The Numeric Understanding Measures (NUM): Developing and Validating Adaptive and Non-Adaptive Numeracy Scales

Michael C. Silverstein, Pär Bjälkebring, Brittany Shoots-Reinhard, Ellen Peters

Michael C. Silverstein, MS*
Center for Science Communication Research
Department of Psychology
University of Oregon
1715 Franklin Blvd.
Eugene, OR 97403
United States
Email: msilver2@uoregon.edu
https://orcid.org/0000-0001-5497-3578

Pär Bjälkebring, PhD
Department of Psychology, University of Gothenburg
Center for Science Communication Research, University of Oregon
Email: pop-par.bjalkebring@psy.gu.se
https://orcid.org/0000-0002-4430-4044

Brittany Shoots-Reinhard, PhD
Department of Psychology, The Ohio State University
Center for Science Communication Research, University of Oregon
Email: shoots-reinhard.1@osu.edu
https://orcid.org/0000-0002-2844-4995

Ellen Peters, PhD
Center for Science Communication Research
School of Journalism and Communication
Department of Psychology
University of Oregon
Email: ellenpet@uoregon.edu
https://orcid.org/0000-0003-0702-6169

*Address correspondence to:
Michael C. Silverstein
1715 Franklin Blvd, Room 146
Eugene, OR 97403
msilver2@uoregon.edu

Abstract

Numeracy–the ability to understand and use numeric information—is linked to good decision making. Several problems exist with current numeracy measures, however. Depending on the participant sample, some existing measures are too easy or too hard; also established measures often contain items well-known to participants. The current paper aimed to develop new Numeric Understanding Measures (NUM) including a one-item (1-NUM), four-item (4-NUM), and four-item adaptive measure (A-NUM).

In a calibration study, two participant samples (n=226 and 264 from Amazon's Mechanical Turk [MTurk]) each responded to half of 84 novel numeracy items. We calibrated items using two-parameter logistic item response theory (IRT) models. Based on item parameters, we developed the three new numeracy measures. In a subsequent validation study, 600 MTurk participants completed the new numeracy measures, the adaptive Berlin Numeracy Test, and the Weller Rasch-Based Numeracy test, in randomized order. To establish predictive and convergent validities, participants also completed judgment and decision tasks, Raven's progressive matrices, a vocabulary test, and demographics.

Confirmatory factor analyses suggested that the 1-NUM, 4-NUM, and A-NUM load onto the same factor as existing measures. The NUM scales also showed similar association patterns to subjective numeracy and cognitive ability measures as established measures. Finally, they effectively predicted classic numeracy effects. In fact, based on power analyses, the A-NUM and 4-NUM appeared to confer more power to detect effects than existing measures. Thus, using IRT, we developed three brief numeracy measures, using novel items and without sacrificing construct scope. The measures can be downloaded as Qualtrics files (https://osf.io/pcegz/).

Key words: Numeracy, numeric literacy, numeric reasoning, adaptive test, validation, decision making

Objective numeracy—called numeracy throughout the rest of this paper—refers to the ability to understand and use numeric information (Steen, 1990; Peters et al., 2006; Reyna et al., 2009; Peters & Bjälkebring, 2015). With lower levels of numeracy, people can complete basic mathematical processes such as counting, sorting, basic arithmetic, and understanding simple percentages (Peters, 2020). Greater numeracy is required for successful completion of unfamiliar or less explicit numeric tasks, data interpretation, and problems involving multiple steps. As reviewed below, studies have linked these numeric abilities to decision making and life outcomes, making it important to assess numeracy well. Overall, effective measures should: 1) avoid ceiling and floor effects by including a broad range of easy and difficult items, 2) provide a more fine-grained, precise numeracy assessment along the full difficulty continuum of interest without gaps; 3) be brief for time efficiency, and 4) have novel, unfamiliar items to reduce memory and learning effects. The use of item response theory, in particular, then can maximize the information from each item, thus creating briefer scales that will save time for researchers and participants. Finally, any new measure should demonstrate predictive validity similar to or better than existing measures. In this paper, we develop and fully test three new numeracy measures: an adaptive measure, a four-item, non-adaptive measure, and a single-item measure.

Numeric abilities are linked with better decision making

Being numerate is important for effective decision making (Peters, 2020). Even everyday problem solving like weighing numerical information, using statistical information in text and figures, comprehending risk, and weighing numerical information in decisions all require some numeric proficiency (Peters et al., 2006). Despite its importance, the Organization for Economic Cooperation and Development estimated that 29% of US adults are able to do only simple processes with numbers like counting and sorting; they cannot perform math involving two or

more steps and are unable to understand and use percentages, fractions, simple measurements, and figures (Desjardins et al., 2013). For example, among people with diabetes, less numerate people were worse at identifying abnormally high or low blood glucose levels and managing their health as measured by their levels of hemoglobin A1c (Cavanaugh et al., 2008; Zikmund-Fisher et al., 2014). Furthermore, experimentally improving numeracy protected healthy behaviors and financial literacy of college students across a semester (Peters et al., 2017) and produced greater consistency of risk perceptions in an online sample (Chesney, Shoots-Reinhard, & Peters, 2021).

Numerous reasons likely exist for better outcomes among the highly numerate. People higher in numeracy are more likely to understand and respond consistently to numeric information (Woloshin et al., 2001; McAuliffe et al., 2010; Del Missier et al., 2012; Sinayev & Peters, 2015). They also have more precise emotional responses to numbers that appear to allow them to use numbers more in judgments (Peters et al., 2006; Västfjäll et al., 2016). People higher in numeracy also perform more mathematical operations in judgments and decisions that may help them ascertain the meaning of numbers for their decisions (Peters & Bjälkebring, 2015). In contrast, less numerate people tend to rely more on qualitative information (e.g., anecdotes, emotions incidental to a decision, and heuristics; Burns et al., 2012; Dieckmann et al., 2009; Hart, 2013; Peters et al., 2009). For example, less numerate individuals are more susceptible to attribute framing effects (e.g., the difference in ratings of meats marked 75% lean v. 25% fatty; Peters et al., 2006). Finally, more numerate people use more and more complex information in judgments and decisions involving quantities compared to the less numerate (Pachur & Galesic, 2013; Peters & Levin, 2008). In the present validation study, we used tasks focused on comprehension and judgments involving numeric information to test the predictive validity of

our numeracy measures. As in many of the previous studies, we also adjust for other forms of intelligence given that numeracy correlates with them. Of course, none of the scales measure with perfect reliability or validity, making this attempt imperfect. Nonetheless, tests of specific abilities are useful.

Measures of Numeracy

Several scales exist to measure numeracy; however, each scale has distinct problems. First, many measures do not include a sufficient range of difficulty. Some measures include too many easy problems, thus leading to ceiling effects that make it difficult to distinguish among participants higher in ability (e.g., Numeracy Assessment; Schwartz, Woloshin, Black, & Welch, 1997; Numeracy Scale; Lipkus, Samsa, & Rimer, 2001). Another factor affecting existing measures may be familiarity. The Cognitive Reflection Task (CRT) has been widely published (Frederick, 2005), and was expanded to combat familiarity as a result. However, these and other numeracy items have been reused in newer scales, making all of them problematic in terms of the use of familiar items (e.g., Rasch-Based Numeracy Scale; Weller et al., 2013). Other measures are too difficult for many populations (e.g., older adults) which can result in floor effects (Berlin Numeracy Test; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Cognitive Reflection Task; Frederick, 2005; Toplak, West, & Stanovich, 2011).

Other measures may be too long to be cost-effective (Expanded Numeracy Scale; Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007). In this respect, adaptive measures are particularly useful as they focus participants on questions just above, within, and below their ability so that they do not take more time than needed. In general, researchers prefer measures that use less participant time and effort while measuring numeracy in a reliable and valid manner.

One possible solution to these issues is measuring numeracy without a math test (Fagerlin et al., 2007). Fagerlin and colleagues asked participants to self-report their mathematical ability and math preferences and used their combination as a proxy for numeracy. However, Peters and Bjälkebring (2015) concluded that this Subjective Numeracy Scale captured a related but separate numeric competency, suggesting that self-assessment of numeric ability cannot replace objective measures of numeracy. Subjective numeracy appears to relate more to motivations to use numeric ability rather than the ability itself (Peters et al., 2019; Peters & Shoots-Reinhard, 2022; Miron-Shatz et al., 2014; Choma et al., 2019).

An ideal measure would be developed by applying item response theory (IRT)—rather than classical test theory—to a large range of novel items not easily found online. By doing so, one can create short-form measurements that discriminate amongst wide ranges of ability (Smith et al., 2000). IRT also produces more consistent item-difficulty parameters across samples and less measurement error relative to classical testing theory (Magno, 2009). Where possible, adaptive measures can further reduce retest effects as participants see different question items (Arendasy & Sommer, 2017). Moreover, adaptive testing allows for accurate ability estimates using fewer items and less time (Legree et al., 1998). Such finer measurement of numeracy is important to improve understanding of how numeracy relates to decisions and behaviors and how it interacts with other numeric competencies, such as numeric self-efficacy (i.e., numeric confidence; Peters & Bjälkebring, 2015; Peters, Tompkins, et al., 2019). Moreover, more precise measurement also may contribute to understanding how numeracy as a specific ability can improve prediction in tasks involving numbers (i.e., Coyle & Greiff, 2021).

The Current Paper

The current paper introduces a new online adaptive test (A-NUM) and a non-adaptive test of numeracy (4-NUM). The goal of these measures is to assess numeracy effectively while allowing for granular measurement across a meaningful range of ability. Because researchers are sometimes interested in using as brief a measure as possible, we also explore the feasibility of measuring numeracy with a single item (1-NUM). We present the Numeric Understanding Measures (NUM) for numeracy developed using item response theory and compare their performance to currently used numeracy measures. The hypotheses, methods, and analyses for the validation study were pre-registered on Open Science Framework (https://osf.io/9tjgz), although not all hypotheses and their tests are reported herein. In particular, associations with Big 5 Personality Measures for discriminant validity can be found in the supplement (Appendix G). For the data, please visit https://osf.io/kv7cn.

We hypothesized that the new measures would load onto the same latent factor as older numeracy measures. Additionally, we expected the new measures would be positively associated with subjective numeracy and two non-numeric intelligence measures: Raven's progressive matrices, and vocabulary (Cokely et al., 2012; Peters et al., 2010; Peters & Bjälkebring, 2015). Together, these patterns of association would provide convergent validation. Further, to evaluate predictive validity, we examined common decision-making tasks previously shown to be associated with numeracy. We expected to replicate these numeracy associations for existing measures and the new measures. Specifically, we expected the new measures to predict behavior similarly to established measures regarding probability interpretation and benefit perceptions (as in Cokely et al., 2012), attribute framing effects (as in Peters et al., 2006), the effect of a small loss on bet attractiveness (as in Peters et al., 2006; Peters & Bjälkebring, 2015), and risk consistency (as in Del Missier et al., 2012).

Together, support for these hypotheses would indicate the Numeric Understanding Measures to be valid. To foreshadow our results, the NUM measures demonstrated good convergent validity; they loaded onto a single factor with established measures and showed patterns of correlations similar to them. The A-NUM demonstrated predictive validity for all tasks that established measures also predicted. The 4-NUM showed similar patterns of predictive power, and the 1-NUM demonstrated predictive power for most tasks despite being comprised of a single item. Unlike established measures, these new measures offer a fresh start of sorts, by using unfamiliar items whose answers are not easily available online.

Study 1: Calibration Study

Participants

Due to the large number of new items being calibrated, two participant samples were recruited, and each one completed half of the new items. The first sample consisted of 264 participants (53.4% female; $\bar{x}_{age}$=40.47; 79.2% Caucasian, 9.5% Asian, 7.2% African American, 6.8% Hispanic), randomly selected from a cohort of about 1,000 Mechanical Turk workers. The second sample was randomly selected from the same cohort (no overlap of participants existed between the two samples) and included 226 participants (50.0% female; $\bar{x}_{age}$= 41.37; 85.0% Caucasian, 6.6% Asian, 4.4% African American, and 6.2% Hispanic).

Procedures

Following informed consent, each participant was assigned to complete one of two blocks of the total 84 items (42 items each), with each participant responding to only one block. New items were largely generated based on previous scales. After generation, the authors discussed and

refined the items, then sorted them into bins based on perceived difficulty. The authors also attempted to include items from various domains in the final set of items (e.g., medical, financial, etc.) and to cover a wide range of math processes (e.g., arithmetic, probability transformation, cumulative risk, etc.). The final set was piloted with research assistants to ensure they were neither too hard nor too easy and were easy to understand. Each block was designed to contain both easy and hard questions. To ease participant burden and prevent a potential uneven dropout of participants with high math anxiety or low numeric self-efficacy, each 42-item block was broken into sub-blocks of 21 items completed in two separate sessions, spaced approximately 7 days apart. Overall, 91.7% and 88.6% of participants returned for the second part of the study in the first and second samples, respectively. The order of the sub-blocks was counterbalanced to prevent any order effects. All items were open-ended and only numeric responses were allowed. Some items were modified from previous numeracy scales and additional new questions were created; none of their answers could be found online. In addition, all participants completed a common question to test if the samples differed in ability. Participants did not significantly differ in terms of their accuracy on this common question, χ^2 (df = 2, N = 490) = 5.14, p = .08.

Item Calibration

The 84 items were calibrated for difficulty and discriminability using a two-parameter logistic model for dichotomous responses (correct v. incorrect) using FlexMIRT (Cai, 2013). A two-parameter model estimates how well an item discriminates between levels of ability (α_i) and at what level of ability half of participants are expected to get an item correct (β_i; An & Yung, 2014). The probability that an examinee at some ability level, θ_j, will get an item correct is represented by the following equation:

$$P(x_i = 1|\theta_j) = \frac{1}{1 + e^{\alpha_i(\theta_j - \beta_i)}}$$

Rather than assuming that the abilities of the participants are perfectly normally distributed, the distribution of ability was empirically estimated using quadratures (De Boeck & Wilson, 2004; Woods, 2007). To aid convergence of the two-parameter logistic model (2PL), discriminabilities were estimated using Bayesian methods. A three-parameter model is not employed because all responses to the numeracy items were open-ended and none of the responses are commonly guessed values (e.g., 0, 1, 50, 100).

Within FlexMIRT (Cai, 2013), two separate 2PLs were estimated with a lognormal prior on the discriminability parameter—one for each sample and set of 42 items. The prior distribution was determined by fitting a 1PL (creating a equality constraint across the items' discriminability to estimate the average discriminability; i.e., a's = 1.54 and 1.62), then mean of the lognormal distribution was set to the log(a) (i.e., means = 0.432 and 0.476) and the standard deviation was set to (log(a*2) – log(a/2))/4 (i.e., sd's = 0.347 and 0.347). For the first set of items, ten items demonstrated a lack of fit with the model and were excluded as potential items (G^2 *(df* = 177*)* = 6175.36, $p < .001$, *RMSEA* = .36). For the second set of items, eight items demonstrated a lack of fit with the model and were excluded (G^2 *(df* = 138*)* = 4842.93, $p < .001$, *RMSEA* = .39). Poor item-level goodness of fit suggests that the estimated parameters for a particular item do not accurately capture a plausible data-generating process. This left 66 candidate numeracy items with difficulties ranging from β = -2.69 to β = 4.91 (for all item parameters, see Appendix A).

Measure Construction

Adaptive Numeric Understanding Measure (A-NUM). The adaptive measure was developed from the 66 remaining items and simulated using CatR (Magis et al., 2018). We assumed that

participants' numeracies to be distributed normally about average ability (z = 0). For intermittent ability and final estimates, expected a posteriori estimates were used; it further used the global-discrimination index to select items (Kaplan et al., 2015). The measure began with a fixed item chosen for its average difficulty and high discriminability ($\alpha = 2.17$, $\beta = 0.00$) and ended once the participant had responded to four items. Thirteen unique questions comprise the resulting adaptive measure, from which any one test taker would be asked to respond to four items (constituting seven unique exams). This approach categorizes participants into one of nine categories or 16 unique estimated a posteriori (EAP) scores. Participant ability is calculated as a Z-score based on responses to viewed items and the item parameters (range z = -2.57 and z = 2.21; see Figure 1 for measure structure and Table 1 for the items). A simulation was conducted to test whether the setup of the A-NUM (i.e., a measure using the same item parameters, structure, and length) could accurately assess ability. 5,000 thetas were generated from a standard normal distribution. Measure responses were then simulated for all of the thetas using catR (Magis et al., 2018); then, the theta estimates (using EAP) were compared to the true thetas. Ability estimates based on simulated responses were highly correlated with the randomly generated abilities ($r = .85$, $p < .001$; See Appendix B for further details). There is an alternative scoring method for the A-NUM. Rather than using EAP scores, participants can be sorted into groups as shown in Figure 1. It categorizes people into 9 levels based on the structure of the measure and the pattern of responses (possible scores range from 1 to 9) This method is similar to the scoring used by the adaptive version of the BNT but sorts participants into more categories. A Qualtrics file (Qualtrics Survey Format) for the A-NUM can be downloaded from OSF (https://osf.io/frq2n). This file can be downloaded and imported to any Qualtrics account for easy use of the adaptive scale.

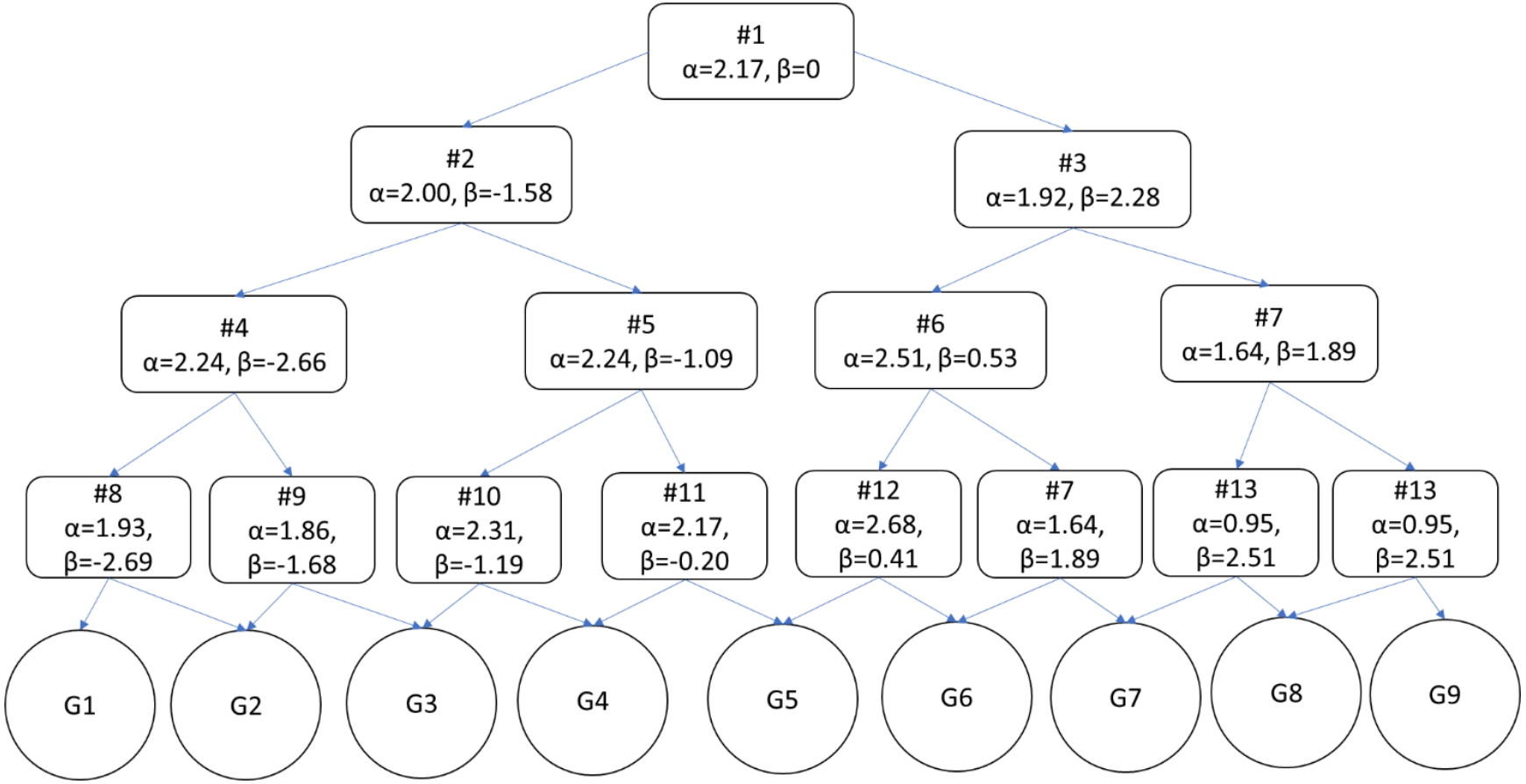

Figure 1. The Structure of the Adaptive Numeric Understanding Measure (A-NUM) with Item Parameters Included (from Study 1; N's=224 and 264)

Table 1. Items that Compose the A-NUM in Figure 1's Order

#	Item	Percentile Difficulty
1	Imagine that you have a five-sided die (the sides of which show 1, 2, 3, 4, 5), and we throw it 150 times. On average, out of these 150 throws how many times would this five-sided die show an odd number (1, 3, 5)? _____ throws	50%
2	A medical study will either give people medicine A or medicine B. Each person has an equal chance to get medicine A or B. If there are 536 people in the study, about how many are expected to get medicine A? ___ people	5.7%
3	If the probability of getting the common cold is 60% in 1 year, what is the probability of getting the common cold in 2 years? _____%	98.9%
4	If Person A's risk of getting a disease is 7% in twenty years, and Person B's risk is double that of A's, what is B's risk of getting the disease in twenty years? ____ % in twenty years	0.4%
5	If the chance of getting a disease is 60 out of 300, this would be the same as having what percent chance of getting the disease? _____ %	13.8%
6	The town of Jamesville has a pole that is red, blue and green standing in the center of town. One-third of a pole is painted red, one-half of the pole is painted blue, and three feet of the pole is painted green. What is the height of the pole? _____ feet	70.2%

7	In a lake 20% of fish are red. A red fish is poisonous with a probability of 20%. A fish that is not red is poisonous with a probability of 15%. What is the probability that a poisonous fish in the lake is red? _____%	97.1%
8	What is 74% of 100 people? _____ people	0.4%
9	If a class of 200 people includes 50 men, this would be the same as the class being what percent men? ____%	4.6%
10	If 70% of basketball players on a college basketball team are over six feet tall and there are 20 players on the team, how many players on the team are shorter than six feet tall? ___ players	11.7%
11	In a field containing 1000 squirrels, 40% of squirrels are striped and a striped squirrel is rabid with a probability of 20%, on average, how many squirrels are there in the field that are rabid and have stripes? ___ squirrels	42.1%
12	Allenton College has a column that is green, white, and yellow (the school's colors) standing in front of the campus library. One-third of the column is painted green, one-half of the column is painted white, and four feet of the column is painted yellow. What is the height of the column? _____ feet	65.9%
13	In a box of cookies, 25% have chocolate chips, 25% have raisins, and 50% are plain. 40% of the chocolate chip and plain cookies aren't fresh, and 30% of the raisin cookies aren't fresh. What percentage of cookies that aren't fresh are raisin cookies? __ %	99.4%

Four-Item Non-adaptive Numeric Understanding Measure (4-NUM). We removed the 13 items used in A-NUM from the pool of items so that A-NUM and 4-NUM did not share any items. The remaining candidate items lacked items with difficulties between $\beta = 0.83$ and $\beta = 1.89$ (approximately the 80th and 97th percentiles, respectively), so previously created and calibrated items modified from BNT items with unpublished answers were also considered. The 4-NUM items chosen from these items were highly discriminating (ranging from $\alpha = 1.62$ to $\alpha = 2.16$) and covered a wide range of ability levels with item difficulties ranging between $\beta = -1.41$ and $\beta = 1.58$ (see Table 2 for the items). Like for the A-NUM, a simulation study was conducted for the 4-NUM to test if the setup of the 4-NUM (i.e., a measure using the same item parameters, structure, and length) could accurately assess ability. Using catIRT (Nydick & Nydick, 2013), 5,000 thetas were generated from a standard normal distribution. Measure responses were then simulated for all of the thetas using catIRT (Nydick & Nydick, 2013) and the theta estimates (using EAP) were compared to the true thetas. Ability estimates based on simulated responses were highly correlated with the randomly generated abilities ($r = .80$, $p < .001$; See Appendix C for simulation details). As an alternative to EAP scores, the 4-NUM can be scored as a sum of correct responses. A Qualtrics file (Qualtrics Survey Format) for the 4-NUM can be downloaded from OSF (https://osf.io/4tk2f). This file can be downloaded and imported to any Qualtrics account for easy use of the adaptive scale.

Table 2. The Items that Compose the 4-NUM in Order of Difficulty (Easiest to Hardest) with Item Parameters (from Study 1, N's=224 and 264)

#	Item	α	β	Percentile Difficulty
1	Suppose that you are buying a gallon of milk at the grocery store. There are two options for the same brand of milk: buying 4 quarts at $2.50 per quart or buying 1 gallon for $8.00. What is the cost per quart (1 gallon=4 quarts) of the better priced milk? $_____ per quart	1.84	-1.41	7.9%
2	Imagine you are throwing a fair six-sided die (the sides of which show 1, 2, 3, 4, 5, 6) 120 times. On average, how many times would you expect this die to show a number less than 5 (1, 2, 3 or 4)? _____ out of 120 throws.	2.16	-0.38	35.2%
3	Out of 300 fruits, 200 are apples and 100 are bananas. Out of the 200 apples, 90 are green. Out of the 100 bananas, 30 are green. What is the probability that a randomly picked green fruit will be an apple? ____ %	1.99	0.52	69.8%
4	In a field 40% of snakes are striped, 30% brown and 30% black. A striped snake is poisonous with a probability of 10%. A snake that is not striped is poisonous with a probability of 20%. What is the probability that a poisonous snake in the field is striped? __ %	1.62	1.58	94.3%

Study 2: Validation

Participants

Based on power analysis using results from Weller and colleagues (2013), 614 MTurk participants were recruited. After our pre-registered data cleaning (https://osf.io/9tjgz), 14 participants were excluded from the analyses (all for indicating that they looked up answers or used a calculator), leaving a final N = 600 (47.17% female; $\bar{x}_{age}$ = 41.19; 75.7% Caucasian, 8.2% Asian, 8.5% African American, and 3.2% Hispanic).

Procedures

After providing consent, participants were randomly assigned to conditions of four decision tasks and completed other measures to test further the predictive, convergent, and divergent validity of the new numeracy measures; random assignment occurred separately for each task. First, they completed either a positively or negatively framed judgment task of students' grades. Then, they rated the attractiveness of either a bet with no loss or a similar bet with a small loss. Participants further completed a probability interpretation task and a benefit perception task. To assess convergent/discriminant validity, participants were then asked to complete a short Big 5 personality measure and the Subjective Numeracy Scale. Next, participants were asked to complete a risk-consistency task by estimating the probability of 4 events occurring in the next year and then, separately, the probability of the same 4 events in the next 5 years.

In a randomized order, participants then completed the adaptive form of the Berlin Numeracy Test (Cokely et al., 2012), a Rasch-Based Numeracy scale (Weller et al., 2013), the new 4-item numeracy measure (4-NUM), and the new adaptive numeracy measure (A-NUM). Following the first completed measure, participants were asked about their experience with the numeracy

assessment up to that point. Following the four numeracy measures, participants completed two non-numeric cognitive measures. Finally, they completed an exploratory mood measure for an unrelated purpose and that is not included in any of the analyses here.

Measures

Numeracy. Numeracy was assessed using four measures. Two established measures were used to test for the convergent validity of our two new measures and to evaluate comparative predictive validity. Moreover, the sets of established and new measures each included one non-adaptive measure (i.e., a measure in which all participants complete the same items) and one adaptive measure (i.e., a measure which modifies the items participants see based on previous responses). An additional, single-item measure (i.e., the 1-NUM) was explored to see how well it would perform. Missing responses were counted as incorrect for all numeracy scales.

Weller. Weller and colleagues' (2013) Rasch-Based Numeracy Scale (Weller) is an 8-item non-adaptive measure assessing numeracy. It was scored as the number of items answered correct. Possible scores range from 0 to 8.

BNT. The Berlin Numeracy Test (BNT) is an adaptive measure assessing numeracy in which each participant responds to 2-3 items (Cokely et al., 2012). Based on the structure of the test and the pattern of responses, participants are categorized into four quartiles. Possible scores range from 1 to 4.

A-NUM. The Adaptive Numeric Understanding Measure (A-NUM) is the new adaptive measure assessing numeracy described above.

4-NUM. The Four-Item Numeric Understanding Measure (4-NUM) is the new non-adaptive measure assessing numeracy described above.

1-NUM. Lastly, the Single-Item Numeric Understanding Measure (1-NUM) is a one-question measure assessing numeracy; it is the first item of the A-NUM. This item has high discriminability ($\alpha = 2.17$; $\lambda = .79$) and is of average difficulty ($\beta = 0$), making it ideal for discriminating between participants above and below average in numeracy. Specifically, the question is, "Imagine that you have a five-sided die (the sides of which show 1, 2, 3, 4, 5), and we throw it 150 times. On average, out of these 150 throws how many times would this five-sided die show an odd number (1, 3, 5)?" It is scored as correct or not so that possible scores are 0 or 1. All participants answered this question as part of the A-NUM.

Negative Subjective Exam Experience. Directly following the first numeracy measure, participants were asked, "How is your experience of answering the math questions so far?" Participants indicated their agreement with 4 statements (i.e., "The questions are tedious"; "The questions are stressful"; "The experience is negative"; "The experience is positive") on a 7-point Likert scale (1=Strongly Disagree to 7=Strongly Agree). Responses were averaged after reverse scoring some items so that higher scores indicated a more negative subjective exam experience.

Decision-Making Tasks to Test Predictive Validity

Framing Task. Modified from Peters et al. (2006), participants were presented with the exam scores and course levels from three courses (200, 300, or 400—indicating varying difficulty levels of classes) of five students and were asked to rate the quality of each student's work on a 7-point scale (-3=very poor to +3 = very good). The frame was manipulated between subjects by presenting the grades as either percent correct or percent incorrect (average percent correct scores for the five students over the three courses were 66.3, 78.3, 79.0, 83.0, and 87.3). For example, "Mike" was described as receiving either 78% correct on his exam or 22% incorrect in a 200-level course.

Bets Task. Based on Peters et al. (2006), a random half of participants rated the attractiveness of a no-loss gamble (7/36 chances to win $9; otherwise, win $0); the other half rated a similar gamble with a small loss (7/36 chances to win $9; otherwise lose 5¢). Participants indicated their preference on a slider scale from 0 (not at all attractive) to 20 (extremely attractive).

Probability Interpretation Task. Based on Cokely et al. (2012), participants indicated their response to a probability interpretation task with multiple-choice options. Participants were asked to select the correct interpretation of a weather forecast. Participants were asked, "Imagine there is a 30% chance of rain tomorrow. Please indicate which of the following alternatives is the most appropriate interpretation of the forecast." They were given 3 options and the correct answer was "It will rain on 30% of the days like tomorrow."

Benefit Perception Task. Based on Cokely et al. (2012), participants indicated their response to a benefit perception task with multiple-choice options. Participants were asked to choose which piece of additional information would most inform about a toothpaste. Participants read an ad for "Zendil" which they were told caused a "50% reduction in occurrence of gum inflammation." Then, they were asked "Which one of the following would best help you evaluate how much a person could benefit from using Zendil?" They were given six options and the correct answer was "The risk of gum inflammation for people who do not use Zendil."

Risk Inconsistency. From Bruine de Bruin et al. (2007), participants were asked to indicate the likelihood of 4 events occurring in the next year and then again in the next 5 years. They responded using a slider from 0 (no chance) to 100 (certainty). Each repeated pair was scored as correct if the probability for the event happening the next year was no larger than for it happening in the next 5 years. Within each time frame, one item is a subset of another (e.g., dying in a terrorist attack is a subset of the superset dying from any cause). To be scored as

correct, the probability of a subset event should not exceed that of its more general event. Therefore, the maximum score is 8 and the minimum is 0.

Convergent Validity Measures

Subjective Numeracy. This 8-item measure from Fagerlin et al. (2007) has two sub-scales: numeric confidence (e.g., "How good are you at working with fractions?") and preference for numeric information (e.g., "How often do you find numerical information to be useful?"). Participants responded on 6-point Likert-type scales. Subjective numeracy was calculated as an average of all items. Numeric self-efficacy was calculated as the average of the first four items and numeric preference as the average of the last four items.

Big 5 Personality Traits. Participants responded to the 30 items of The Big Five Inventory–2 Short Form (BFI-2-S) using a 5-point Likert scale from 1 (Disagree strongly) to 5 (Agree strongly; Soto & John, 2017). Items for each personality trait were averaged together. Associations with Big 5 Personality Measures can be found in the supplement (Appendix G).

Raven's Progressive Matrices. Participants completed 10 six-alternative multiple-choice questions in which they had to complete the pattern in a matrix (Raven, 1989). The number of correct responses was summed.

Vocabulary. Participants took a 12-item vocabulary test. The items were 6-alternative multiple-choice questions with the 6th option being “Skip.” The items were created by Ekstrom and Harmon (1976). The original 36-item test was shortened based on IRT analysis. The number of correct responses was summed.

Results

Descriptive. Of the 600 participants, 149 participants completed the A-NUM (that includes the 1-NUM) first, 150 completed the BNT first, 153 completed the Weller first, and 148 completed the 4-NUM first. An exploratory multivariate ANOVA revealed no significant differences in scores on the five measures by which measure was completed first (F (3, 596) = 0.98, p = 0.47; *Wilk's lambda* = 0.98). The average score on the A-NUM was 4.70 (SD = 1.49; 52%) on the 1-9 scale and -0.14 (SD = 0.95) using expected a posteriori (EAP) estimation. The two scoring methods were highly correlated (r = .99). Thus, further analyses only examined scores calculated using the 1-9 scale. The average score on the 4-NUM was 1.84 (SD = 1.17; 46%) on a 0-4 scale (Guttman's λ6 = .59). The average score on the BNT was a 2.32 (SD = 1.19; 58%) on a 1-4 scale. The average score on the Weller was 5.30 (SD = 1.98; 66%) on a 0-8 scale (Guttman's λ6 = .74). About half (48%) of participants answered the exploratory 1-NUM correctly.

Test of whether all numeracy measures would load onto a single latent variable. This hypothesis was investigated using a robust confirmatory factor analysis (CFA) which specified a model in which the scores from each numeracy measure load onto a single factor. The model resulted in an adequate fit, $\chi^2(2) = 7.97$, p = .02, *CFI* = .99, *TLI* = 0.97, *RMSEA* = .07, *SRMR* = .02. See Figure 2. An exploratory factor analysis also supported a single-factor solution, $\chi^2(2) = 9.08$, p = .01, *RMSEA* =.08, *TLI* = .98. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .83. Bartlett's test of sphericity indicated that the correlation structure is adequate for factor analyses, χ2 (6) = 1329.28, p < .001.

A follow-up exploratory factor analysis was conducted, replacing the A-NUM with the 1-NUM because they share one item; it also supported a single-factor solution, $\chi^2(2) = .06$, p = .97, *RMSEA* < .001, *TLI* = 1. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .82. Bartlett's test of sphericity indicated that correlation structure is

adequate for factor analyses, $\chi 2$ (6) = 1024.76, $p < .001$. Latent numeracy accounted for 45.7% of variance in the 1-NUM. Latent numeracy accounted for 71.8% of variance in the 4-NUM, 60.1% of variance in the BNT, and 60.4% of variance in the Weller. The A-NUM, 4-NUM, and 1-NUM loaded onto a single latent factor with other numeracy measures. Additionally, latent numeracy accounted for a larger proportion of their variance than it did for the BNT and Weller scales, a point to which we return in the discussion.

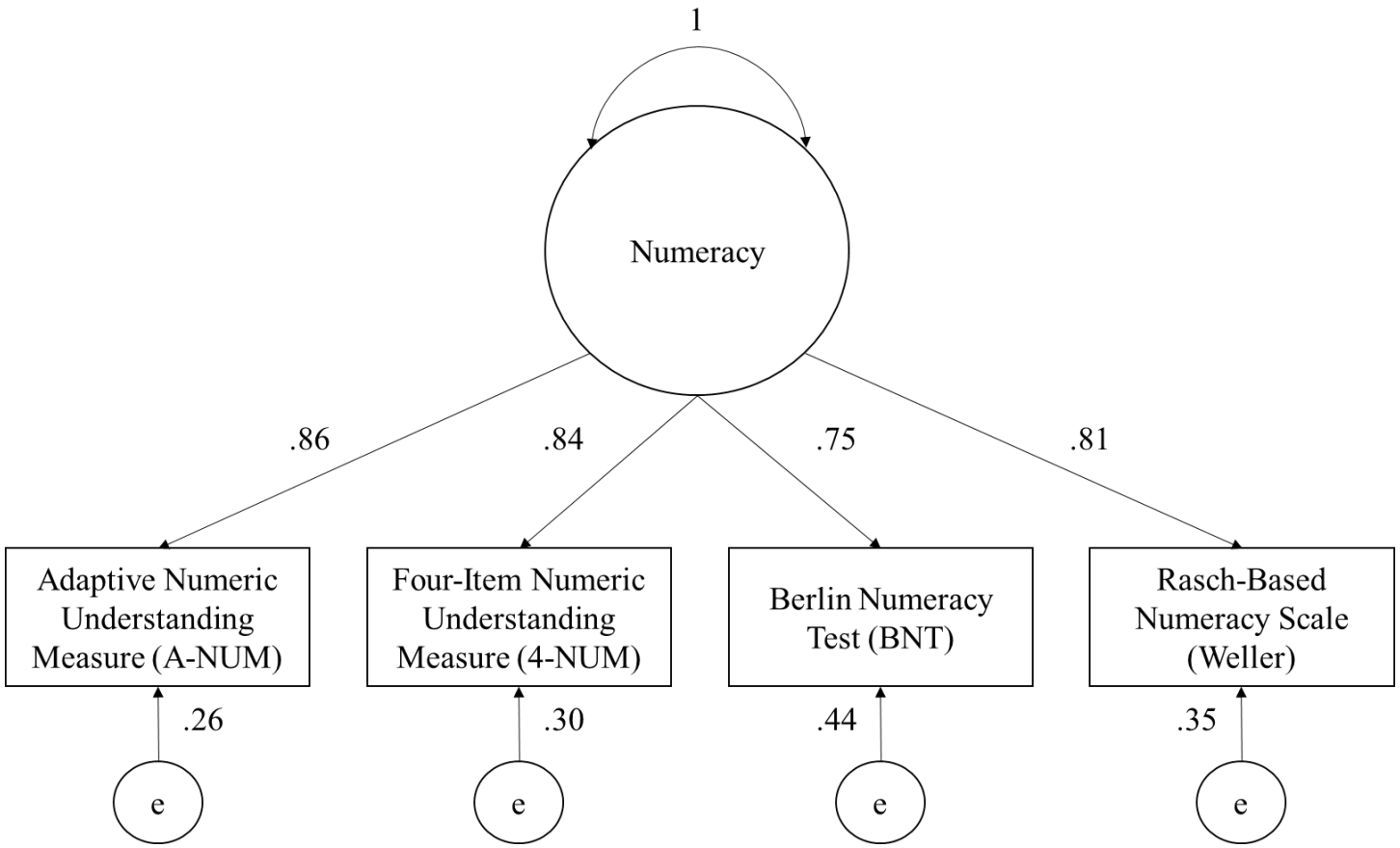

Figure 2. The Results of the Confirmatory Factor Analysis of the A-NUM, 4-NUM, BNT, and Weller Numeracy Measures. Values on the lines from numeracy to the measures represents the loading of each measure onto the latent factor. The values below the measures represents the error variance of the measures (i.e., the variance in scores unexplained by the latent factor).

Note: p-values are < .001 for all paths

Test of whether the new measures would demonstrate convergent validity. To test convergent validity, the new measures (i.e., A-NUM, 4-NUM, and exploratory 1-NUM) were correlated with the two existing measures (i.e., Weller and BNT), subjective numeracy and its subscales (i.e., numeric self-efficacy and numeric preference) and two non-numeric intelligence measures - Raven's progressive matrices and vocabulary (See Table 3 for all bivariate correlations). We expected the new numeracy measures to correlate strongly and positively with the Weller and BNT numeracy measures. Moreover, we expected moderately strong relations between numeracy and subjective numeracy as well as its subscales of numeric self-efficacy and numeric preference. We expected moderate positive associations between numeracy and both vocabulary and Raven's matrices. Overall, we expect the pattern of associations for the new measures to be like those for the other numeracy measures. Indeed, the new numeracy measures strongly correlated with established measures and had moderate to strong relations with subjective numeracy and the two non-numeric intelligence measures.

Table 3. Means, Standard Deviations, and Correlations for all Constructs used to Test Convergent Validity

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
Numeracy measures											
1. A-NUM (Categories)	4.70	1.49									
2. 4-NUM	1.84	1.17	.72**								
3. 1-NUM	0.48	0.50	.75**	.58**							
4. BNT	2.32	1.19	.63**	.66**	.53**						
5. Weller	5.30	1.98	.71**	.66**	.52**	.60**					
6. Subjective Numeracy	4.61	0.96	.44**	.39**	.38**	.34**	.40**				
7. Numeric Self-Efficacy	4.38	1.23	.44**	.42**	.37**	.36**	.40**	.91**			
8. Numeric Preference	4.84	0.96	.31**	.24**	.29**	.23**	.28**	.84**	.53**		
Non-numeric Intelligence measures											
9. Ravens Matrices	5.34	1.70	.38**	.36**	.26**	.30**	.39**	.20**	.15**	.20**	
10. Vocabulary	6.49	2.03	.35**	.29**	.22**	.26**	.31**	.14**	.13**	.11**	.15**

Note. M and *SD* are used to represent mean and standard deviation, respectively. * indicates $p < .05$. ** indicates $p < .01$. A-NUM: Adaptive Numeric Understanding Measure; 4-NUM: Four-Item Numeric Understanding Measure; 1-NUM: Single-Item Numeric Understanding Measure; BNT: Berlin Numeracy Test; Weller: Rasch-Based Numeracy Scale.

Predictive Validity

Test of whether people scoring higher on the NUM and other numeracy measures would have superior probability interpretation than those scoring lower in numeracy. Overall, 51.0% of participants responded correctly to the probability interpretation task. To assess if greater numeracy predicted more correct responses in the probability interpretation task, multiple binary logistic regressions were used to predict correctness from each of the four measures one-at-a-time. As hypothesized, the new numeracy measures and the established measures predicted correct probability interpretation. The strongest predictor of correct interpretation was the 1-NUM, possibly due to the similar difficulty of the numeracy item and the interpretation task (i.e., both nearly split the sample in half; see Table 4 for effect sizes). For full models, see Appendix D (Table S1).

Table 4. Effect Sizes of Focal Numeracy Effects in Predictive Validity Tasks Without and With Adjusting for Assessed Raven's Progressive Matrices and Vocabulary

Task		A-NUM	4-NUM	1-NUM[ǂ]	BNT	Weller
Probability Interpretation	No Covariates[a]	2.18***	1.91***	3.00***	1.88***	2.10***
	With Covariates[a]	2.06***	1.77***	2.56***	1.73***	2.00***
Benefit Perception	No Covariates[a]	1.26*	1.24*	1.28	1.37***	1.37**
	With Covariates[a]	1.24*	1.22	1.20	1.36**	1.38**
Framing Effect	No Covariates[b]	.05	.05	.03	.04	.04
	With Covariates[b]	.05	.06	.03	.04	.04
Bets Task	No Covariates[b]	.09*	.13**	.08*	.09*	.06
	With Covariates[b]	.09*	.12**	.08*	.09*	.06
Risk Consistency	No Covariates[c]	.16***	.19***	.15***	.13***	.18***
	With Covariates[b]	.09*	.14***	.10**	.07	.12**

[a] *Standardized Odds Ratio;* [b] *Partial R of the Focal Effect;* [c] *R. * indicates p < .05. ** indicates p < .01. *** indicates p < .001.*
[ǂ]Analyses using the 1-NUM are exploratory. A-NUM: Adaptive Numeric Understanding Measure; 4-NUM: Four-Item Numeric Understanding Measure; 1-NUM: Single-Item Numeric Understanding Measure; BNT: Berlin Numeracy Test; Weller: Rasch-Based Numeracy Scale.

Test of whether people scoring higher on the NUM and other numeracy measures would have more accurate benefit perceptions than those scoring lower. Overall, 25.7% of participants responded correctly to the benefit perception task. To assess if greater numeracy predicted correct responses for the benefit perception task, multiple binary logistic regressions were used in predicting correctness from each of the 4 measures. The A-NUM, BNT, and Weller predicted correct benefit perceptions similarly. Although the 4-NUM and 1-NUM did not attain significance as a predictor of benefit perceptions after adjusting for assessed vocabulary and Raven's matrices, the effect size of numeracy with covariates was similar to its effect size without covariates (see Table 4 for effect sizes). For full models, see Appendix D (Table S2)

Test of whether the more numerate would demonstrate smaller framing effects. The framing effect was evaluated using multilevel linear regressions predicting participant's ratings of the student from frame condition (positive vs. negative), numeracy as assessed by the four measures, a frame-by-numeracy interaction, and random intercepts for each student being evaluated and each participant. Although the frame significantly influenced judgments, numeracy (as operationalized by any of the measures) did not significantly modify the effect of frame on judgments of the students as indicated by the non-significant Frame × Numeracy interactions (See Table 4 for effect sizes). For full models, see Appendix D (Tables S3 and S4). Appendix E contains exploratory analyses examining the pattern of results across student scores indicating that the numeracy by frame results appeared when the proportions used were more extreme (e.g., 87% vs. 63% correct; Peters et al., 2006). It may be that the use of higher average grades across the stimuli would have resulted in a replication of the original effect.

Test of whether the highly numerate—more than the less numerate—would rate the small-loss condition of the bets task as more attractive than the no-loss condition. The bets task was

evaluated with a simple linear regression predicting attractiveness ratings from the loss condition (no-loss v. small-loss), numeracy as assessed by the 4 measures, and a loss-by-numeracy interaction. In each case where the interaction was significant, the highly numerate rated the loss bet as more attractive than the no-loss bet whereas the less numerate rated them more similarly. The significant measures appear to demonstrate similar effect sizes (see Table 4 for effect sizes). For full models, see Appendix D (Table S5).

Test of whether the highly numerate would show more risk consistency than the less numerate. Risk consistency was evaluated using simple linear regressions predicting the number of risk consistent responses (out of 8) from numeracy as assessed by the 4 measures. The significant measures appear to demonstrate similar effect sizes (see Table 4 for effect sizes). Adjusting for assessed vocabulary and Raven's matrices reduced the effect sizes for numeracy suggesting that these other cognitive abilities may, in part, account for the predictive power of the numeracy measures. For full models, see Appendix D (Table S6).

Overall, established associations between decision making task and numeracy were largely replicated using the A-NUM and mostly replicated using the other numeracy measures. However, numeracy as measured by each of the included assessments did not moderate the framing effect in the present study. Adjusting for non-numeric intelligence measures had little to no influence on the effect sizes of numeracy in predicting the behaviors evaluated in this paper. This pattern of results could suggest that the assessed traits for covariates largely did not account for the effect of numeracy on these behaviors. However, the models with covariates were not corrected to account for imperfect reliability of measures. One exception to the noted pattern was in predictions of risk consistency where numeracy's effect size was reduced when adjusting for

assessed Raven's matrices and vocabulary (see Table 4). Together, these results demonstrate predictive validity for the new numeracy measures.

Negative Subjective Exam Experience. We were also interested in whether participants' subjective exam experiences differed across the measures. More negative subjective experiences were evaluated using a multivariate analysis of variance (MANOVA) predicting the four subjective-experience responses after the first numeracy measure completed. The MANOVA revealed no significant differences overall between numeracy measures for reported negative test experience (overall means were 3.74, 3.48, 3.56, and 3.76, respectively, on a 7-point scale for A-NUM, 4-NUM, BNT, and Weller; F (3, 596) = 1.60, p = 0.08; *Wilk's lambda* = 0.968), indicating that the experience across measures was generally neither positive nor negative.

Table 5. Summary of Measure Metrics

Task	A-NUM	4-NUM	1-NUM	BNT	Weller
Number of items	4	4	1	2-3	8
Loading to Latent Numeracy	.86	.84	.68[ǂ]	.75	.81
Item Difficulty (β) range (as part of population)	-2.69 to 2.51	-1.41 to 1.58	0	0.21 to 1.61*	-1.78 to 1.32
Internal Consistency (Guttman's λ6)	-	.59	-	-	.74
Significant predictor of JDM tasks (with covariates)	4/5	3/5	3/5	3/5	3/5
Negative Subjective Exam Experience	3.74/7	3.48/7	-	3.56/7	3.76/7
Estimated Sample Size for Replication of Predictive Validity Tasks[±]	1,159	1,305	2,619	2,093	2,619

*Item parameters from Allan, 2021; [ǂ]1-NUM loading based on analysis excluding the A-NUM. [±]See Appendix F for details. A-NUM: Adaptive Numeric Understanding Measure; 4-NUM: Four-Item Numeric Understanding Measure; 1-NUM: Single-Item Numeric Understanding Measure; BNT: Berlin Numeracy Test; Weller: Rasch-Based Numeracy Scale.

Discussion

In the present paper, we sought to develop new measures of numeracy—developed using item response theory (IRT)—that provided granular measurement across a meaningful range of ability and with a minimum number of items. Using largely newly-developed math problems, we ultimately produced three numeracy measures that fulfilled our aim: a non-adaptive four-item measure, an adaptive measure (on which participants would respond to four out of 13 items), and a single-item measure. All three measures demonstrated convergent and predictive validity (Campbell & Fiske, 1959; Cronbach & Meehl, 1955). Thus, by using IRT, we were able to develop brief measures of numeracy that captured many levels of difficulty, without sacrificing the scope of the construct. Our new one- and four-item numeracy measures and our new online adaptive numeracy measure all provide researchers with short-form assessments of numeracy using unfamiliar items that address retest-effect issues caused by using the same items for all participants over many studies (these measures can be downloaded for Qualtrics from https://osf.io/pcegz/). Unlike other current numeracy measures, answers for these items also cannot be located easily online at this time.

To address construct validity, we first sought to demonstrate that the new measures assess the same latent trait as established numeracy measures. Indeed, factor analysis suggested one latent trait was responsible for most of the variance in responses to each of the numeracy measures. The A-NUM and 4-NUM appeared to reflect latent numeracy better than the two established measures. One potential reason for this greater shared variance is participants did not have prior experience with the items so that less of the variability in scores can be explained by either prior experience or the ability to find the correct answers online. Alternatively, it is possible that the greater difficulty range in items allowed for more shared variance between the new measures and

latent numeracy. Further, we attempted to demonstrate convergent validity by assessing the associations of the new measures with established numeracy measures and measures of subjective numeracy as well as non-numeric intelligences. Generally, the new measures correlated with other measures as expected. Interestingly, the new single-item measure has the potential to divide samples in half by numeracy level, and even it performed similarly to the longer measures although more weakly than them.

Next, we were able to generally demonstrate predictive validity of our measures using established numeracy tasks. As expected, numeracy measures predicted performance in the probability interpretation and benefit perceptions tasks (Cokely et al., 2012). Numeracy also moderated the relative attractiveness ratings of small-loss and no-loss bets (Peters et al., 2006). Participants higher in numeracy made 1-year and 5-year risk judgments that were more consistent with each other than did those lower in numeracy (Del Missier et al., 2012). New numeracy measures demonstrated similar predictive power to established measures. Results of a framing task did not support numeracy-by-frame interactions for any numeracy measure (Peters et al., 2006; see Appendix E for exploratory analyses indicating that stimuli choices may have played a role). Together, support for these hypotheses suggest construct validity by demonstrating convergent and predictive validity for all three of the Numeric Understanding Measures (NUM; see Table 5 for metrics of each measure).

While all measures provide predictive validity, the most consistent predictor of decision performance across tasks was the A-NUM that placed participants in one of nine categories based on their responses to four out of thirteen new math problems (compared to 4 categories for the adaptive Berlin Numeracy Test). These findings could suggest that its more granular and accurate assessment provided more power to identify effects in studies involving numeracy. The

effect size of A-NUM was largely consistent as a predictor of the effects after adjusting for assessed vocabulary and Raven's matrices. This pattern suggests that the new A-NUM and 4-NUM measures are ideal for assessing the unique covariance of numeracy with behaviors.

Another strength of the A-NUM is the difficulty range it covers. The A-NUM is expected to measure ability in MTurk and similar populations ranging between the 0.5th percentile and the 98.6th percentile (about 98% of participants). By comparison, a participant who gets the minimum score of 1 on the BNT—which is geared toward the more highly numerate—is expected to achieve a score of 3.67 on the A-NUM (using the 1 to 9 categorical scale; a score of -0.81 using the expected a posteriori method). The BNT has a difficulty range falling between $\beta = 0.21$ and $\beta = 1.61$ (Allan, 2021) compared to A-NUM's difficulty range falling between $\beta = -2.69$ and $\beta = 2.51$. The BNT's truncation means little distinction exists among examinees below approximately the 58th percentile in numeracy, based on the calibration by Allan (2021). This limited range is problematic when assessing numeracy in lower ability populations (e.g., older age and less educated populations) and when changes in numeracy at lower parts of the ability spectrum predict changes in outcomes (Desjardins et al., 2013; Peters, Fennema, & Tiede, 2019). Overall, the A-NUM is expected to have the capacity to assess numeracy at lower levels than the BNT.

The present study was limited in several ways. First, this validation used an online convenience sample in the US and thus may not generalize widely. However, American MTurkers do not significantly differ from the US population in global cognitive ability as measured by the 16-item International Cognitive Ability Resource (Merz et al., 2022). Additionally, unlike classical test theory, IRT can produce unbiased estimates of item parameters without a representative sample (Embretson & Reise, 2013). The present study did not assess the length of time to take each

assessment; however, we have minimized the number of items with limited sacrifice of the breadth of numeracy assessed. The present paper only reports internal consistency for the 4-NUM and the Weller. Internal consistency cannot be calculated for the other measures (see Figure S3 for the A-NUM test information curves as an indicator of reliability). However, the new numeracy measures do address reliability; they are expected to reduce learning and memory effects stemming from high usages of numeracy measures in two ways. First, the new NUM measures use new items that are not easily found and are relatively unknown to participants. Second, the adaptive nature of the A-NUM has the potential to reduce learning and memory effects since a participant who correctly answers an item which they had previously gotten wrong is presented an item novel to them. To ensure reliability, highly discriminating items were used whenever possible. The present study does not specifically investigate the dimensionality of numeracy as a construct. However, its results suggest that the included numeracy measures load onto a common factor despite being comprised of different items, thus supporting construct validity.

Although the A-NUM reduces many problems with previous measures (such as those assessed in this study), it does have limitations. First, no items were available in some parts of the difficulty range; the calibrated items and items drawn for the measures were lacking items with difficulties between $\beta = 0.83$ and $\beta = 1.89$ (approximately the 80th and 97^{th} percentiles, respectively; see Appendix A). Thus, the standard error of the measure is larger in this range (See Figure S1 in the supplement). This ability range corresponds approximately to an ability level needed to complete a novel, multi-step numeric problem (Desjardins et al., 2013). Nevertheless, the pattern of responses to items surrounding this missing range provide some information about ability range therein. Future research should develop more items to assess numeracy in this ability range.

Second, the A-NUM prioritized a short test length resulting in a larger error of the estimate when calculating ability using estimated a posteriori (EAP) scores. Rather than using EAP estimation, we have sorted participants into one of nine ability levels which combine similar ability levels from EAP scores. The EAP method would provide a measure of ability based on Z-score and require no other standard for comparison, whereas, sum scores require comparison to a standard (e.g., the average within the population of interests; Embretson & Reise, 2013). However, with a larger error in the estimate, discriminating between close scores may result in inaccurate rank-order estimates. Sorting participants into categories as we have done throughout the paper produces a less granular measure but should allow for more accurate rank-order estimates, address concerns with estimation errors, and provide a simpler way to score the measure. The A-NUM uses highly discriminating items which means that the items are more informative (i.e., they more closely assess latent numeracy). However, the most difficult item of the A-NUM (i.e., item 13; see Table 1) has a relatively lower discriminability. Several possible explanations exist for the lower discriminability. One is that the item simply is not as reflective of latent numeracy. Another possibility that we favor is that the type of problem (i.e., a Bayesian reasoning problem) could be solved in multiple ways: employing logic or employing Bayes' theorem. We suspect that people of different numeracy levels employ different strategies to solve the Bayesian reasoning problem (Pachur & Galesic, 2013). Nevertheless, the discriminability of item 13 is adequate with an equivalent factor loading of $\lambda = .49$ (as discriminability in a 2PL and factor loadings can be calculated from each other directly).

One way that current measures can be further improved is through ongoing creation and calibration of new numeracy items. Items with similar parameters could replace current items or, with sufficient items, the measure could randomly draw from pools of items with similar

difficulties. This approach would go further to address retest effects in numeracy assessment. However, the creation of new items to create such a measure is complicated by the need to empirically investigate the difficulty of items prior to use. In studies with large samples, new uncalibrated numeracy items could be included and calibrated to continually grow the bank of potential numeracy items. Then, promising items could be re-calibrated together and used to improve the A-NUM and create new measures. Many different labs could potentially create and calibrate new items for this purpose using similar methods as those described in the present paper. Future research should seek to calibrate new items to address current information gaps and build up interchangeable items. Translation and validation of these measures in other languages and other countries also are needed. In addition to developing more questions, researchers should examine the effect of different contexts, such as financial and health contexts, on understanding across countries.

Conclusion

Numeracy appears to be an important construct in judgments and decision making as well as in health and financial outcomes and wellbeing (Bjälkebring & Peters, 2021; Peters, 2020; Peters, Tompkins, et al., 2019). Moreover, the study of numeracy has led to greater understanding of motivated reasoning (Kahan et al., 2017; Shoots-Reinhard et al., 2021). However, current measures of numeracy are flawed. Many measures assess truncated levels of numeric ability (e.g., Numeracy Scale; Lipkus, Samsa, & Rimer, 2001; Berlin Numeracy Test; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) whereas others are comprised of well-known items that might mask participants' true ability (e.g., Cognitive Reflection Task; Frederick, 2005; Rasch-Based Numeracy Scale; Weller et al., 2013).

In the current research, we developed and validated three novel measures of numeracy: an adaptive measure in which participants respond to four out of thirteen items (A-NUM), a four-item non-adaptive (4-NUM) measure, and a single-item non-adaptive (1-NUM) measure. Both the A-NUM and 4-NUM demonstrated convergent and predictive validity, suggesting they have good construct validity (Cronbach & Meehl, 1955). Thus, using IRT, we provided short measures of numeracy without sacrificing construct scope (Smith et al., 2000; Smith & McCarthy, 1996). Moreover, both four-item measures provided researchers with short-form assessments of numeracy using novel items to address retest effects caused by using the same items for all participants over many studies. Lastly, even the single-item numeracy measure (1-NUM) measured numeracy adequately and could be useful when time is limited.

Acknowledgement: A special thank you to Martin Tusler for his work in developing new numeracy items.

Funding: This work was supported by the National Science Foundation (2017651) and the Decision Sciences Collaborative at The Ohio State University.

References

Allan, J. N. (2021). Statistical numeracy norms and decision vulnerability benchmarks: A norm-referenced method for estimating the risk literacy difficulty level of choices and risk communications. https://shareok.org/handle/11244/330238

An, X., & Yung, Y. F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS Institute Inc. SAS364-2014*, *10*(4).

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence, 62*, 89–98. https://doi.org/10.1016/j.intell.2017.03.003

Best, R., Carman, K. G., Parker, A. M., & Peters, E. (2022). Age declines in numeracy: An analysis of longitudinal data. *Psychology and Aging, 37*(3), 298–306. https://doi.org/10.1037/pag0000657

Bjälkebring, P., & Peters, E. (2021). Money matters (especially if you are good at math): Numeracy, verbal intelligence, education, and income in satisfaction judgments. *PLOS ONE, 16*(11), e0259331. https://doi.org/10.1371/journal.pone.0259331

Brooks, M. E., & Pui, S. Y. (2010). Are individual differences in numeracy unique from general mental ability? A closer look at a common measure of numeracy. *Individual Differences Research, 8*(4), 257–265.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology, 19*(3), 273–293. https://doi.org/10.1207/S15326942DN1903_3

Burns, W. J., Peters, E., & Slovic, P. (2012). Risk perception and the economic crisis: A longitudinal study of the trajectory of perceived risk. *Risk Analysis: An International Journal, 32*(4), 659–677.

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Castel, A. D. (2007). Aging and memory for numerical information: The role of specificity and expertise in associative memory. *The Journals of Gerontology: Series B, 62*(3), P194–P196. https://doi.org/10.1093/geronb/62.3.P194

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Cavanaugh, K., Huizinga, M. M., Wallston, K. A., Gebretsadik, T., Shintani, A., Davis, D., Gregory, R. P., Fuchs, L., Malone, R., & Cherrington, A. (2008). Association of numeracy and diabetes control. *Annals of Internal Medicine, 148*(10), 737–746.

Chesney, D. L., Shoots-Reinhard, B., & Peters, E. (2021). The Causal Impact of Objective Numeracy on Judgments: Improving Numeracy via Symbolic and Non-Symbolic Approximate Arithmetic Training Yields More Consistent Risk Judgments. *Journal of Numerical Cognition, 7*(3), 351–367. https://doi.org/10.5964/jnc.6925

Choma, B. L., Sumantry, D., & Hanoch, Y. (2019). Right-wing ideology and numeracy: A perception of greater ability, but poorer performance. *Judgment and Decision Making*, *14*(4), 412–422. https://doi.org/10.1017/S1930297500006100

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. Judgment and Decision Making, 7(1), 25–47.

Coyle, T. R., & Greiff, S. (2021). The future of intelligence: The role of specific abilities. *Intelligence, 88*, Article 101549. https://doi.org/10.1016/j.intell.2021.101549

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

De Boeck, P., & Wilson, M. (2004). Explanatory item response models: A generalized linear and nonlinear approach (Vol. 10). Springer.

Dehaene, S. (2011). The number sense: How the mind creates mathematics. OUP USA.

Del Missier, F., Mäntylä, T., & de Bruin, W. B. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25*(4), 331–351. https://doi.org/10.1002/bdm.731

Desjardins, R., Thorn, W., Schleicher, A., Quintini, G., Pellizzari, M., Kis, V., & Chung, J. E. (2013). OECD skills outlook 2013: First results from the survey of adult skills.

Dieckmann, N. F., Slovic, P., & Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. *Risk Analysis: An International Journal, 29*(10), 1473–1488.

Dieckmann, N., Peters, E., Leon, J., Benavides, M., Baker, D., & Norris, A. (2015). The role of objective numeracy and fluid intelligence in sex-related protective behaviors. *Current HIV Research, 13*(5), 337–346.

Ekstrom, R. B., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672–680.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Hart, P. S. (2013). The role of numeracy in moderating the influence of statistics in climate change messages. *Public Understanding of Science, 22*(7), 785–798.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy, 1*(1), 54–86.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Kovacs, K., & Conway, A. R. A. (2019). What Is IQ? Life Beyond "General Intelligence." *Current Directions in Psychological Science*, *28*(2), 189–94. https://doi.org/10.1177/0963721419827275

Legree, P. J., Fischl, M. A., Gade, P. A., & Wilson, M. (1998). Testing word knowledge by telephone to estimate general cognitive aptitude using an adaptive test. *Intelligence, 26*(2), 91–98. https://doi.org/10.1016/S0160-2896(99)80056-7

Magis, D., Raiche, G., & Magis, M. D. (2018). Package 'catR.' R Package, Version, 3.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1-11.

McAuliffe, T. L., DiFranceisco, W., & Reed, B. R. (2010). Low numeracy predicts reduced accuracy of retrospective reports of frequency of sexual behavior. *AIDS and Behavior, 14*(6), 1320–1329. https://doi.org/10.1007/s10461-010-9761-5

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Merz, Z. C., Lace, J. W., & Eisenstein, A. M. (2022). Examining broad intellectual abilities obtained within an mTurk internet sample. *Current Psychology, 41*(4), 2241–2249. https://doi.org/10.1007/s12144-020-00741-0

Miron-Shatz, T., Hanoch, Y., Doniger, G. M., Omer, Z. B., & Ozanne, E. M. (2014). Subjective but not objective numeracy influences willingness to pay for BRCA1/2 genetic testing. *Judgment and Decision Making*, *9*(2), 152–158. https://doi.org/10.1017/S1930297500005519

Nydick, S. W., & Nydick, M. S. W. (2013). Package 'catIrt.'R Package, Version, 3.

Pachur, T., & Galesic, M. (2013). Strategy selection in risky choice: The impact of numeracy, affect, and cross-cultural differences. *Journal of Behavioral Decision Making, 26*(3), 260–271. https://doi.org/10.1002/bdm.1757

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science, 21*(1), 31–35.

Peters, E. (2020). Innumeracy in the wild: Misunderstanding and misusing numbers. Oxford University Press.

Peters, E., & Bjälkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802–822.

Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, *3*(6), 435-448.

Peters, E., Baker, D. P., Dieckmann, N. F., Leon, J., & Collins, J. (2010). Explaining the effect of education on health: A field study in Ghana. *Psychological Science, 21*(10), 1369–1376. https://doi.org/10.1177/0956797610381506

Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C. K., Slovic, P., & Hibbard, J. H. (2009). Bringing meaning to numbers: The impact of evaluative categories on decisions. *Journal of Experimental Psychology: Applied, 15*(3), 213–227. https://doi.org/10.1037/a0016978

Peters, E., Fennema, M. g., & Tiede, K. E. (2019). The loss-bet paradox: Actuaries, accountants, and other numerate people rate numerically inferior gambles as superior. *Journal of Behavioral Decision Making, 32*(1), 15–29. https://doi.org/10.1002/bdm.2085

Peters, E., Shoots-Reinhard, B., Tompkins, M. K., Schley, D., Meilleur, L., Sinayev, A., Tusler, M., Wagner, L., & Crocker, J. (2017). Improving numeracy through values affirmation enhances decision and STEM outcomes. *PloS One, 12*(7), e0180674.

Peters, E., Tompkins, M. K., Knoll, M. A. Z., Ardoin, S. P., Shoots-Reinhard, B., & Meara, A. S. (2019). Despite high objective numeracy, lower numeric confidence relates to worse financial and medical outcomes. *Proceedings of the National Academy of Sciences, 116*(39), 19386–19391. https://doi.org/10.1073/pnas.1903126116

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407–413.

Raven, J. C. (1998). Raven's progressive matrices and vocabulary scales. Oxford Pyschologists Press.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327

Shoots-Reinhard, B., Goodwin, R., Bjälkebring, P., Markowitz, D. M., Silverstein, M. C., & Peters, E. (2021). Ability-related political polarization in the COVID-19 pandemic. *Intelligence, 88*, Article 101580. https://doi.org/10/gmjnbh

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. Calculation in decision making. *Frontiers in Psychology, 6*, 1-16. https://doi.org/10.3389/fpsyg.2015.00532

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Smith, P., & McCarthy, G. (1996). The development of a semi-structured interview to investigate the attachment-related experiences of adults with learning disabilities. *British Journal of Learning Disabilities, 24*(4), 154–160.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Steen, L. A. (Ed.). (1990). On the shoulders of giants: New approaches to numeracy. National Academy Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147–168.

Valerius, S., & Sparfeldt, J. R. (2014). Consistent g- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence, 44*, 120–133. https://doi.org/10.1016/j.intell.2014.04.003

Västfjäll, D., Slovic, P., Burns, W. J., Erlandsson, A., Koppel, L., Asutay, E., & Tinghög, G. (2016). The arithmetic of emotion: Integration of incidental and integral affect in judgments and decisions. *Frontiers in Psychology, 7*, 325.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making, 26*(2), 198–212.

Woloshin, S., Schwartz, L. M., Moncur, M., Gabriel, S., & Tosteson, A. N. A. (2001). Assessing values for health: Numeracy matters. *Medical Decision Making, 21*(5), 382–390. https://doi.org/10.1177/0272989X0102100505

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement, 67*(1), 73–87. https://doi.org/10.1177/0013164406288163

Zikmund-Fisher, B. J., Exe, N. L., & Witteman, H. O. (2014). Numeracy and literacy independently predict patients' ability to identify out-of-range test results. *Journal of Medical Internet Research, 16*(8), e187.