Chapter 3
How to Measure Risk Comprehension in Educated Samples*

Edward T. Cokely, Saima Ghazal, Mirta Galesic, Rocio Garcia-Retamero, and Eric Schulz

Abstract The Berlin Numeracy Test is a psychometrically sound instrument designed to quickly assess statistical numeracy and risk comprehension in educated samples (e.g., college students or medical and business professionals). The test is available in multiple languages and formats including an online adaptive test that automatically scores data (http://www.riskliteracy.org). In this chapter, we review results of a validation study (n = 300) documenting convergent (e.g., cognitive

*In this chapter, we partially reproduced the article 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*, 25–47.

E.T. Cokely, Ph.D. (✉)
Department of Cognitive and Learning Sciences, Michigan Technological University, 212 Meese Center, Houghton, MI, USA

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, Berlin 14195, Germany
e-mail: ecokely@mtu.edu

S. Ghazal
Department of Cognitive and Learning Sciences, Michigan Technological University, 212 Meese Center, Houghton, MI, USA
e-mail: sghazal@mtu.edu

M. Galesic, Ph.D.
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, Berlin 14195, Germany
e-mail: galesic@mpib-berlin.mpg.de

R. Garcia-Retamero, Ph.D.
Departamento de Psicología Experimental, Facultad de Psicología, University of Granada, Campus Universitario de Cartuja s/n, Granada 18071, Spain

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, Berlin 14195, Germany
e-mail: rretamer@ugr.es

R. Garcia-Retamero and M. Galesic (eds.), *Transparent Communication of Health Risks: Overcoming Cultural Differences*, DOI 10.1007/978-1-4614-4358-2_3,

ability, numeracy), discriminant (e.g., personality, life satisfaction), and predictive validity (e.g., numerical and non-numerical risky choices). The Berlin Numeracy Test was found to be the strongest predictor of a battery of everyday risky decisions (e.g., evaluating claims about medical treatments, consumer goods, and interpreting forecasts), providing more than twice the predictive power of other numeracy instruments. The Berlin Numeracy Test also accounted for unique variance beyond other related cognitive tests (e.g., cognitive reflection, working memory, and intelligence). Twenty additional validation studies (n = 5,036) indicated that the Berlin Numeracy Test maintained psychometric discriminability across 15 countries (e.g., China, England, Germany, Japan, India, Pakistan, Spain, Sweden, and the USA) and various samples (i.e., community samples, Mechanical Turk web panels, medical professionals). Discussion centers on construct validity and the benefits and limits of adaptive testing.

3.1 Introduction and Background

Efforts to measure individual differences in statistical numeracy come primarily in three forms. Some research examines risky decisions in relation to individual differences in overall educational attainment, cognitive abilities, or cognitive styles (Frederick 2005; Stanovich and West 2000, 2008). Other research primarily focusing on clinical and health domains has developed a valid subjective instrument for self-reported estimations of statistical numeracy (Zikmund-Fisher et al. 2007). Most common, however, is the use of direct performance measures of numeracy—i.e., psychometric tests (for a list of tests see Reyna et al. 2009; see also Black et al. 1995; Galesic and Garcia-Retamero 2010; Lipkus et al. 2001; Peters et al. 2006; Schwartz et al. 1997; Weller et al. 2012).

In this chapter, we describe the most widely used statistical numeracy instruments (Lipkus et al. 2001; Schwartz et al. 1997; see also Chaps. 2 and 15), examining their successes and psychometric limits. We then introduce a new test of statistical numeracy for risk literacy: the Berlin Numeracy Test.[1] This test can be

[1] The Berlin Numeracy Test is named to reflect the international, interdisciplinary development effort initiated in 2007 at Center for Adaptive Behavior and Cognition in the Max Planck Institute for Human Development. For additional discussion and similar public outreach efforts concerning expertise, ethics, and philosophical judgment see philosophicalcharacter.org (Feltz and Cokely 2009, 2012; Schulz et al. 2011).

E. Schulz
Cognitive, Perceptual, and Brain Sciences Department, University College London, 26 Bedford Way, London WC1H 0AP, UK

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany
e-mail: hanshalbe@googlemail.com

used in multiple formats (i.e., computer-adaptive, paper-and-pencil, single-item median-split, multiple-choice) and provides a fast, valid, and reliable tool for research, assessment, and public outreach. We show that the new test offers unique predictive validity for everyday risky decisions beyond other cognitive ability (e.g., cognitive reflection, working memory span, and fluid intelligence) and numeracy tests. Further, we show that the Berlin Numeracy Test dramatically improves psychometric discriminability among highly educated individuals (e.g., college students, graduates, and medical professionals) and across diverse cultures and different languages. We close the chapter with a discussion of implications of the current results for construct validity as well as discussion of the merits of fast and accurate measurement of numeracy (e.g., custom-tailored interactive risk communication).

3.2 Numeracy in Educated Samples

In 2001, Lipkus et al. published the numeracy test for highly educated samples, which was an extension of previous work by Schwartz et al. (1997). Lipkus et al. (2001) conducted a series of four studies ($n = 463$) on community samples of well-educated adult participants (at least 40 years of age) in North Carolina. Among other tasks, all participants answered 11 numeracy questions including (a) one practice question, (b) three numeracy questions taken from the work of Schwartz et al. (1997), and (c) seven other questions (one of which had two parts) that were framed in the health domain (e.g., if the chance of getting a disease is 10% how many people would be expected to get the disease: (a) Out of 100, (b) Out of 1,000; see also Chaps. 2 and 15). Two questions had multiple-choice options while all others were open-ended. All questions were scored (0 or 1) with data aggregated across several studies and entered into a factor analysis. The analysis showed that a one factor solution was appropriate. Overall, results indicated that the refined test of Lipkus et al. (2001) was a reliable and internally consistent measure of western high-school and college educated individuals' statistical numeracy.

The results of Lipkus et al. (2001) were interesting for a number of reasons. First, they provided additional evidence that even among highly educated US community samples some sizable proportion of individuals was likely to be statistically innumerate (e.g., 20% failed simple questions dealing with risk magnitude). Such findings were and continue to be important as many efforts designed to support informed and shared decision-making rest on an erroneous assumption that decision-makers are numerate (or at least sufficiently statistically numerate, see Chap. 13; see also Guadagnoli and Ward 1998 and Schwartz et al. 1997). Second, results indicated that domain framing (e.g., medical vs. financial vs. abstract gambles) did not necessarily differentially affect test performance or comprehension. This finding suggests that various domain-specific items (e.g., items framed in terms of financial or medical or gambling risks) can provide a reasonable basis for the assessment of general statistical numeracy skills that can transfer across domains. Overall, for nearly a decade, the Lipkus et al. (2001) test, and its predecessor from Schwartz

et al. (1997) have provided relatively short, reliable, and valuable instruments that have been used in more than 100 studies on topics such as medical decision making, shared decision making, trust, patient education, sexual behavior, stock evaluations, credit-card usage, graphical communication, and insurance decisions, among many others (see Lipkus and Peters 2009).

3.3 Psychometric Limits of Previous Measures of Numeracy

Despite its many successes and its influential role in advancing risky decision research, as anticipated by Lipkus et al. (2001), a growing body of data suggests some ways that the current numeracy instrument could be improved (for an item response theory based analysis see Schapira et al. 2009; see also Weller et al. 2012). For example, one major concern is that the Lipkus et al. (2001) test is not hard enough to adequately differentiate among the higher-performing, highly educated individuals who are often studied (e.g., convenience samples from major research universities). To illustrate, in one study of college students at Florida State University (a public research university in the USA), data indicated that the Lipkus et al. (2001) test was a significant predictor of risky decisions. The test, however, showed extensive negative skew with scores approaching the measurement ceiling (e.g., most participants answered more than 80% of items correctly, see Cokely and Kelley 2009; for similar results see also Peters et al. 2006, 2007a, 2008, and Schapira et al. 2009; for similar findings in physicians-in-training see Hanoch et al. 2010). Another recent study by Galesic and Garcia-Retamero (2010) using large probabilistic national samples of the whole populations of two countries (i.e., the USA and Germany) revealed negative skew in numeracy scores even among participants from the general population (see also Chap. 2).

A second psychometric concern is that there is relatively little known about the relations between either the Lipkus et al. (2001) or Schwartz et al. (1997) numeracy test and other individual differences, such as basic cognitive abilities (Liberali et al. 2012). To illustrate, one might argue that statistical numeracy is a useful predictor of risky choice simply because it serves as a proxy for fluid intelligence. It is well known that tests of general intelligence, particularly those designed to measure fluid intelligence, are valid and reliable predictors of a wide variety of socially desirable cognitive, behavioral, occupational, and health-related outcomes (Neisser et al. 1996).[2] Fluid intelligence tests such as Raven's Standard or Advanced Progressive Matrixes tend to be more time consuming yet also confer considerable benefits in terms of psychometric rigor and cross-cultural fairness (Raven 2000). To date,

[2] The underlying cognitive mechanisms that give rise to these effects are debated and remain unclear (Cokely et al. 2006; Ericsson et al. 2007; Fox et al. 2009; Neisser et al. 1996).

however, there are few tests that have investigated the extent to which the Lipkus et al. (2001) or Schwartz et al. (1997) instruments provide unique predictive power beyond other cognitive ability instruments either within or across cultures (see Chaps. 2, 9 and 11; see also Cokely and Kelley 2009; Galesic and Garcia-Retamero 2010; Garcia-Retamero and Galesic 2010a, 2010b; Liberali et al. 2012; Okan et al. 2012).

A third psychometric concern is that even if numeracy is compared with other abilities, the observed measurement skew and ceiling effects will complicate comparative evaluations (e.g., intelligence vs. statistical numeracy). Consider a recent study designed to investigate the extent to which each of several individual differences (e.g., executive functioning, cognitive impulsivity, and numeracy) influenced decision-making competence (Del Missier et al. 2010, 2012). The study found that numeracy was less related to some decision-making competencies as compared to measures of executive functioning or cognitive impulsivity, measured by the cognitive reflection test (Frederick 2005). However, it is possible that, at least in part, some negative skew in numeracy scores among the college student sample could have limited differentiation of those individuals with the highest levels of numeracy. In contrast, both executive functioning and the cognitive reflection tests are known to prove discrimination even among highly educated individuals. To be clear, our reading of the individual differences study by Del Missier et al. (2012) is that it represents precise and careful research using many of the best available methods and tools. However, the potential psychometric limits inherent in the now 10-year-old numeracy test leave open important questions. To the extent that a numeracy instrument does not adequately or accurately estimate variation in the sub-populations of interest it is not an efficient basis for theory development or policy evaluations.

3.4 Development and Validation of the Berlin Numeracy Test

Building on the work of Lipkus et al. (2001) and Schwartz et al. (1997), we endeavored to develop a new psychometrically sound statistical numeracy test that could be used with highly educated, high-ability samples. Here, our goal was not to develop a high-fidelity comprehensive test of statistical numeracy or of its subskills. Rather, the goal was to develop a brief, valid, and easy-to-use instrument, with improved discriminability. The development of the Berlin Numeracy Test began with pre-testing on a pool of items including all items from both the Lipkus et al. (2001) and Schwartz et al. (1997) tests along with other items that were internally generated. Following a protocol analysis in which participants solved all numeracy items while thinking aloud (see also Fox et al. 2011), we analyzed responses and selected 28 candidate questions for inclusion in the next stage of test development (i.e., 12 original items plus 16 new items).

3.4.1 Participants

We tested a community sample of 300 participants (57% women) from Berlin, Germany at the Max Planck Institute for Human Development. Participants were primarily current or former undergraduate or graduate students from the Humboldt, Free, and Technical Universities of Berlin. The mean participant age was approximately 26 years old (i.e., 25.9, SD = 4.0; range = 18–44). Each participant completed about 6 hours of testing over the course of 2–3 weeks in exchange for 40€ (ca. $55).

3.4.2 Stimuli and Procedure

A number of different instruments were used to provide convergent, discriminant, and predictive validity for the Berlin Numeracy Test. All comparative instruments are listed and described in Table 3.1. Participants were tested in three separate phases. In phase 1, all participants were tested individually via computer and/or with the assistance of a laboratory technician as required by the particular instrument. The first testing session lasted for approximately 2 hours and consisted primarily of cognitive ability instruments and cognitive performance tasks, including assessment of all candidate numeracy items. During this session calculators were not allowed; however, participants were provided with paper and pens/pencils for notes. In phase 2, participants completed an online assessment from their home including a variety of self-report personality and other survey instruments. All participants agreed to complete the online portion of the study in one session in which they sat alone, in a quiet room. In phase 3, participants returned about 2 weeks after phase 1 and completed another 2 hours of testing. All participants were again tested individually via computer and/or with the assistance of a laboratory technician as required by the particular instrument/task. The final 2 hours of testing involved new cognitive performance tasks including a battery of everyday risky decision-making questions that served as a means of assessing predictive validity.

3.4.3 Test Construction and Test Items

Our goal was to create a brief test that would score each participant on a 1–4 point interval scale corresponding to that participant's quartile rank relative to other highly educated individuals (i.e., higher scores are associated with higher quartiles). Performance quartiles for all participants were assessed according to performance on all 28 candidate statistical numeracy questions. A subset of five questions with a four-level tree structure was identified using the decision tree (i.e., categorization tree) application from the predictive modeling and forecasting software DTREG (Sherrod 2003). The tree structure was constructed such that participants arriving at each branch of the tree had approximately a 50% probability of answering correctly/

Table 3.1 Descriptions and references for tests used to establish psychometric validity of the Berlin Numeracy Test

Measure	Description	Reference
Fluid intelligence (RAPM)	Short form Raven's Advanced Progressive Matrices—a 12 item test of fluid intelligence	Bors and Stokes (1998)
Cognitive reflection (CRT)	The Cognitive Reflection Test uses 3 math questions to assess cognitive impulsivity	Frederick (2005)
Crystallized intelligence (vocabulary)	A 37 item "spot-a-word" German vocabulary test	Lindenberger et al. (1993)
Working memory capacity (span)	A multi-item performance measure of one's ability to control attention when simultaneously solving math operations and remember words	Turner and Engle (1989)
Understanding everyday risks	A multi-item test of one's understanding of information about consumer products, medical treatments, and weather forecasts	Cokely et al. (2012)
Maximizing–satisficing	A 13 item scale measuring one's tendency to maximize vs. satisfice during decision making	Schwartz et al. (2002)
Persistence	The Grit-S is an 8 item brief measure designed to assess persistence in the face of adversity	Duckworth et al. (2011)
Achievement motivation	The AMS-R is a 10 item trait assessment of one's general achievement motivation (e.g., one's desire to achieve good grades or performance evaluations)	Lang and Fries (2006)
Self-efficacy	A10 item self-report measure of one's general sense of self-efficacy	Schwarzer and Jerusalem (1995)
Personality	A 10 item assessment of the Big Five personality traits	Gosling et al. (2003)
Test anxiety	The TAI-G is a 20 item assessment of test-taking anxiety	Hodapp and Benson (1997)
Implicit theories	A 4 item measurement of the extent to which one believes that intelligence is stable vs. changeable	Blackwell et al. (2007)
Satisfaction with life	A 5 item instrument measuring self-reported levels of one's satisfaction with life	Diener et al. (1985)

incorrectly. The test's tree structure was subjected to cross-validation and showed less than 10% misclassification.[3] Subsequent analyses indicated that reducing the four-level solution to a simpler three-level solution (i.e., removing one problem) did not affect test classification performance or validity yet reduced test-taking time

[3] Although some misclassification is unavoidable, the algorithm rarely misclassified a participant by more than one quartile. The assessment is similar to an item response theory analysis, in that it identifies items with maximal discriminability across the range of item difficulty, with a guessing parameter of zero.

(i.e., 10% reduction), increased test format flexibility (i.e., simplified the paper-and-pencil format), and provided improved discriminability among new samples (see Sect. 3.6). All final Berlin Numeracy Test formats are based on the four questions used for the optimal three-level categorization tree as follows (see also Chap. 15):

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in a choir 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. _____ (*correct answer*: 25%)

2a. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? ____ out of 50 throws (*correct answer*: 30)

2b. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? _____out of 70 throws (*correct answer*: 20)

3. In a forest 20% of mushrooms are red, 50% brown, and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? _____ (*correct answer*: 50%)

3.4.4 Test Formats and Scoring

Different research environments have different constraints on factors such as computer-access, group-testing options, data-security requirements, etc. Accordingly, we designed the test to be flexible by offering multiple formats.

3.4.4.1 Computer-Adaptive Test Format

In this format, 2–3 questions (of 4 possible questions) are asked to participants. Questions are adaptively selected based on participants' past success in answering previous questions using an adaptive scoring algorithm (see Fig. 3.1 for test structure). The adaptive structure means that all questions have about a 50% probability of being answered correctly with subsequent questions adjusted on the basis of participants' prior answers. If an answer is correct/incorrect then a harder/easier question is automatically provided that again has a 50% probability of being right/wrong. A participant's skill-level can then be determined from answers to only 2–3 questions in roughly half the time normally required for the Lipkus et al. (2001) numeracy test (less than 3 min; see Table 3.2). To facilitate access, the computer-adaptive Berlin Numeracy Test is available online in a format that automatic scores participants' responses and reports data to researchers in terms of estimated partici-

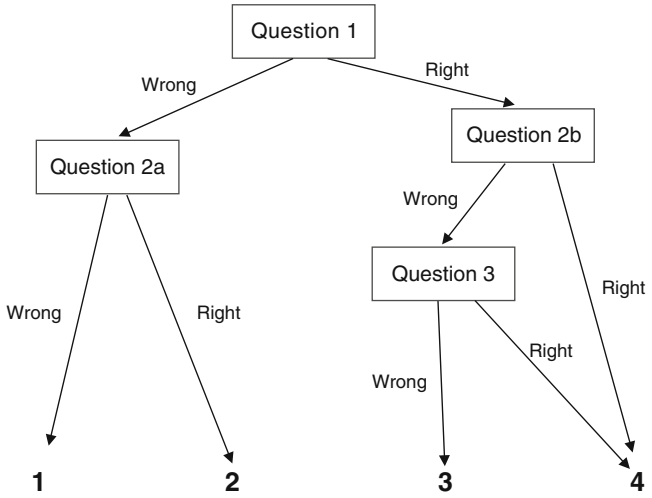

Fig. 3.1 The structure of the computer-adaptive Berlin Numeracy Test. Each question has a 50% probability of being right/wrong. If a question is answered right/wrong a harder/easier question is provided that again has a 50% probability of being right/wrong

pant quartile scores. This version of the test can also be accessed via internet ready hand-held devices (e.g., smart phones) for work in clinics or in the field. The online forum provides an option for the public to complete the test and receive feedback on their performance along with information about potential challenges they may face when making risky decisions. The test can be accessed at the following internet address: http://www.riskliteracy.org. Before completing any test items, the portal seamlessly redirects participants to a secure online location. Online data collection is managed and hosted via the unipark survey software system designed for academic research (unipark.de). We recommend that researchers use the computer-adaptive Berlin Numeracy Test whenever possible as this format provides an efficient balance between speed and psychometric accuracy, and allows us to continue to collect data to further refine the test.

3.4.4.2 Traditional (Paper-and-Pencil) Format

The alternative, traditional format requires that participants answer all four questions of the Berlin Numeracy Test in sequence. Scoring involves totaling all correct answers (i.e., 0–4 points possible). In this format, the structure of the adaptive test is ignored, although the adaptive scoring algorithm can be applied following data collection as might be useful for comparison with other samples. This alternative standard format may be useful when computerized testing is impractical (e.g., group

Table 3.2 Psychometric properties of the numeracy tests: Basic attributes, reliability, and discriminability

			Berlin Numeracy Test (Cokely et al. 2012)		
	Schwartz et al. (1997) 3 items	Lipkus et al. (2001) 11 items	Computer-adaptive test format	Paper-and-pencil format	Single-item format
Basic attributes					
Range of possible scores	0–3	0–11	1–4	0–4	0–1
Range of achieved scores	0–3	5–11	1–4	0–4	0–1
Average score					
Mean	2.4	9.7	2.6	1.6	0.52
Median	3	10	3	2	1
Standard deviation	0.82	1.38	1.13	1.21	0.50
Length					
Number of items	3	11	2–3	4	1
Mean duration in minutes	1.2	4.5	2.6	4.3	1.1
Reliability					
Cronbach's alpha	0.52	0.54	–[a]	0.59	–[a]
Discriminability					
Item % correct (mean)	0.82	0.89	–[b]	0.41	0.52
Mean score of					
1st quartile	0.8	7.3	1.0	0.0	0.0
2nd quartile	2.0	9.0	2.0	1.0	
3rd quartile	3.0	10.0	3.0	2.0	1.0
4th quartile	3.0	11.0	4.0	3.3	

[a] Cronbach's alpha cannot be computed
[b] Approximately 50%, conditional on previous responses

testing, limited computer access). Testing requires about as long as the original Lipkus et al. (2001) numeracy test (i.e., less than 5 min).

3.4.4.3 Single-Item (Median) Format

When time is extremely limited, it is possible to use only the first item of the test (question 1; see Sect. 3.4.3) as a means of estimating median splits. Those who answer the question right are estimated to belong to the top half of highly educated participants while all others are assigned the bottom half. Note that the use of median splits can be problematic. Therefore, given the relatively small time savings over the adaptive format, we recommend this option be avoided whenever practical. Generally, this test format takes about as long as the Schwartz et al. (1997) instrument (i.e., about 1 min).

3.5 Results and Discussion

3.5.1 Psychometric Properties

Results of psychometric analyses are presented in Tables 3.2, 3.3, 3.4 and 3.5. The three formats of the Berlin Numeracy Test (i.e., computer-adaptive, paper-and-pencil, and single-item) are compared with the standard numeracy test by Lipkus et al. (2001) as well as with the brief three-item test by Schwartz et al. (1997).

Table 3.3 Psychometric properties of the numeracy tests: Convergent and discriminant validity

			Berlin Numeracy Test (Cokely et al. 2012)		
	Schwartz et al. (1997) 3 items	Lipkus et al. (2001) 11 items	Computer-adaptive test format	Paper-and-pencil format	Single-item format
Convergent validity					
Numeracy tests					
Lipkus et al. 11 items	0.75**				
Berlin Numeracy (computer-adaptive)	0.45**	0.49**			
Berlin Numeracy (paper-and-pencil)	0.50**	0.50**	0.91**		
Berlin Numeracy (single-item)	0.39**	0.42**	0.90**	0.75**	
Cognitive abilities/styles					
Fluid intelligence	0.41**	0.37**	0.48**	0.53**	0.41**
Cognitive reflection	0.40**	0.41**	0.51**	0.56**	0.41**
Crystallized intelligence	0.25**	0.21**	0.24**	0.25**	0.22**
Working memory span	0.14*	0.11	0.21**	0.20**	0.16**
Discriminant validity					
Motivation measures					
Maximizing–satisficing	0.01	0.04	0.05	0.04	0.05
Persistence (Grit-S)	0.02.	0.03	−0.05	−0.07	−0.03
Achievement motivation	−0.08	−0.10	−0.02	0.00	−0.01
Self-efficacy	0.00	−0.01	−0.01	0.02	0.03
Personality traits					
Emotional stability	−0.10	−0.05	0.01	0.05	−0.02
Conscientiousness	−0.09	−0.04	−0.09	−0.08	−0.06
Agreeableness	−0.03	−0.07	−0.14*	−0.08	−0.17**
Extraversion	−0.07	−0.06	−0.05	−0.05	−0.06
Openness to experience	−0.14*	−0.16**	−0.18**	−0.14*	−0.16**
Other measures					
Test anxiety	−0.15*	−0.16*	−0.12	−0.16*	−0.09
Implicit theories	−0.15*	−0.13**	−0.07	−0.10*	−0.04
Satisfaction with life	0.14*	0.08	0.12	0.16	0.07

*$p < 0.05$; **$p < 0.01$

Table 3.4 Psychometric properties of the numeracy tests: Predictive validity

			Berlin Numeracy Test (Cokely et al. 2012)		
	Schwartz et al. (1997) 3 items	Lipkus et al. (2001) 11 items	Computer-adaptive test format	Paper-and-pencil format	Single-item format
Predictive validity					
Understanding everyday risks	0.20**	0.18**	0.27**	0.31**	0.23**
Mean proportion correct of					
1st quartile	0.72	0.68	0.68	0.66	0.70
2nd quartile	0.74	0.66	0.70	0.70	
3rd quartile	0.78	0.78	0.74	0.78	0.78
4th quartile	0.78	0.78	0.84	0.84	

**$p<0.01$

Table 3.5 Explanatory value of the numeracy tests over and above Raven Advanced Progressive Matrixes and cognitive reflection test scores (beta coefficients from hierarchical regression analyses)

			Berlin Numeracy Test (Cokely et al. 2012)		
	Schwartz et al. (1997) 3 items	Lipkus et al. (2001) 11 items	Computer-adaptive test format	Paper-and-pencil format	Single-item format
As single predictor	0.20**	0.20	0.29**	0.34**	0.25**
With CRT	0.09	0.08	0.17**	0.23**	0.14*
With Raven	0.14*	0.15*	0.24**	0.31**	0.19**

*$p<0.05$; **$p<0.01$

3.5.2 *Basic Attributes*

In our highly educated sample, scores on the standard Lipkus et al. (2001) numeracy scale show dramatic negative skew (see Table 3.2). Although possible scores range from 0 to 11, the lowest observed score was 5 (45% correct). Both the mean and median are close to the measurement ceiling (i.e., 88% and 91% correct, respectively). Similar levels of skew are observed for the Schwartz et al. (1997) test. In contrast, scores on the Berlin Numeracy Test are distributed evenly across the whole range of possible scores regardless of format. In addition, all Berlin Numeracy Test formats typically take less time to complete than the standard Lipkus et al. (2001) numeracy scale.

3.5.3 *Convergent and Discriminant Validity*

If the Berlin Numeracy Test is successful in assessing levels of statistical numeracy, it should correlate with other numeracy tests and with measures of cognitive ability

(i.e., convergent validity). Moreover, to the extent the Berlin Numeracy Test primarily measures statistical numeracy it should not correlate with essentially unrelated constructs, such as motivation, personality, beliefs, or attitudes (i.e., discriminant validity). As Table 3.3 shows, both requirements—high correlations with related constructs and low with unrelated constructs—are satisfied for all three forms of Berlin Numeracy Test.

3.5.4 Predictive Validity

One of the intended purposes of the Berlin Numeracy Test is predicting people's understanding of risks in everyday contexts. To investigate the predictive validity of the Berlin Numeracy Test, we administered a short battery of items dealing with information about risks related to common consumer, health, and medical choices (e.g., evaluating toothpastes, cancer screenings), as well as information about probabilities typically used in forecasts (see Chap. 7; Galesic and Garcia-Retamero in press). Table 3.4 shows correlations of the different numeracy tests with the overall accuracy of answers to these items. All formats of the Berlin Numeracy Test were superior to the previous numeracy tests, essentially doubling the predictive resolution.

We further investigated the extent to which the Berlin Numeracy Test explained additional variance in risk understanding after controlling for the strongest alternative predictors of performance (i.e., fluid intelligence and cognitive reflection). As Table 3.5 shows, all formats of the Berlin Numeracy Test explain a substantial portion of additional variance after these others tests are included in a hierarchical regression model. In contrast, both the standard numeracy test by Lipkus et al. (in press) and the brief three-item test by Schwartz et al. (1997) lose most (or all) of their predictive power when intelligence or cognitive reflection tests are included. Overall, results indicate that the Berlin Numeracy Test is a reliable and valid test of statistical numeracy offering higher levels of discriminability and overcoming key psychometric limitations of previous numeracy tests.

3.6 Cross-Cultural Validation Studies

The initial validation of the Berlin Numeracy Test was completed on a sample of highly educated people living in a major metropolitan city in Germany. As a means of out-of-sample validation, we sought to assess the extent to which the test generalized to other highly educated samples from different cultures, presented in different languages. Specifically, we examined test performance in studies conducted in 14 different countries with diverse cultural backgrounds. Studies were conducted by different research groups, examining college-student samples at research-active universities, primarily drawn from introduction to psychology participant pools. Studies

were conducted in China (Tsinghua University), Japan (University of Tokyo), India (Thapar University), Pakistan (University of Punjab), Norway (University of Oslo),[4] Sweden (Uppsala University), England (University College London), France (Universite de Lausanne), Germany (Max Planck Institute for Human Development), Switzerland (University of Basel), Poland (Wroclaw University), Portugal (University of Porto),[5] Spain (University of Granada), and the USA (Michigan Technological University).[6] In total, data from 2,379 college students was examined. All reported data are scored via the adaptive Berlin Numeracy Test algorithm, where 2–3 questions (out of 4) are used to estimate statistical numeracy quartiles for each participant.[7]

Overall results show that the test generally discriminated within desirable tolerances (i.e., ±10%) for each quartile (see Table 3.6). Aggregating across all samples, the mean test score was 51.7% correct, which closely approximated the ideal score of 50%. This score indicates that on average, across all countries, the first question of the Berlin Numeracy Test achieved the intended 50% discriminability. Across all countries, we also observed modest underestimation of the third quartile and commensurate overestimation in the top quartile (i.e., the fourth quartile). In part, higher top quartile scores may reflect the fact that several of our samples were collected from some elite, highly selective universities (e.g., University College London; Tsinghua University in China). Visual inspection reveals some positive and negative skewing of scores across various countries.[8] For example, Spain, Pakistan, and India all show positive skew. In contrast, the sample from China was the highest performing group, showing extreme negative skew. Overall, however, when all groups were averaged together differences approximated the intended quartiles. The observed distributions indicate that with only 2–3 statistical numeracy questions the Berlin Numeracy Test achieves good discriminability across most countries even when presented in different languages or when used at elite or technological/engineering universities.

[4] Data collection in Norway used a standard rather than adaptive form of the Berlin Numeracy Test. Data reported in the table are calculated using the adaptive scoring algorithm, which was highly correlated with overall score, $r_{154} = 0.90$. In the standard format the average score was 62% correct showing modest skew (0.29).

[5] Data collection in Portugal used a modified Berlin Numeracy Test. Therefore, data were only available for the single-item test and are not presented in Table 3.6. Overall 46.4% of participants ($n = 306$) from Portugal answered the first question right (theoretical ideal test score = 50%).

[6] We thank Nicolai Bodemer, Siegfried Dewitte, Stefan Herzog, Marcus Lindskog, Hitashi Lomash, Yasmina Okan, Jing Qian, Samantha Simon, Helena Szrek, Masanori Takezawa, Karl Teigen, Jan Woike, and Tomek Wysocki for assistance with cross cultural data collection.

[7] Translation involved iterative cycles of back-translation with revision.

[8] The Berlin Numeracy Test estimates quartiles and so caution is required when interpreting standard assessments of skew.

Table 3.6 Percentage of people in each quartile from 14 different countries estimated by the computer-adaptive test format of the Berlin Numeracy Test. Countries are ordered by their percentage of top quartile scores

Country	Language	*N*	1st quartile	2nd quartile	3rd quartile	4th quartile
China	English	166	0.04	0.07	0.14	0.75
Poland	Polish	205	0.14	0.20	0.22	0.44
England	English	420	0.20	0.31	0.14	0.35
Japan	Japanese	63	0.06	0.36	0.24	0.34
Sweden	Swedish	47	0.21	0.28	0.17	0.34
France	French	86	0.30	0.13	0.23	0.34
USA	English	55	0.20	0.29	0.20	0.31
Switzerland	German	503	0.26	0.23	0.23	0.28
Germany	German	173	0.29	0.21	0.22	0.28
Norway	Norwegian	156	0.25	0.24	0.25	0.26
Belgium	Dutch	50	0.30	0.30	0.16	0.24
India	English	83	0.19	0.52	0.08	0.21
Pakistan	English	114	0.29	0.41	0.19	0.11
Spain	Spanish	258	0.48	0.41	0.07	0.04
Total		2,379	0.23	0.28	0.18	0.31

3.7 Validation Across Different Populations

3.7.1 Numeracy in Physician Assistants

One goal for the Berlin Numeracy Test is to offer an instrument that can quickly assess statistical numeracy in working professionals. Of particular interest are those professionals who commonly make risky decisions and communicate risks. One such group in the USA is physician assistants. Physician assistants are independently licensed medical professionals who diagnose and treat patients, and provide care similar to that provided by a physician across all medical subspecialties (e.g., emergency medicine, family practice, surgery). Physician assistants' training typically involves 2 or 3 years of postgraduate study and clinical rotations, usually leading to a terminal master's degree.

As noted, previous studies of physicians-in-training in the UK (Hanoch et al. 2010) revealed dramatic skew in responses to the Lipkus et al. (2001) test. Specifically, in one sample of physician-in-training, Hanoch and colleagues found that the average Lipkus et al. (2001) test score was 95% correct, with 64% of participants answering all questions correctly. Here, we assessed performance of the Berlin Numeracy Test by administering the paper-and-pencil format to a group of physician assistant students ($n=51$) who were completing their final semester of training at the University of Oklahoma.[9] Results of the study indicated that the mean test score was 44.3% correct, which reasonably approximated the ideal score of

[9] We thank Robert Hamm for data collection.

Table 3.7 Percentage of people in each quartile from three different samples estimated by the computer-adaptive test format of the Berlin Numeracy Test

Sample	*N*	1st quartile	2nd quartile	3rd quartile	4th quartile
Graduating US physician assistants	51	0.16	0.39	0.29	0.16
General population of Sweden	213	0.20	0.36	0.24	0.20
USA web-panel sample (M-Turk)	1,612	0.49	0.27	0.12	0.13
Total	1,876	0.28	0.34	0.22	0.16

50%. Results also revealed very modest positive skew (0.16) indicating the test was generally well calibrated. A similar distribution was observed when the adaptive scoring algorithm was applied (Table 3.7). Note that in contrast to other highly educated samples, these data show slightly more central clustering of scores. To the extent this pattern generalizes, it suggests physician assistants are somewhat less likely to have either very low or high levels of statistical numeracy. Overall, results indicate that the Berlin Numeracy Test is well suited for use with these and other professionals and individuals with post-graduate educations. Ongoing research is assessing test performance among other professional groups (e.g., judges, lawyers, physicians, dieticians, financial advisors).

3.7.2 Numeracy in the General Population

The Berlin Numeracy Test was designed for, and normed with, highly educated individuals. However, considering the observed skew in scores from the Lipkus et al. (2001) test, the Berlin Numeracy Test may also be suitable for use with some well-educated general populations. As part of a larger validation and translation study, data were collected from 213 adults in Sweden who were sampled to be representative of the general population (see Lindskog et al. 2012).[10] The test was presented in Swedish and was administered using the computer-adaptive test format. Results show that the average test score was 48.8% correct, which closely approximated the theoretically ideal score of 50%. Distributions of estimated quartiles were somewhat concentrated around the middle quartiles, particularly the second quartile (see Table 3.7). This suggests that compared to other highly educated groups of individuals, there are moderately fewer people in Sweden with either very low or very high levels of statistical numeracy.

In addition, participants in this study also completed the Lipkus et al. (2001) test. As expected, results showed rather profound skew in scores with an average score

[10] This research was financed by the Swedish Research Council. We thank Marcus Lindskog and colleauges for these data.

of 83.5% correct and clear negative skew (−1.94). We compared the scores in the Lipkus et al. (2001) test in this study with those in the study of Galesic and Garcia-Retamero (2010) using probabilistic national samples in the USA and Germany (see Chap. 2). Results indicate that Swedish residents' scores showed considerably more negative skew reflecting significantly higher levels of numeracy compared to the populations in Germany, $t_{1,209} = 9.29$, $p = 0.001$, skew = −0.55, and the USA, $t_{1,375} = 13.51$, $p = 0.001$, skew = −0.33.

Overall, results indicate that the Berlin Numeracy Test is well suited for estimating numeracy among the general population of Sweden and other similar highly numerate countries. However, because the general population of Sweden is more numerate than that of either the USA or Germany, we can expect positive skew in general population samples from the USA, Germany, and other similar countries. Accordingly, when assessing statistical numeracy in most general populations we suggest including at least one other test in addition to the Berlin Numeracy Test (e.g., Weller et al. 2012). One promising strategy that adds only about 1 min in testing time is to combine the three-item Schwartz et al. (1997) test with the Berlin Numeracy Test data (for an example see Sect. 3.7.3). Ongoing studies are examining this potential strategy along with performance of the Berlin Numeracy Test in probabilistic national samples of residents in the USA.

3.7.3 Numeracy in Web-Panel Data

Behavioral scientists are increasingly using paid web panels for data collection and hypothesis testing. One popular option for data collection is Amazon.com's Mechanical Turk web panel (for a review see Paolacci et al. 2010). The first published study to assess numeracy among participants from Mechanical Turk was published in 2010. In this study, Paolacci et al. (2010) assessed numeracy using a subjective numeracy scale (see Chaps. 2 and 15; see also Fagerlin et al. 2007), which is known to correlate with the Lipkus et al. (2001) test. Results revealed an average subjective numeracy score of 4.4 (i.e., about 67% of maximum), which is in line with previously reported scores (e.g., participants recruited from a university hospital with a modest skew of −0.3; see Fagerlin et al. 2007). Similarly, we recently investigated numeracy using the Schwartz et al. (1997) test on a convenience sample using Mechanical Turk ($n = 250$; Okan et al. 2012). Consistent with results from the subjective numeracy test, results showed an average score of 2.1 (i.e., 70% correct), which revealed moderate negative skew (−1.2). A total of 42% of the sample also answered 100% of the questions correct.

To evaluate the performance of web panelists on the Berlin Numeracy Test, we administered the computer-adaptive test format to a large Mechanical Turk web-panel convenience sample ($n = 1{,}612$). All reported data were scored via the adaptive algorithm, where 2–3 questions (out of 4) are used to estimate statistical numeracy quartiles for each participant. As anticipated, we observed positive

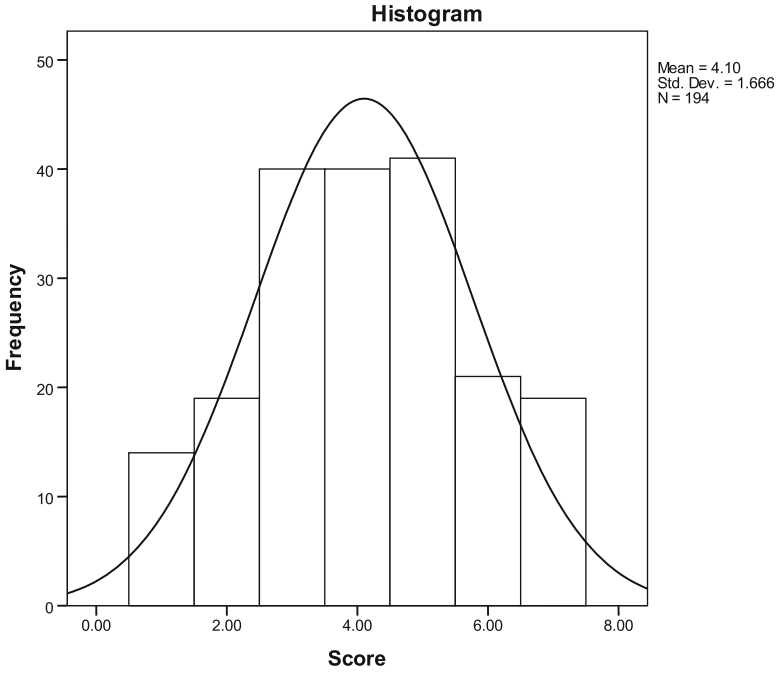

Fig. 3.2 Distribution of combined scores (Mechanical Turk web-panel sample) on the Berlin Numeracy Test and the Schwartz et al. (1997) three numeracy items

skew (0.90) in the sample scores indicating that the test was somewhat too difficult (see Table 3.7).[11] This finding of positive skew is not surprising given that the Berlin Numeracy Test was designed to measure numeracy among highly educated samples.

In the web-panel studies we mentioned above, we observed positive skew for the Berlin Numeracy Test and negative skew for the Schwartz et al. (1997) test. It stands to reason that combining the two tests would yield a better distribution, providing increased discriminability. Therefore, we conducted a new study including both the Schwartz et al. (1997) test and the Berlin Numeracy Test with a convenience sample of participants on Mechanical Turk ($n = 206$). When scored separately, we replicated the negative (−0.62) and positive (0.48) skewing of scores on the two tests. However, simply adding the two scores together yielded a normal distribution with no evidence of skew (−0.016; Fig. 3.2). In summary, combining the Berlin Numeracy Test with the Schwartz et al. (1997) test provides a very fast assessment (<4 min) with

[11] To the extent our data generalize, results suggest that our single question 2a (see Sect. 3.4.3) may allow for a rough approximation of a median split among Mechanical Turk participants. This question is simpler/easier than question 1 (see Sect. 3.4.4.3), and therefore was a good approximation of a median split in less highly educated samples.

good discriminability that is well suited for use with Mechanical Turk. In addition, combining both tests should also be appropriate for measuring numeracy in other general samples (e.g., older adults).

3.8 A Multiple-Choice Format

In some cases researchers may require more flexibility than the current Berlin Numeracy Test formats provide. For example, many psychometric tests are given in a multiple-choice format. Unfortunately, providing potential answers to participants increases the benefits of simple guessing. With four options, guessing would be expected to yield a score of approximately 25% correct. In contrast, in all other "fill in the blank" formats of the Berlin Numeracy Test, the contribution of a guessing parameter is essentially zero. To address this issue, we developed a multiple-choice format of the test, which began with an analysis of patterns of incorrect responses to previous tests from participants in the aforementioned Mechanical Turk study ($n = 1{,}612$). For each question, we selected the most frequently listed incorrect options (recorded in 8–20% of incorrect answers). We then included the correct answer, the two highest frequency incorrect answers, and a "none of the above" option.

Next, we collected data from participants at the Michigan Technological University ($n = 269$). Participants included convenience samples primarily from Departments of Psychology, Mechanical Engineering, and Computer Science. The majority of participants were undergraduate students, with a small proportion of the sample composed of either graduate students or faculty. Participants were either sent a link asking them to complete a survey via internal listservs or tests were administered in classes. Participants were presented with one of the two versions of the multiple-choice format differing only in the wording of question 1 (see Sect. 3.4.3).[12] This manipulation was conducted because we received feedback that some professional groups may be more willing to participate if questions seemed related to their areas of expertise (e.g., some medical doctors will see more face validity in questions about genetic mutations as compared to choir membership). Accurate responses to the new ($M = 0.56$) vs. old ($M = 0.60$) question did not reliably differ $\chi_1^2 = 0.26$. Distributions of scores did not significantly differ between tests either, $t_{267} = 1.38$, $p = 0.17$, and so data sets were combined for subsequent analyses. Overall, the mean multiple-choice test score was 55% correct, which reasonably approximated the ideal score of 50%. Analysis of distributions of responses indicated that the multiple-choice format showed no skew (−0.01). Results indicate that the multiple-choice format provided good discriminability and remained well balanced even when used with highly numerate individuals (e.g., computer science students).

[12] The exact wording of the alternative question is as follows: "Out of 1,000 people in a small town, 500 have a minor genetic mutation. Out of these 500 who have the genetic mutation, 100 are men. Out of the 500 inhabitants who do not have the genetic mutation, 300 are men. What is the probability that a randomly drawn man has the genetic mutation?"

3.9 Discussion and Conclusions

Over the last decade, the Schwartz et al. (1997) and Lipkus et al. (2001) numeracy tests have proven useful and even essential for some aspects of theory development, as well as for applications in risk communication. However, as anticipated by Lipkus et al. (2001), in the 10 years since publication of their test, research has identified a number of limitations and opportunities for improvement in measures of statistical numeracy. Building on the work of Lipkus et al. (2001), Schwartz et al. (1997), and many others (e.g., Peters et al. 2006, 2007b; Reyna et al. 2009), we developed and validated a flexible, multi-format test of statistical numeracy for risk literacy in educated samples: The Berlin Numeracy Test, which measures the range of statistical numeracy skill that is important for accurately interpreting and acting on information about risk. With the help of colleagues from around the world, we conducted 21 validation studies showing that a very short, adaptive format of the Berlin Numeracy Test provides sound assessment with dramatically improved discriminability across diverse populations, cultures, education levels, and languages. Content validity is clear in the types of questions included in the test—i.e., math questions involving ratio concepts and probabilities. Convergent validity was documented by showing high intercorrelations with other numeracy tests, as well as with other assessments of general cognitive abilities, cognitive styles, and education. Discriminant validity was documented by showing that the test was unrelated to common personality and motivation measures (e.g., uncorrelated with emotional stability). Predictive validity was documented by showing that the Berlin Numeracy Test provided unique predictive validity for both numeric and non-numeric everyday risky decision-making. This unique predictive validity held when statistically controlling for all the existing numeracy tests and other general ability and cognitive-style instruments. Taken together, results converge and contribute to our evolving understanding of the construct validity of numeracy.[13]

Going forward, more research is needed to document the causal linkages between numeracy and risky decision making (for a detailed discussion see Cokely et al. 2012). Theoretically, improving some types of math skills will improve risk literacy and risky decision making. However, the evidence of such benefits along with quantification of the magnitudes of benefits is surprisingly limited (e.g., how much study time is required to improve decisions). As well, despite the utility of current theoretical frameworks, our theoretical understanding underlying mechanisms is underspecified. Research is likely to benefit by more closely aligning with current research in mathematics and general literacy education, as well as research on mathematics development (e.g., Siegler 1988), mathematics expertise, and training

[13] According to Cronbach and Meehl's (1955) review of construct validity "a construct is some postulated attribute of people, assumed to be reflected in test performance." Similarly, contemporary views hold that construct validity "…is not a property of the test or assessment as such, but rather of the meaning of the test scores" which is established by integrating and evaluating multiple lines of evidence (Messick 1995).

for transfer. Additionally, there is a need for validated tests that provide larger item pools and parallel forms that can be administered multiple times to assess learning. Related development efforts are currently underway for the Berlin Numeracy Test.

It is important to again note that the Berlin Numeracy Test is designed specifically for educated samples (e.g., college students, business, medical, and legal professionals). Discriminability will be reduced when assessing individuals who have lower levels of educational attainment or when administered to groups that come from considerably less selective universities (i.e., the Berlin Numeracy Test will show some positive skew in less educated samples). When this is a concern, researchers can include an additional instrument such as the fast three-item test by Schwartz et al. (1997). The results of our Mechanical Turk's web-panel study (see Sect. 3.7.3) show that this strategy can produce excellent discriminability with virtually no skew providing a 4-min assessment that is sensitive to both low and high levels of statistical numeracy.

Because the Berlin Numeracy Test provides a broad estimate of variation in statistical numeracy it is not able to provide detailed assessment of differences in specific numeracy skills, such as identifying deficits in reasoning about probability as compared to proportions or multiplication. As noted, factor analytic research by Liberali et al. (2012) indicates that, at least with respect to some risky decisions and judgments, component numeracy skills (e.g., multiplication vs. probability) may be differentially beneficial.[14] We also currently do not have any theoretical account systematically linking component numeracy skills and competencies with the many various types of risky decisions people commonly face. There is a need for larger scale cognitive process tracing and factor analytic assessments to be conducted across all aspects of numeracy, risk literacy, and risky decision making. Initial studies may benefit by examining relations between established numeracy tests, component math skills, and other established instruments such as the advanced decision-making competency tests (Bruine de Bruin et al. 2007; Parker and Fischhoff 2005).

Future research will need to use methods that provide details about the ecological frequencies of problematic risky decisions related to numeracy, including techniques like representative sampling (Dhami et al. 2004). This type of epidemiological data could then be used to start to quantify the economic, personal, and social impact of specific weaknesses in numeracy and risk literacy (e.g., is denominator neglect a dangerous factor in high-stakes risky decisions and to what extent does numeracy inoculate? For related discussion see Chap. 10; see also Garcia-Retamero et al. 2012). This ecological approach would provide essential input for relative prioritization of different interventions (i.e., which kind of problems do the most harm and which kinds of interventions will produced the biggest benefits). Unfortunately, because there may be many numeracy skills a test of all component skills may turn out to be very long. In this case, and perhaps even if a comprehensive test is not particularly long, adaptive testing is likely to offer many benefits (Thompson and Weiss 2011).

[14] The factor structures varied across two studies, which complicate interpretation. Nevertheless, the results are suggestive.

Research on all these topics is ongoing in our laboratories. As new tools, interactive activities, and improved tests become available they will be added to the content on http://www.riskliteracy.org (for other individual difference measures see also Appelt et al. 2011; http://www.sjdm.org/dmidi/).

References

Appelt, K. C., Milch, K. F., Handgraaf, M. J. J., & Weber, E. U. (2011). The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making, 6*, 252–262.

Black, W. C. W., Nease, R. F. R., & Tosteson, A. N. A. (1995). Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. *Journal of the National Cancer Institute, 87*, 720–731.

Blackwell, L., Trzesniewski, K., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*, 246–263.

Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*, 382–398.

Bruine de Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938–956.

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*, 25–47.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*, 20–33.

Cokely, E. T., Kelley, C. M., & Gilchrist, A. H. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin and Review, 13*, 991–997.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2010). Executive functionions in decision making: An individual differences approach. *Thinking and Reasoning, 16*, 69–97.

Del Missier, F. T., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.731

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Diener, E., Emmons, R. A., Larsen, R., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*, 7716–7720.

Ericsson, A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*, 114–121.

Fagerlin, A., Zikmund-Fisher, B., 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*, 672–680.

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 18*, 342–350.

Feltz, A., & Cokely, E. T. (2012). The philosophical personality argument. *Philosophical Studies.* doi: 10.1007/s11098-011-9731-4

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*, 316–344.

Fox, M. C., Roring, R., & Mitchum, A. L. (2009). Reversing the speed-IQ correlation: Intra-individual variability and attentional control in the inspection time paradigm. *Intelligence, 37*, 76–80.

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

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine, 170*, 462–468.

Galesic, M., & Garcia-Retamero, R. (in press). Using analogies to communicate information about health risks. Applied Cognitive Psychology. doi: 10.1002/acp.2866.

Garcia-Retamero, R., & Galesic, M. (2010a). How to reduce the effect of framing on messages about health. *Journal of General Internal Medicine, 25*, 1323–1329.

Garcia-Retamero, R., & Galesic, M. (2010b). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science and Medicine, 70*, 1019–1025.

Garcia-Retamero, R., Okan, Y., & Cokely, E. T. (2012). Using visual aids to improve communication of risks about health: A review. *The Scientific World Journal.* Article ID 562637. Doi:10.1100/2012/562637

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*, 504–528.

Guadagnoli, E., & Ward, P. (1998). Patient participation in decision-making. *Social Science and Medicine, 47*, 329–339.

Hanoch, Y., Miron-Shatz, T., Cole, H., Himmelstein, M., & Federman, A. D. (2010). Choice, numeracy and physicians-in-training performance: The case of Medicare part D. *Health Psychology, 29*, 454–459.

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress, and Coping, 10*, 219–244.

Lang, J. W. B., & Fries, S. (2006). A revised 10-item version of the Achievement Motives Scale: Psychometric properties in German-speaking samples. *European Journal of Psychological Assessment, 22*, 216–224.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.752

Lindenberger, U., Mayr, U., & Kliegl, R. (1993). Speed and intelligence in old age. *Psychology and Aging, 8*, 207–220.

Lindskog, M., Kerimi, N., Winman, A., & Juslin, P. (2012). A Swedish validation study of the Berlin Advanced Numeracy Test (manuscript in preparation).

Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical insights. *Health Education & Behavior, 36*, 1065–1081.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly-educated samples. *Medical Decision Making, 21*, 37–44.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *The American Psychologist, 51*, 77–101.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making.* doi: 10.1002/bdm.751

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual differences approach. *Journal of Behavioral Decision Making, 18*, 1–27.

Peters, E., Dieckmann, N. F., Dixon, A., Slovic, P., Mertz, C. K., & Hibbard, J. H. (2007a). Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64*, 169–190.

Peters, E., Hibbard, J. H., Slovic, P., & Dieckmann, N. F. (2007b). Numeracy skill and the communication, comprehension, and use of risk and benefit information. *Health Affairs, 26*, 741–748.

Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making, 3*, 619–635.

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

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1–48.

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*, 943–973.

Schapira, M. M., Walker, C. M., & Sedivy, S. K. (2009). Evaluating existing measures of health numeracy using item response theory. *Patient Educational and Counseling, 75*, 308–314.

Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition, 20*, 1722–1731.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*, 1178–1197.

Schwartz, L. M. L., Woloshin, S. S., Black, W. C. W., & Welch, H. G. H. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*, 966–972.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio* (pp. 35–37). Windsor: Nfer-Nelson.

Sherrod, P. H. (2003). *DTREG: Predictive Modeling Software*. http://www.dtreg.com, accessed June 30, 2012.

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General, 117*, 258–275.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *The Behavioral and Brain Sciences, 23*, 701–726.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*, 1–9.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

Weller, J. A., Dieckmann, N., Tusler, M., Mertz, C. K., Burns, W., & Peters, E. (2012). Development and testing of an abbreviated numeracy scale: A rasch analysis approach Journal of Behavioral Decision Making. doi: 10.1002/bdm.1751

Zikmund-Fisher, B., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making, 27*, 663–671.