.org/10.1186/1471-2458-13-116

Schapira, Marilyn M., cindy M. Walker, Kevin J. Cappaert, Pamela S. Ganschow, Kathlyn E. Fletcher, Emily L. McGinley, Sam Del Pozo, Carrie Schauer, Sergey Tarima, and Elizabeth A. Jacobs. 2012. "The Numeracy Understanding in Medicine Instrument: A Measure of Health Numeracy Developed Using Item Response Theory." *Medical Decision Making* 32.6: 851–865. https://doi.org/10.1177/0272989X12447239

Schwartz, Lisa M., Steven Woloshin, and H. Gilbert Welch. 2005. "Can Patients Interpret Health Information? An Assessment of the Medical Data Interpretation Test." *Medical Decision Making* 25.3: 290–300. https://doi.org/10.1177/0272989X05276860

Schonlau, Matthias, Laurie Martin, Ann Haas, Kathryn Pitkin Derose, and Rima Rudd. 2011. "Patients' Literacy Skills: More than Just Reading Ability." *Journal of Health Communication* 16.10: 1046–1054. https://doi.org/10.1080/10810730.2011.571345

Shapiro, Danielle N., Jesse Chandler, and Pam A. Mueller. 2013. "Using Mechanical Turk to Study Clinical Populations." *Clinical Psychological Science* 1.2: 213–220. https://doi.org/10.1177/2167702612469015

Steen, Lynn Arthur, ed. 2001. "The Case for Quantitative Literacy." *Mathematics & Democracy: The Case for Quantitative Literacy*. Washington, DC: National Council on Education and the Disciplines.

Taylor, Anne A., and Lucie M. Byrne-Davis. 2016. "Clinician Numeracy: The Development of an Assessment Measure for Doctors." *Numeracy* 9.1: 5. https://doi.org/10.5038/1936-4660.9.1.5

Weiss, Barry D., Mary Z. Mays, William Martz, Kelley Merriam Castro, Darren A. DeWalt, Michael P. Pignone, Joy Mockbee, and Frank A. Hale. 2005. "Quick Assessment of Literacy in Primary Care: The Newest Vital Sign." *The Annals of Family Medicine* 3.6: 514–522. https://doi.org/10.1370/afm.405

Woloshin, Steven, and Lisa M. Schwartz. 2011. "Communicating Data about the Benefits and Harms of Treatment: A Randomized Trial." *Annals of Internal Medicine* 155.2: 87–96. https://doi.org/10.7326/0003-4819-155-2-201107190-00004

Appendix: The Bhala Test

Here is the full text of the survey instrument used for this study. The numbered passages were provided via audio recordings, while the lettered questions based on each recording were readable to subjects on their computer screen.

The different representations of quantitative information used in the two versions of the survey are shown in brackets, separated by a vertical bar, with the percentage version given first. For example, the passage [89%|89 out of 100] indicates that in the percentage version the audio recording included "eighty-nine percent," while the natural frequency version of the recording included "eighty-nine out of one hundred."

When multiple choice responses were offered, they are indicated here with open bullets following the question.

Survey Questions

1. PAXCID, also known as paxoprasole, is a prescription drug for heartburn in adults. It is recommended for men and women bothered by heartburn or acid reflux disease. It should not be taken by women who are pregnant or breast feeding. Side effects include diarrhea and headaches. QUESTOR is a competing heartburn drug to PAXCID. QUESTOR is another option for the same group of people who are eligible to take PAXCID.

a. What is the main reason to take PAXCID or QUESTOR? Select one answer only.

○ To relieve diarrhea
○ To relieve heartburn
○ To prevent stomach cancer
○ To prevent gallstones

b. Who should NOT take PAXCID or QUESTOR?

○ People with heartburn
○ People with congenital heart failure
○ People with dizziness
○ People who are pregnant or breastfeeding

c. Who CAN take PAXID or QUESTOR?

○ Men
○ Women
○ Both men and women
○ Neither men nor women

2. When people were given 20 mg of PAXCID a day, [89%|89 out of 100] reported their heartburn went away. When people were given 20 mg of QUESTOR a day, [19%|19 out of 100] reported their heartburn went away.

a. Would you be more likely to take PAXCID or QUESTOR for your heartburn?

b. What is the percentage point difference between the effects of PAXCID and QUESTOR?

3. In clinical trials, [2% of|2 out of 100] people given QUESTOR had diarrhea. [6% of|6 out of 100] people given PAXCID had diarrhea.

a. If avoiding diarrhea is a central concern, would you take QUESTOR or PAXCID for your heartburn?

b. How many times more likely is it that you'll get diarrhea with PAXCID than with QUESTOR?

4. In clinical trials, [0.005% of|1 out of 20,000] people given PAXCID had a heart attack. [0.01% of|1 out of 10,000] people given QUESTOR had a heart attack.

a. Which drug puts you at higher risk for getting a heart attack?

b. How many more times likely is it that you'll get a heart attack if you choose the drug with the higher risk?

5. When people were given 20 mg of PAXCID a day, [89%|89 out of 100] reported their heartburn went away. When people were given 20 mg of QUESTOR a day, [19%|19 out of 100] reported their heartburn went away. [0.005% of|1 out of 20,000] people given PAXCID had a heart attack. [0.01% of|1 out of 10,000] people given QUESTOR had a heart attack.

a. Which drug would you be more likely to take if you needed relief from heartburn?

b. If your only other option was not receiving treatment for your heartburn, would the possible benefit of Paxcid be worth the risk of side effects?

c. If your only other option was not receiving treatment for your heartburn, would the possible benefit of Questor worth the risk of side effects?

6. You are at a [5 millionths of a percent|5 out of 100 million] chance of contracting an aggressive cancer. However, the test which screens for the disease results in 24 hours of nausea and indigestion.

a. Would you sign up for a screening?

7. You are at a [0.005%|5 out of 100,000] chance of contracting an aggressive cancer. However, the test which screens for the disease results in 24 hours of nausea and indigestion.

a. Would you sign up for a screening?

8. You are at [5%|5 out of 100] chance of contracting an aggressive cancer. However, the test which screens for the disease results in 24 hours of nausea and indigestion.

a. Would you sign up for a screening?

9. Do you think you're generally good with numbers?
Choose one.

○ Way Below Average
○ Below Average
○ Average
○ Above Average
○ Way Above Average

10. How many different types of prescription medication are you currently using?

Survey Scoring Guidelines

The scoring guidelines, indicating the value of each question, and the rules for awarding partial value to responses, when this was permitted, are given below.

Verbal Score (indicates verbal literacy)

Q1a. 2 points if "To relieve heartburn," 0 points otherwise
Q1b. 2 points if "People who are pregnant or breastfeeding," 1 point if "People with heartburn," 0 points otherwise
Q1c. 2 points if "Both men and women," 0 points otherwise
Verbalsum = sum of Q1, Q2, and Q3 is a score from 0 to 6 that will indicate verbal literacy.

Note: incomplete answers are counted as wrong. However, there were no incomplete answers in this portion of the survey.

Objective Score (indicates ability to perform calculations, weigh magnitudes, and make logical decisions based on numerical information)

Q2a. 1 point if "PAXCID," 0 points otherwise

Note: Incomplete answers are counted as wrong. There were a few incomplete answers in this portion of the survey.

This question requires the participant to perform a comparison between magnitudes and to make a logical decision from that comparison. In this question, the participant is comparing which drug has the larger benefit. Therefore, the correct answer is the drug with the larger number.

Q2b. 3 points if "70," 2 points if correct mathematical operation was present (89/100 minus 19/100) with correct data but no answer/wrong answer (indicate the subtraction sign or the word difference), 1 point if repeat data (say 89 and 19) without doing calculation, 0 points if no answer/incorrect

This question requires you to do a calculation between magnitudes.

- "89 versus 19" was counted as 1 point because the word "versus" doesn't indicate a specific mathematical operation (could be a difference, could be a ratio, etc.)
- "97 vs 19 out of 100" and "81% vs 19%" and "Less percentage (only 19) said their heartburn went away with QUESTOR versus PAXCID" were all given 0 points (both numbers need to be heard and stored correctly in order to do a mental calculation and comparison of the data)
- "85-19=66" was given zero points. The first point can only be achieved if the participant has the correct data. Only then can the participant achieve 2 points by knowing that he or she needs to take a difference. If the participant gets the problem completely correct, he or she will get 3 points.

Q3a. 1 point if "QUESTOR," 0 points otherwise. No answer is counted as incorrect.

This question requires the participant to do a comparison between magnitudes and to make a logical decision from that comparison. This question is different from question 4 because in this question the participant is comparing which drug has less side effects. Therefore, the correct answer is the drug with the smaller number.

Q3b. 2 points if "3," 1 points if correct math equation was stated but not carried out or if "1/3" was stated (the inverse answer), 0 if incorrect/no answer

This question requires the participant to do a comparison and then perform a ratio calculation.

Q4a. 1 point if "QUESTOR," 0 points otherwise

This question requires the participant to perform comparisons with fractions/decimals while evaluating a serious side effect. The number that was a smaller denominator will have the higher risk.

Q4b. 2 points if "2," 1 point if "0.5" (the inverse calculation), 0 otherwise

This question requires the participant to do a ratio calculation using complex numbers (fractions/decimals).

Q6a/7a/8a. 1 point if "Consistent," 0 points otherwise

These questions assess the participant's ability to make logical decisions.

ObjSum=This score will tell us the participant's numerical literacy when taking into account the ability to make comparisons, assess ratios, perform common mathematical calculations, and make logical decisions. This score can range from 0–11. The higher the score, the greater the participant's numerical literacy,

Perception Score (PerceptionSum)

Q6a. 1 point if "Yes," 0 points otherwise

This question asks whether the participant will sign up for a preventive screening, given an extremely low risk of contracting the disease.

Q7a. 1 point if "Yes," 0 points otherwise

This question asks whether the participant will sign up for a preventive screening, given a moderate risk of contracting the disease.

Q8a. 1 point if "Yes," 0 points otherwise

This question asks whether the participant will sign up for a preventive screening, given a high risk of contracting the disease.

PerceptionSum=total score of subjective section. This score can range from 0 to 3. The higher PerceptionSum, the more likely the patient is to choose to request healthcare services.

Self-Report

Q9. 0 points if way below average, 1 point if below average, 2 points if average, 3 points if above average, and 4 points if way above average

This question indicates the participant's self-reported comfort with numbers. The score can range from 0 to4. A higher score indicates the participant thinks that he or she is numerically competent.

Real Choices

Q10. This question asks how many prescription medications the participant takes.

Note: Q5 was excluded in the analyses due to its inability to measure its intended construct, based on peer review.