Judgment and Decision Making, Vol. 9, No. 1, January 2014, pp. 15–34

Predicting biases in very highly educated samples: Numeracy and metacognition

Saima Ghazal* Edward T. Cokely,*† Rocio Garcia-Retamero†‡

Abstract

We investigated the relations between numeracy and superior judgment and decision making in two large community outreach studies in Holland (n=5408). In these very highly educated samples (e.g., 30–50% held graduate degrees), the Berlin Numeracy Test was a robust predictor of financial, medical, and metacognitive task performance (i.e., lotteries, intertemporal choice, denominator neglect, and confidence judgments), independent of education, gender, age, and another numeracy assessment. Metacognitive processes partially mediated the link between numeracy and superior performance. More numerate participants performed better because they deliberated more during decision making and more accurately evaluated their judgments (e.g., less overconfidence). Results suggest that well-designed numeracy tests tend to be robust predictors of superior judgment and decision making because they simultaneously assess (1) mathematical competency and (2) metacognitive and self-regulated learning skills.

Keywords: numeracy, risk literacy, individual differences, cognitive abilities, superior decision making, judgment bias, metacognition, confidence, dual systems.

1 Introduction

Statistical numeracy—i.e., one's practical understanding of probabilistic and statistical problem solving—is one of the strongest domain-general predictors of superior judgment and decision making in both numerical and non-numerical tasks (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Cokely & Kelley, 2009; Kutner, Greenberg, Jin, & Paulsen, 2006; Lipkus & Peters, 2009; Peters, 2012; Peters & Levin, 2008; Peters et al., 2006; Reyna, Nelson, Han, & Dieckmann, 2009). Numeracy also tends to be a substantial independent predictor of superior performance when compared with tests of fluid intelligence, cognitive reflection, and attentional control (Cokely et al., 2012; Låg, Bauger, Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Lindberg, & Friborg, 2013; Schapira et al., 2012; Weller, Dieckmann, Tusler, Mertz, Burns, & Peters, 2013).[1] Research indicates that the link between numeracy and superior decision making does not primarily reflect differences in abstract reasoning or neoclassically normative decision strategies.[2] Instead, numeracy's predictive power often reflects differences in (1) heuristic-based deliberation (e.g., deep elaborative processing, Cokely & Kelley, 2009; Cokely et al., 2012); (2) affective numerical intuition (e.g., precise symbolic number mapping, Peters, 2012; Peters et al., 2006); and (3) meaningful intuitive understanding (e.g., gist-based representation and reasoning; Reyna, 2004, 2012; Reyna & Brainerd, 2005b; Reyna et al., 2009).

There are now many established and newer numeracy tests validated for use with diverse samples (e.g., the "Numeracy Understanding in Medicine Instrument" (NUMi) for older-adult patient samples; Schapira et al., 2012). However, most numeracy tests are not appropriate for the

Financial support for this research was provided by grants from the National Science Foundation (SES–1253263) and the Ministerio de Economía y Competitividad (entitled "Helping Doctors and Their Patients Make Decisions About Health", PSI2011–22954). We are grateful to Han van der Maas, Marthe Straatemeier, and colleagues and staff at the University of Amsterdam, and Hans van Maanen and staff from *de Volkskrant* newspaper for their support and assistance with data collection.

*Department of Cognitive and Learning Sciences, Michigan Technological University. Authorship is equal for the first two authors. Correspondence concerning this article should be addressed to Edward T. Cokely, Department of Cognitive and Learning Sciences, Michigan Technological University. Email: ecokely@mtu.edu.

†Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development.

‡Department of Psychology, University of Granada

[1] Some innovative research shows that executive functions can outpredict numeracy under some conditions (Del Missier, Mäntylä, & Bruine de Bruin, 2012; Del Missier, Mäntylä, Hansson, Bruine de Bruin, & Parker, 2013). However, as noted by the authors (see also Cokely et al., 2012), highly sensitive numeracy tests were not yet widely available and could not be used in these studies. Related replication and extension studies are currently ongoing in our laboratory.

[2] Although numeracy tends to predict superior performance, numeracy is also positively correlated with some non-normative biases. See Peters et al. (2006) for the seminal example of numeracy's link with heuristic processes that can give rise to both normatively superior and inferior judgment and decision making. See Cokely and Kelley (2009) for a cognitive process tracing study detailing qualitative and quantitative differences in heuristic search and elaborative encoding.

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

measurement of statistical numeracy in highly educated participants such as professionals working in medicine and finance (for a review of available numeracy tests see Cokely, Ghazal, & Garcia-Retamero, 2013, in press). One exception is the Berlin Numeracy Test (BNT), which has been found to provide superior psychometric sensitivity in moderate to very highly numerate participants (e.g., college students, professionals, computer literate adults; Cokely et al., 2012). Accordingly, we build on previous research investigating the mechanisms, robustness, and generalizability of numeracy by examining the performance of two very highly educated community samples on a small set of paradigmatic judgment and decision making tasks. We begin with a literature review providing an overview of some of the notable findings and numeracy assessment tools that are now available. We then present results of two new studies conducted as part of our RiskLiteracy.org outreach efforts (e.g., a study included in a newspaper report about the importance of statistics for decision making). We conclude with a discussion of the links between numeracy, metacognition, and superior judgment and decision making.

1.1 Numeracy

Experts do not agree on an exact and uncontroversial theoretical definition of mathematics. Fortunately, quantitative skills are easier to operationalize and measure. For more than 50 years, researchers have studied the causes and consequences of *numeracy* (Huff & Geis, 1954; Paulos, 1988), including extensive longitudinal studies conducted in large diverse samples such as the National Assessment of Adult Literacy (NAAL; Kutner et al., 2006) and the Program for International Student Assessment (PISA; OECD, 2012). There is wide agreement that the theoretical construct of "numeracy" is not synonymous with pure mathematical skill but instead refers to mathematical or quantitative literacy (Steen, 1990; see also Nelson, Reyna, Fagerlin, Lipkus, & Peters, 2008, and Reyna et al., 2009), reflecting an emphasis on "mathematics in context" as described in the *US Common Core State Standards Initiatives*. Specifically, the construct "numeracy" refers to the "array of mathematically related proficiencies that are evident in adults' lives ... including a connection to context, purpose, or use... for active participation in the democratic process and... in the global economy" (Ginsburg, Manly, & Schmitt, 2006). At the more basic levels, numeracy involves the "real number line, time, measurement, and estimation" whereas higher levels focus on an "understanding of ratio concepts, notably fractions, proportions, percentages, and probabilities" (Reyna et al., 2009).

Within the decision sciences, efforts to understand and measure numeracy involve both subjective and performance assessments. For example, one validated subjective assessment of numeracy often used in health and medical domains asks participants eight questions in which they judge their personal levels of numeracy (e.g., "How good are you at working with fractions;" Fagerlin et al, 2007; Zikmund-Fisher, Smith, Ubel, & Fagerlin, 2007; and for subjective graph literacy see Garcia-Retamero, Cokely, & Ghazal, 2014b). Several studies indicate moderate-to-high correlations between objective and subjective measures (Fagerlin et al., 2007; Liberali et al., 2012; Weller et al., 2013; Zikmund-Fisher et al., 2007). Studies further show the subjective test can provide unique predictive power beyond intelligence test scores (Låg et al., 2013). Nevertheless, other research indicates that people can be highly overconfidence in reporting their subjective numerical ability. For example, Sheridan, Pignone, and Lewis (2003) showed that 70% of subjects reported that they consider themselves to be "good with numbers", while only 2% of those respondents correctly answered three objective numeracy questions (see also Dunning, Heath, & Suls, 2004).[3]

Performance based numeracy assessments are the most commonly used methods in the allied decision sciences. The longest-standing and most widely used assessments of numeracy are based on classical testing theory, which estimates theoretical differences in abilities based on one's relative test score (Novick, 1966; see also Cokely et al., 2013, in press; Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006; Schapira, Walker, & Sedivy, 2009; Schwartz et al., 1997). To illustrate, in 1997, Schwartz et al. (1997) conducted a seminal randomized cross-sectional numeracy study investigating the relations between numeracy and relative risk perceptions. Five hundred women were initially mailed the study stimuli and asked to participate. Respondents included 287 mostly older adult women (mean age 68 years) who were veterans with modest incomes (e.g., less than $25,000 per year). The majority of participants had also completed high school (96%) and about a third had completed at least some college. Numeracy was assessed with three items that were similar to and based on items used in the NAAL survey (see previous section). Once scored, these items were used to predict the women's understanding of data presented in one of four formats (e.g., relative risk reduction versus absolute risk reduction with baseline). The women were asked to interpret the material provided and to report on the risks/benefits of mammography screening (e.g., "Imagine 1000 women exactly like you. Of these women what is your best guess about how many will die from breast cancer during the next 10 years if they are not screened every year for breast cancer?"). Results indicated that about

[3]The three items were from the test by Schwartz et al. (1997). These results suggest that subjective instruments are likely best suited for specific purposes, including rapid, rough numeracy assessment among people who have some math anxiety.

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

half of the women (i.e., 54%) accurately answered two questions, while only 20% accurately answered all three (i.e., most could not convert 1 in 1000 to 0.1%). As expected, results also revealed a moderate positive correlation between participants' final score and their relative risk reduction interpretations, providing evidence of decision-related criterion validity for the brief assessment.

The results of Schwartz et al. (1997) and the subsequent results provided by Lipkus et al. (2001) were timely for a number of reasons (for reviews see Cokely et al., 2012, in press).[4] First, the results provided additional evidence that among community samples in the United States some sizable proportion of individuals were likely to be statistically innumerate (e.g., 20% failed questions dealing with risk magnitude), a result that accords with findings from the NALS and NAALS National Surveys. Such findings are important, as many efforts designed to support informed and shared decision making rest on an assumption that decision-makers are numerate (or at least sufficiently statistically numerate; see also Edwards & Elwyn, 2009, and Guadagnoli & Ward, 1998). Second, results indicated that domain framing (e.g., medical, financial, or abstract gambles) did not tend to affect test performance or comprehension. This finding indicates that various domain-specific items (e.g., items framed in terms of financial, medical or gambling risks) can provide a reasonable basis for the assessment of domain-general statistical numeracy skills, although it is theoretically possible that domain familiarity will confer some additional decision performance advantages (Levy, Ubel, Dillard, Weir, & Fagerlin, 2014).

1.2 Advances in numeracy assessment

After more than a decade of research using classical tests of numeracy, research in the decision sciences has turned to modern psychometric testing paradigms—i.e., Item Response Theory (IRT) and its variants. In contrast to classical testing theory, item response theory requires modeling of probabilistic distributions over test taker's responses to specific items. The focus of test development is on the *item* rather than on the pooled responses to items as in classical testing theory. A full description of the theory is beyond the scope of this paper (see Lord, 1980; Van der Linden & Hambleton, 1997); however, it is useful to note that IRT tests improve predictive performance by eliminating item redundancy with estimated parameters including item difficulty (e.g., how hard is any particular item for a given trait level), discrimination (e.g., how sharply and consistently does an item distinguish individuals at higher versus lower trait levels), and guessing (e.g., true/false items will be guessed correctly 50% of the time). To illustrate, Schapira et al. (2012) developed the Numeracy Understanding in Medicine Instrument (NUMi) to provide a higher-fidelity assessment of basic health numeracy among less educated patient samples. The 20 item test was developed using a two parameter IRT approach integrating four numeracy sub-skills (e.g., graph literacy, statistical numeracy). Results reveal that the NUMi test is robust and provides good psychometric sensitivity that is suitable for use with less numerate individuals (e.g., older adult patient samples). Results also provided evidence of construct validity and unique predictive power (e.g., independent of the predictions of general intelligence tests).

Using a Rasch analysis, which is akin to a one parameter IRT-type approach, Weller et al. (2013) developed an eight item numeracy measure optimized for use with the general population of the United States. Test development involved comparison of 18 items taken from existing measures of numeracy and a cognitive reflection test. Specifically, items were drawn from tests developed by Lipkus et al. (2001) (which includes the items of Schwartz et al., 1997), and tests developed by Peters et al. (2007), and Frederick (2005). The resulting scale provides greatly improved psychometric discriminability when used with the general population of the United States. Evidence also indicates that the test provides stronger predictive validity for risk judgments (i.e., Låg et al., 2013; Lipkus et al., 2001). Despite these notable improvements, one limitation of the Weller et al. (2013) abbreviated numeracy scale, as well as the test items analyzed by Låg and colleagues (2013), is that they combine two distinct types of test items with differential ranges of sensitivity to improve psychometric sensitivity of the numeracy assessments. In particular, they include: (1) some relatively difficult items designed to measure cognitive impulsivity/reflection (i.e., the CRT by Frederick, 2005) and (2) some relatively easy items designed to measure statistical numeracy.[5]

[4] There are also a number of performance measures of numeracy that assess one's approximate number system—a related but independent theoretical construct. For a recent example of these tests see Lindskog, Winman, and Juslin (2013).

[5] Although confirmatory factor analysis has indicated that the constructs can be considered one factor, there is reason to be cautious with this interpretation. The two types of items have been found to dissociate in theoretically notable ways, differentially predicting financial judgments, reasoning, and risk comprehension (Cokely et al., 2012; Cokely, Parpart, & Schooler, 2009; Di-Girolamo, Harrison, Lau, & Swarthout, 2014; Låg et al., 2013; Liberali et al., 2012). Recent results also indicate the two types of items can load on different factors (Liberali et al., 2011) and that statistical numeracy alone can capture all reliable variance associated with the CRT in some tasks involving highly educated individuals (Låg et al., 2013). Differences in item type are also responsible for differences in psychometric discrimination at different ranges (e.g., CRT items are harder and numeracy items are easier; Låg et al., 2013; Weller et al., 2013).

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

1.3 The Berlin Numeracy Test

Building on the work of Lipkus et al. (2001) and Schwartz et al. (1997), Cokely and colleagues (2012) developed a fast psychometric test of differences in *statistical numeracy* among educated samples of adults living in diverse industrialized countries (e.g., college students, working professionals, and computer literate adults). The test was created using new statistical numeracy items selected from a large pool of candidate items. All items were subjected to think aloud protocol analysis to control for potential confounds from factors such as linguistic confusion. The test was then developed using a decision tree application from the predictive modeling software DTREG (Sherrod, 2003). The analysis yielded several versions of the test (see http://www.RiskLiteracy.org for links and test format recommendation tools), including (i) the adaptive test that adjusts item difficulty based on a test-takers previous responses (2–3 items; about 2.5 minutes duration) and (ii) a traditional 4 item paper-and-pencil test (4 items; < 5 minutes duration). Psychometrically the decision tree's assessment approximates an item response theory analysis identifying items with high levels of discriminability across a specified range of item difficulty, with a guessing parameter of zero.

The construct validity, reliability, and psychometric sensitivity of the Berlin Numeracy Test was initially established in 21 studies (n=5336) of participants from 15 countries including assessments of diverse groups (e.g., US medical professionals, community samples, Mechanical Turk web-panels). Validation studies have since been extended to participants from 60 countries and include several patient and physician samples from all over the world (Garcia-Retamero, Cokely, & Ghazal, 2014a; Garcia-Retamero, Wicki, Cokely, & Hanson, in press). Initial and subsequent analyses indicate that the test offers robust sensitivity, with optimal performance among those who have some college education.[6] The test was also found to be the strongest predictor of understanding everyday risks (e.g., evaluating claims about products and treatments; interpreting forecasts), doubling the predictive power of other numeracy instruments and accounting for unique variance beyond other cognitive tests (e.g., cognitive reflection, working memory, intelligence).

The BNT has been validated for the prediction of risk literacy (e.g., accurate interpretation and comprehension of everyday risks). However, relatively few studies have investigated the relationship between performance on the test and in other types of tasks measuring superior performance (for some related examples see Di-Girolamo et al., 2014; Garcia-Retamero, Cokely, Wicki, & Hanson, 2014; Riege & Teigen, 2013; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012). Theoretically, the test should predict performance across the same wide range of domains as other numeracy tests (e.g., the correlation with the test by Schwartz et al., 1997 is around .5). However, unlike other numeracy tests, the BNT is designed to provide greater psychometric sensitivity among moderate-to-very-highly numerate individuals, such as highly educated participants and professionals. Nevertheless, there could be some threshold level of mathematical skill wherein decision-makers are competent enough to accurately interpret and perform all requisite calculations present in judgment and decision task stimuli. In the same way reading ability becomes less predictive of performance once one has achieved college level reading proficiency, numeracy's predictive power may wane or fail among very highly educated participants because they're all numerate enough. To investigate issues in psychometric sensitivity and predictive validity, along with an examination of some key underlying cognitive mechanisms, we conducted a series of two large studies of paradigmatic judgment and decision making tasks in very highly educated samples from the Netherlands.

1.4 Experimenting with public outreach

In 2012, following the publication of the Berlin Numeracy Test and the launch of www.RiskLiteracy.org, we were contacted by a journalist working for *de Volkskrant*—a national daily morning newspaper in Holland.[7] He was interested in details of the BNT for an article about the importance of statistics for decision making. Rather than include a direct link to RiskLiteracy.org, we asked if we could create a unique link to an experiment that would be included in the newspaper article. Along with allowing for the collection of data, the link would provide users with immediate feedback on their relative numeracy levels (e.g., an estimate of their overall risk literacy). Ultimately, with support from editors, technical support, internal review boards, etc., we created a brief online study that newspaper readers could participate in, hosted on the *de Volkskrant* website (ca. 5–8 minutes long). In turn, we provided participants with feedback on their initial performance (i.e., immediate feedback on their Berlin Numeracy Test scores) and later provided a general summary of results included in a second follow-up newspaper article along with additional learning resources.

At a later date, we were invited to take part in the Grand National Numeracy Survey in the Netherlands.[8] Again, one constraint was that our study needed to be very brief

[6] Sensitivity was poorest among students at an elite university in China. About 75% of those participants answered all questions correctly.

[7] We thank Hans van Maanen, editors, and technical support at *de Volkskrant*.

[8] We thank Han van der Maas, Marthe Straatemeire, and other colleagues and participating researchers with the Grand National Numeracy Survey.

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 1: Demographic data on reported occupation and education level in Study 1. Data represented as proportions.

		Education	
Occupation (proportion)		College degree	Masters/PhD
Banking/Finance	.04	.66	.41
Statistics/Math.	.06	.92	.73
Computer/Engineer.	.17	.78	.54
Humanities /Art	.05	.85	.68
Medicine/Health	.12	.86	.68
Management/Admin.	.12	.74	.49
Customer services	.01	.46	.24
Students	.08	.57	.34
Others	.36	.69	.47

and include the Berlin Numeracy Test with performance feedback. For both studies, we selected paradigmatic judgment and decision criterion tasks based on previous research. Each task was selected to provide a small but representative window (1–2 items) into central topics in judgment and decision making, presented either in the context of finance (i.e., gain/loss lotteries and intertemporal choice) or medicine/health (evaluating clinical trials with differing group sizes; subjective confidence in judgment). We also collected data on decision latencies using a relatively insensitive but convenient response time metric (i.e., how long was the internet window open during financial decisions). Study 2 (Dutch National Numeracy Survey) provided a replication and extension of Study 1 (*de Volkskrant*) in which participants completed all the same tasks and also completed the numeracy test by Schwartz et al. (1997).

2 Study one: *de Volkskrant* newspaper study

2.1 Participants

About 4500 visitors responded to the newspaper article presented in *de Volkskrant* in 2012. After removing participants who did not complete the entire study, the final data set used for analysis included 3990 respondents, 64% of whom were male. The mean participant age was 48 years (SD = 13.5). Demographic data on reported education and occupational fields are presented in Table 1.

2.2 Materials, procedures, and hypotheses

All materials were presented in Dutch.[9] Data were collected using online survey software (unipark.de) with recruitment via a link hosted on the *de Volkskrant* website, which was included in both online and print versions of a newspaper article. Upon logging onto the website, participants were redirected to the online survey on the secure unipark server and were subsequently presented with an approved electronic informed consent for review and approval. Next participants read brief instructions and completed an adaptive version of the Berlin Numeracy Test, wherein participants were asked 2–3 questions that were selected based on the accuracy of their previous answers (i.e., correct answers led to harder questions, incorrect answers led to easier questions).

Participants were next presented with three tasks in a financial context on a new website page. Two questions were simple lotteries taken from previous research (Cokely & Kelley, 2009; Frederick, 2005; see Appendix A). For example, participants were asked whether they would prefer +/−€100 for certain or 75% chance of +/−€200 (i.e., in either gain or loss frame; see Appendix A for all material). The two lotteries were systematically counterbalanced and presented in randomized order (e.g., gain first, loss first). The third question on the page was an intertemporal choice that has previously been shown to track individual differences in cognitive reflection (Frederick, 2005), namely, "which option would you prefer: €3400 this month or €3800 next month". Overall, we hypothesized that more numerate participants would make more normatively superior choices, showing smaller framing effects (i.e., approximating expected value) and preferring more normative discounting rates. Consistent with previous findings (Cokely & Kelley, 2009), we predicted that total decision latency on the website page featuring all three questions (i.e., a rough proxy for total deliberation) would be related to numeracy and superior performance. We further hypothesized that decision latency would partially mediate the relationship between the Berlin Numeracy Test and superior financial decision making.

For tasks in the medical context, we presented a modified medical scenario known to be associated with denominator neglect, taken from Okan, Garcia-Retamero, Cokely, & Maldonado (2012; see also Garcia-Retamero & Galesic, 2009).[10] Participants were asked to rate the effectiveness of a drug based on fictional results of a clini-

[9]We thank Dafina Petrova and several colleagues at *de Volkskrant* for facilitating translation of the informed consent and basic test materials. The BNT translation employed in Cokely et al. (2012) was used in this study.

[10]We did not assess judgment latencies because both the metacognitive judgment and the denominator neglect question were presented on the same page and we could not control for differences in reading times (e.g., there was a paragraph describing the clinical trials required for the initial judgment, see Appendix).

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

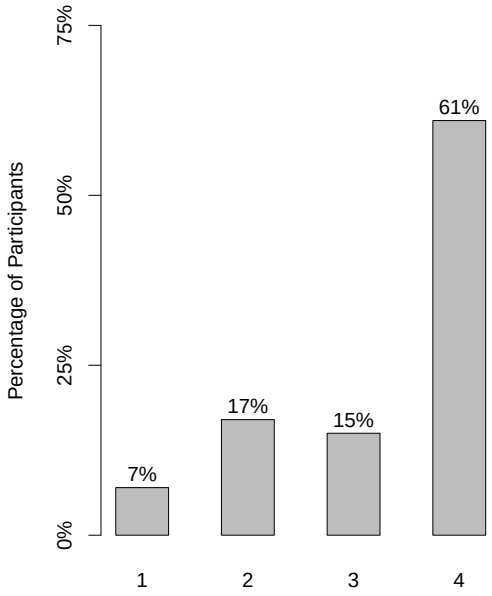
cal trial of a drug designed to reduce heart attack in which "... 80 out of 800 people who did not take the drug died after a heart attack, compared to 16 out of 100 people who took the drug". We then asked participants "How helpful was this drug" on a 7 point scale. Those participants who accurately estimated the ratios would find that 10% of those who did not take the drug died, compared to 16% of those who did take the drug. Thus, the drug was not effective. We hypothesized that less numerate participants who focused on factors like the absolute number of patients who died (16 died if they took the drug versus 80 died if they didn't take the drug) would come to a different, non-normative conclusion (i.e., show denominator neglect bias). Next we asked all participants how confident they were in their previous helpfulness judgment, using a 7 point scale where 1 indicated not at all confident and 7 indicated very confident (see Appendix A and B for exact materials). We hypothesized that accuracy and confidence should have a non-linear relationship. Those who do not effectively self-monitor would tend to be very confident in their inaccurate judgments (i.e., unskilled and unaware phenomena; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). However, as the quality of one's metacognitive self-assessment increased so too should accuracy (i.e., a curvilinear relationship). We further hypothesized that the relationship between scores on the Berlin Numeracy Test and accuracy would be partially mediated by people's ability to accurately assess their own judgment (i.e., degree of overconfidence).

Following all performance tasks, participants were presented with a demographics questionnaire, including questions on their sex, age, education, and professional field. Participants were presented with information about their numeracy score and their relative estimated risk literacy (see RiskLiteracy.org for examples). Finally, participants were thanked and debriefed.

2.3 Results

Our sample from *de Volkskrant* showed a much higher average score on the Berlin Numeracy Test than other past samples of college educated participants, including samples of practicing surgeons in the EU (Figure 1). The high scores were anticipated because participants were (a) reading newspaper articles about statistics for leisure, (b) motivated to log on and test their numeracy skill, and (c) highly educated (72% of the participants reported earning at least one college degree and 50% reported having more than one, see Table 1). Overall 61% of the sample answered all questions correctly and 76% scored above the median on BNT test (see Figure 1). Also, consistent with previous findings, men (t (3960) = 5.9, p = .001) and younger adults (t (3620) = 6.11, p = .001) tended to score slightly higher on the BNT than women and older adults.

Figure 1: Percentage of participants at each level of numeracy as measured by the Berlin Numeracy Test. The four levels represent estimated quartile norms for educated samples from industrialized countries.



2.3.1 Financial choices

A linear regression was used to examine the relation between the BNT and overall score on all three financial choices (i.e., normative accuracy). Regression indicated that BNT was a moderate sized, significant single predictor of normatively superior financial decisions (F (1, 3986) = 282.7, β = .26, p < .001, R^2 = .07—β represents the standardized regression weight). Individuals who scored higher on the BNT made more normatively superior decisions than those with lower BNT scores (see Figure 2).

A significant positive relationship was observed between education and BNT (r (3988) = .21, p =.0001) and between education and performance (r (3988) = .16, p = .0001). To examine further the role of education and other potentially influential variables, we constructed a series of hierarchical linear regression models with gender and age (model 1), education (model 2), and BNT (model 3) as predictors of overall financial decisions. The full model (model 3) significantly predicted performance on the three financial decisions (R^2 = .11, F (4, 3655) = 116.53, p < .001). The BNT remained a moderately sized predictor of superior financial choices with education, age, and gender included ($R^{2\text{change}}$ = .04, β = .20) (see Table 2).[11]

We recorded the time each participant spent on the webpage with the financial decisions as a rough proxy for over-

[11] When only age and gender, not education, were included along with BNT, the coefficient for BNT was little changed (R^2 = .10, $R^{2change}$ = .05, β = .23, p < .001).

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 2: Hierarchical regression predicting performance on financial decision tasks.

Models and variables	β	R	R^2	R^2 change	F change
Model 1		0.23	0.05	0.05	100.57**
Gender	−0.23**				
Age	−0.09**				
Model 2: Educ. added		0.27	0.08	0.25	97.3**
Gender	−0.22**				
Age	−0.09**				
Education	0.16**				
Model 3: BNT added		0.34	0.11	0.04	150.01**
Gender	−0.20**				
Age	−0.06**				
Education	0.11**				
BNT	0.20**				

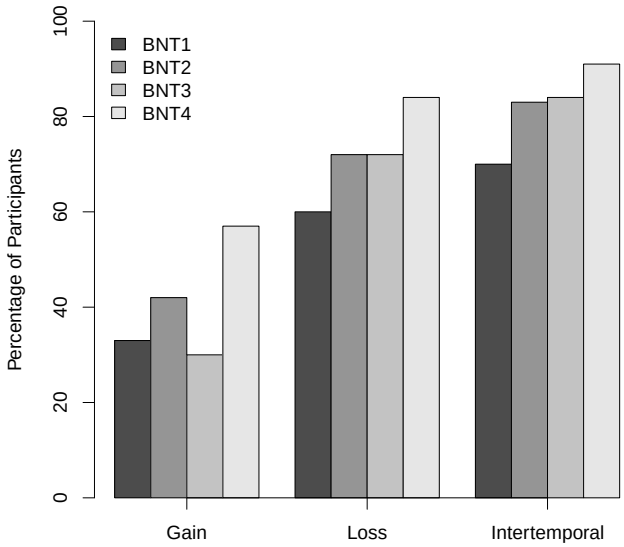
Note: ** p < .001.

all deliberation during risky decision making. We found a positive relationship between decision latency and the BNT (r (3988) = .074, p < .001), between decision latency and superior financial decisions including all three decisions in aggregate (r (3988) = .068, p < .001) and between BNT and superior financial decision (r (3988) = .26, p < .001). A mediation model was developed (Preacher & Hayes, 2004). The direct effects of BNT on performance (path c) and the indirect effects of BNT on performance via decision latency (i.e., deliberation) are presented in Figure 3. Results reveal significant and positive direct effects (path a) of the BNT on latency (B = .09, se = .01, p < .001), and of latency (path b) on superior decision making (B = .09, se = .02, p < .001). An examination of the specific indirect effects (path c') indicates that the relationship between the BNT and superior decision making was partially mediated by decision latency (B = .208, SE = .01, p < .001; *Sobel test value* z = 4.04, p < .001). Note that, although the relationship is significant, the magnitude is modest and smaller than in past studies. We speculate the difference reflects psychometric limits of our rough decision latency assessment (i.e., total website page viewing time for only three choices) as well as restriction of range in our very highly educated sample.

2.3.2 Financial lotteries

Regression was used to examine performance on the two financial lottery questions. The BNT was related to superior risky decision making in the gain frame (r (3988) = .17, p =.001) and in the loss frame (r (3988) = −.17, p < .001).[12] Linear regression indicated that BNT predicted overall performance on combined (gain and loss) decisions (R^2 = .05, F (1, 3986) = 207.4, β = −.22, p < .001). To compare predictive power relative to other potentially influential variables, we constructed hierarchical linear regression models with gender and age (model 1), education (model 2), and BNT (model 3) as predictors of overall risky lottery decisions. The BNT coefficient was largely unchanged when age, gender and education were included (β = −.18, p < .001).

Figure 2: Percentage of respondents at each level of the Berlin Numeracy Test who made more normatively superior financial decisions.



2.3.3 Intertemporal choice

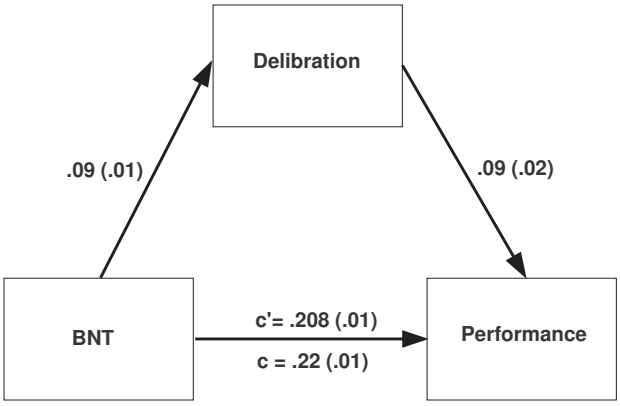
For the intertemporal time preference question 87% of the total sample made normatively superior choices (i.e., preferred €3800 next month rather than €3400 this month). Linear regression indicated BNT was a significant predictor of time preferences (R^2 = .03, F (1, 3986) = 115.6, β = .17, p < .001).[13] BNT remained a predictor for intertemporal choices when age, gender, and education were included in a linear regression (β = .12, p < .001).

[12] We also performed a non-parametric chi-square test to examine the relationship between numeracy and framing effects; we found that highly numerate participants selected more normatively superior decisions for gains (52% vs 39%, χ(1) 46.1, p <.0001) and for losses (82% vs 68%, χ (1) 75.7, p < .0001) as compared to less numerate participants.

[13] We also conducted non-parametric chi-square test; results indicated that highly numerate participants made more patient choices (90% vs 79%, χ (1) 72.4, p <.0001) as compared to less numerate participants.

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Figure 3: Deliberation (i.e., decision latency) partially mediated the relationship between the BNT and superior financial decision making. The Sobel test of mediation was significant, z = 4.04, p < .0001. Unstandardized path coefficients are shown with standard errors in parenthesis.



2.3.4 Medical judgment

Seventy four percent of the total sample made normative judgments on the medical judgment task.[14] Linear regression indicated that the BNT was a significant single predictor of accuracy of the medical judgments (R^2 = .04, F (1, 3986) = 180.86, β = .21, p < .001). Hierarchical linear regression models examined potentially influential variables of age and gender (model 1), education (model 2), and BNT (model 3). The BNT coefficient was essentially unaffected by the inclusion of these variables ($R^{2\ change}$ = .034, β = .19, p < .001; see Table 3).

2.3.5 Confidence

We analyzed the relationship between the BNT, medical judgment accuracy, and confidence in judgment. We found a positive relationship between the BNT and confidence (r (3988) = .09, p < .001). We also found a positive relationship between confidence and accuracy of medical judgments (r (3988) = .26, p < .0001). Curve estimation indicated that the relationship between confidence and accuracy was curvilinear and that a quadratic model fit better than the linear model (R^2 for quadratic = .12, R^2 for linear = .07, $R^{2change}$ = .05). Figure 4 shows the best fitting models. Note that both the decrease in confidence as accuracy increased from 1 to 4 and the increase as accuracy increased from 4 to 7 were highly significant (p < .001). This result suggests the presence of an unskilled and unaware type effect (i.e., participants were highly overconfident at low levels of accuracy yet relatively well calibrated at higher levels of accuracy). As numeracy increased, the total number of participants with perfect calibration also increased, while the proportion of participants who were overconfident decreased (Table 4). We also found that the strength of the relationship between confidence and accuracy increased at higher levels of numeracy, while the strength of the curvilinear model decreased, as did the difference between the linear and curvilinear models (Table 4). These results suggest that participants who are more numerate also tend to have better judgment calibration (e.g., less overconfidence). Path analysis indicated that confidence partially mediated the relationship between BNT and accuracy (Table 5).

[14]Choosing 1 on a 7-point scale, in which 7 means that the drug is very effective and 1 means drug is not effective.

Table 3: Hierarchical regression predicting performance on the medical judgment task.

Models and variables	β	R	R^2	R^2 change	F change
Model 1		0.022	0	0	0.89
Gender	−0.02				
Age	−0.002				
Model 2: Educ. added		0.091	0.008	0.008	28.78**
Gender	−0.02				
Age	0.00				
Education	0.09**				
Model 3: BNT added		0.205	0.042	0.034	128.34**
Gender	0.001				
Age	0.025				
Education	0.045*				
BNT	0.19**				

Note: *p < .05, **p < .001.

2.4 Study 1 discussion

Taken together the results of Study 1 indicate that even in very highly educated and highly numerate community samples (Table 6) the Berlin Numeracy Test is a robust predictor of paradigmatic financial and medical judgment and decision making. Results also indicate that the numeracy test predicts superior performance in part because it predicts differences in metacognitive processes, including differences in deliberation (as evidenced by decision latencies) and differences in the quality of one's self-assessment (as evidenced by differences in overconfidence).

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 4: Proportion of participants who had perfect calibration or were overconfident at each level of numeracy. Results of accuracy regressed on confidence at each level of BNT are also presented.

BNT levels	Prop. perfectly calibrated	Prop. overconf.	R^2 linear	R^2 quadratic	R^2 linear − R^2 quad.
BNT=1	.48	.27	0.03*	0.19**	−0.16
BNT=2	.56	.21	0.07**	0.18**	−0.11
BNT=3	.61	.15	0.05**	0.15**	−0.10
BNT=4	.65	.08	0.08**	0.13**	−0.05

Note: * p < .01, ** p < .001.

Table 5: Mediation through MEDCURVE (Hayes & Preacher, 2010), indirect effects of BNT on accuracy through confidence judgments.

X Values (BNT)	ab (indirect effect)	SE	95% conf. interval
2.3 (–1 SD)	.079	.014	.053–.110
3.3 (Mean)	.084	.016	.055–.115
4.3 (+1 SD)	.088	.018	.057–.120

Note: The table displays results of a medcurve meditational analysis at the mean BNT score and at BNT scores +/– 1 standard deviation from the mean. Indirect effects (i.e., mediation) of the BNT-to-performance relation via confidence judgments are shown to be significant with ab indirect effect coefficients and confidence intervals that do not include zero points.

3 Study 2: Data from the Dutch Grand National Numeracy Survey

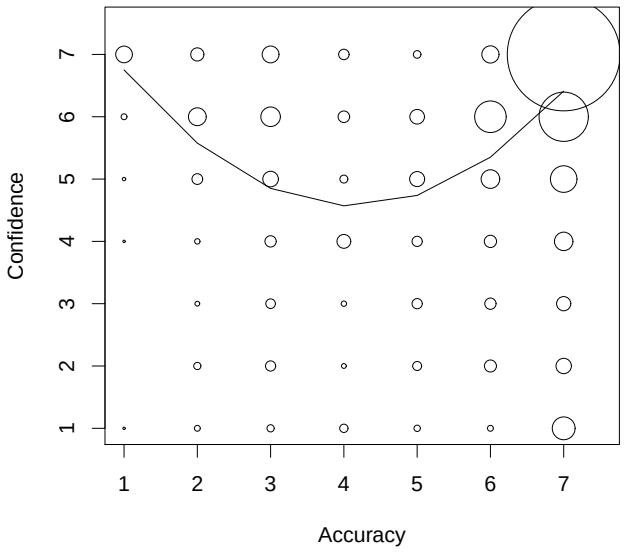
In Study 2 we sought to extend results from Study 1 by comparing the predictive performance of the Berlin Numeracy Test with another commonly used brief numeracy test, namely the Schwartz et al. (1997) three item numeracy test.

3.1 Method

3.1.1 Participants

Data were collected in Holland via an online link included as part of Dutch Grand National Numeracy Survey and associated outreach efforts. The data included 1418 participants with a mean age of 44 years (*SD* = 15). Fifty two percent of the sample was male. Thirty percent of the sample had at least one advanced graduate degree.

Figure 4: Curvilinear relationship between accuracy and confidence. High levels of overconfidence at low levels of accuracy (i.e., lower numbers on the x-axis) become more calibrated at higher levels of accuracy. Circle areas represents the proportion of respondents in each response category.



3.1.2 Material and procedure

All materials and procedures in Study 2 were identical to those used in Study 1 except that we included the Schwartz et al.'s (1997) three item numeracy test immediately before the adaptive Berlin Numeracy Test items.

3.2 Results and discussion

About 38% of the sample scored perfectly on the Berlin Numeracy Test (a score of 4) and 57% of the sample scored above the median point on the BNT (see Figure 5). This suggests that the Study 2 sample was more numerate than the educated samples used to norm the Berlin Numeracy Test yet was considerably less numerate than the sample from Study 1 (76% of which were above the median; see Figure 1).

Analyses followed those presented in Study 1. Linear regression indicated that the BNT predicted superior performance on combined financial decision tasks (β = .24; R^2 = .06, F (1, 1417) = 83.88, p < .001), medical judgments (β = .22; R^2 = .05, F (1, 1417) = 72.18, p < .001), and confidence judgments (β = .23; R^2 = .053, F (1, 1417) = 79.76, p < .001). A series of sets of hierarchical linear

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 6: Overall performance on medical judgments, financial decisions and BNT.

Profession	N	Medical judgments	Financial decisions	BNT
Banking/Finance	139	0.85	0.86	0.79
Statistics/Math.	256	0.90	0.90	0.92
Computer/Eng.	681	0.91	0.89	0.87
Humanities/Art	212	0.89	0.83	0.80
Medicine/Health	459	0.91	0.84	0.83
Mgmt./Admin.	467	0.90	0.84	0.81
Cust. services	67	0.86	0.83	0.74
Students	306	0.88	0.84	0.84
Others	1430	0.89	0.84	0.80

regression models with gender and age (model 1), education (model 2), and BNT (model 3) as predictors of financial decision were used to estimate independent contributions of each factor. In a model including age, gender, and education, the BNT provided unique predictive power for financial decisions ($R^{2\text{change}} = .03$, $\beta = .18$, $p < .001$)[15] (see Table 7). In a model including age, gender and education the BNT was also a good predictor of superior medical judgments ($R^{2change} = .025$, $\beta = .17$, $p < .001$) and confidence judgment ($R^{2\ change} = .036$, $\beta = .20$, $p < .001$; see Tables 8 and 9 for full model). As in Study 1, the BNT coefficient was reduced only a little by the addition of the other predictors.

We again found a curvilinear relationship between accuracy of medical judgments and confidence (R^2 for quadratic = .15, as compared to R^2 for linear = .08, $R^{2change}$ = .074). As numeracy increased, the total number of participants with perfect calibration also increased, while the proportion of participants who were overconfident decreased (Table 10). We also found that the strength of the relationship between confidence and accuracy tended to increase at higher levels of numeracy, while the strength of the curvilinear model tended to decrease, as did the difference between the linear and curvilinear models (Table 10). These results suggest that participants who are more numerate also tend to be better at assessing the accuracy of their judgments (e.g., less overconfidence). Path analysis indicated that confidence partially mediated the relationship between BNT and accuracy.

We analyzed the relationship between decision latency, the BNT, and superior financial decisions as in Study 1.

[15] We also constructed hierarchical linear regression models without entering the education variable into the model (gender and age [model 1] and BNT [model 2]). Excluding education, we found that the model was still a relatively good predictor of superior performance ($R^2 = .09$, R *change* = .035, $p < .001$; $\beta = .19$).

Table 7: Hierarchical regression predicting performance on financial decision tasks (Study 2).

Models and variables	β	R	R^2	R^2 change	F change
Model 1		0.23	0.05	0.05	34.89**
Gender	−0.22**				
Age	−0.07*				
Model 2: Educ. added		0.26	0.065	0.013	18.15**
Gender	−0.22**				
Age	−0.08*				
Education	0.12**				
Model 3: BNT added		0.3	0.09	0.03	37.67**
Gender	−0.18**				
Age	−0.047				
Education	0.078*				
BNT	0.18**				

Note: * p < .05; ** p < .001.

We found that time spent on financial decisions was positively related to performance on financial decisions (β = .09, p = .001). However, the relation between BNT and time was not quite significant (β = .05, p = .08), and our mediational analysis indicated a non-significant trend toward partial mediation (*Sobel test of mediation,* z = 1.43, p = .15). We speculate that this reflects the same psychometric limitations noted in Study 1 (i.e., limited webpage decision latency assessment sensitivity, restriction of range). We note that partial mediation has been seen in other studies (e.g., Study 1, Barton, Cokely, Galesic, Koehler, & Haas, 2009; Cokely & Kelley, 2009; Woller-Carter et al., 2012).

3.3 Psychometric analysis

Regression analysis indicated the Schwartz et al.'s (1997) test was a robust single predictor of financial decisions (β = .20; R^2 = .04, F (1, 1417) = 56.99, $p < .001$) and medical judgments (β = .17; R^2 = .03, F (1, 1417) = 40.04, $p <$ .001). Additional analyses indicated that the BNT doubled the unique predictive power of the Schwartz et al.'s (1997) test for both superior financial and medical decisions (Table 11). A hierarchical linear regression examined the potential additive effects with models of BNT (model 1) and BNT and Schwartz et al.'s (1997) (model 2). Adding the Schwartz et al.'s (1996) test to the BNT provided a modest significant improvement in the predictive power for com-

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 8: Hierarchical regression predicting performance on medical judgment task (Study 2).

Models and variables	β	R	R^2	R^2 change	F change
Model 1		0.07	0.005	0.005	3.13*
Gender	−0.04				
Age	−0.06*				
Model 2: Educ. added		0.15	0.023	0.018	23.57**
Gender	−0.04				
Age	−0.07*				
Education	0.135**				
Model 3: BNT added		0.22	0.05	0.025	33.96**
Gender	0.001				
Age	−0.04				
Education	0.10**				
BNT	0.17**				

Note: * p < .05; ** p < .001.

Table 9: Hierarchical regression predicting performance on subjective confidence task (Study 2).

Models and variables	β	R	R^2	R^2 change	F change
Model 1		0.13	0.02	0.02	11.67**
Gender	−0.13**				
Age	−0.01				
Model 2: Educ. added		0.195	0.04	0.02	26.26**
Gender	−0.13**				
Age	−0.02				
Education	0.14**				
Model 3: BNT added		0.27	0.07	0.035	49.24**
Gender	−0.09*				
Age	0.02				
Education	0.10**				
BNT	0.20**				

Note: * p < .05; ** p < .001.

bined financial decisions ($R^{2\text{change}}$ = .015, β = .13) and for medical judgments ($R^{2\text{change}}$ = .01, β = .10; see Table 13). Following Cokely et al. (2012) we combined the BNT and Schwartz et al.'s (1997) measures together to generate a composite BNT-S score (see Figure 6). As would be expected, the BNT-S score showed considerable skew (Figure 6) yet was a robust predictor of superior financial decisions (β = .27; R^2 = .07, F (1, 1417) = 108.03, p < .001), and medical judgments (β = .24; R^2 = .06, F (1, 1417) = 86.39, p < .001).

4 General discussion

In two large studies conducted with very highly educated samples, the Berlin Numeracy Test was found to be a robust independent predictor of superior judgment and decision making across risky decisions, temporal discounting, class-inclusion illusions (i.e., denominator neglect), and metacognitive judgments (median unique β = .19). The Berlin Numeracy Test doubled the predictive power of the well-established test by Schwartz and colleagues (1997), predicting performance in samples with numeracy scores that were notably higher than those observed in surgeons and medical students (Garcia-Retamero et al., in press, 2014). To put the current observed predictive strength into perspective, the link between the single predictor BNT and overall task performance is stronger than estimates of the link between gender and observed risk-taking behavior. The observed predictive power is about as strong as the meta-analytic estimate of the effect of ibuprofen on pain reduction (Meyer et al., 2001; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). It is noteworthy that the relationship was observed despite conditions of extreme restriction of range (i.e., the use of very highly educated samples) and non-ideal measurement conditions (e.g., few criteria). The current findings suggest that the predictive power of numeracy should tend to be significantly stronger in more diverse samples (e.g., in the general population, among college students), as found in other studies (Cokely et al., 2009, 2012). The current results also provide some of the first evidence that among very highly numerate participants, metacognitive processes continue to partially drive the ability-to-performance relationship (i.e., deliberation and confidence). These results converge with others indicating that the link between numeracy and superior judgment and decision making is not simply a function of differences in "doing the math".

4.1 Numeracy and metacognition

As detailed in the introduction, the theoretical construct of numeracy is multifactorial including (1) a practical understanding of numbers and mathematical procedures, and (2) the skills necessary for effective problem solving and self-regulated learning (e.g., metacognition and thinking about thinking; Flavell, 1979; Garofalo & Lester, 1985;

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 10: Proportion of participants who had perfect calibration or were overconfident at each level of numeracy. Results of accuracy regressed on confidence at each level of BNT are also presented.

BNT levels	Prop. perfectly calibrated	Prop. overconf.	R^2 linear	R^2 quadratic	R^2 linear − R^2 quad.
BNT=1	.32	.29	0.04^{*}	0.22^{**}	−0.18
BNT=2	.45	.26	0.07^{**}	0.18^{**}	−0.11
BNT=3	.56	.16	0.14^{**}	0.30^{**}	−0.16
BNT=4	.62	.16	0.04^{**}	0.11^{**}	−0.07

Note: * p < .05; ** p < .001.

Table 11: Unique predictive power of the two numeracy tests for predicting risky decisions. Standardized beta coefficients presented.

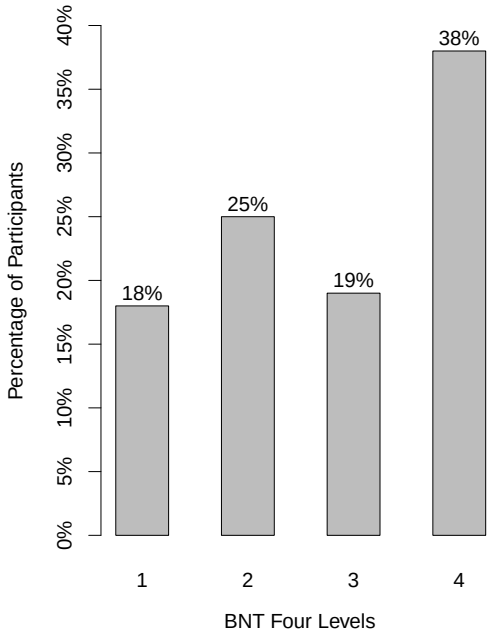
	Financial decisions	Medical judgments
BNT	$.19^{**}$	$.19^{**}$
Schwartz	$.13^{**}$	$.10^{**}$

Note: ** p < .001.

Lucangeli & Cornoldi, 1997; see also Dunlosky & Metcalfe, 2009). Numeracy tests appear to predict a wide range of behavior because they simultaneously assess both mathematical knowledge and the metacognitive processes involved in effective thinking (Halpern, 1998; Schoenfeld, 1992; Schraw, 1998; but for related theory in decision making see Baron, 1985, 2008; Baron, Badgio, & Gaskins, 1986; Stanovich, 2012; Stanovich, West, & Toplak, 2011; Toplak, West, & Stanovich, in press a, in press b). For example, in the current studies, we observed links between numeracy, confidence, deliberation, and superior performance. Because most participants were highly numerate, the differences in performance do not likely reflect differences in the availability of requisite mathematical skills. Nearly all participants were numerate enough to accurately calculate all expected values, discount rates, and relative proportions. Differences are also unlikely to reflect variation in levels of short-term motivation or task goals, as all participants volunteered and logged-on so they could test their numeracy. Rather than differences in goals, motivation, or minimum mathematical understanding, the observed performance differences appear to be more metacognitive in nature.[16] Those participants

[16]Appropriate cognitive representations, rather than explicit math skills, can also play a role in superior performance, as can be seen with the influence of simple visual aids that eliminate large performance differences between more and less numerate participants (Garcia-Retamero & Cokely, 2011, 2013, in press; see also Gigerenzer, Gaissmaier, Kruz-Mickle, Schwartz, Woloshin, 2007, and Peters et al., in press).

Figure 5: Levels of numeracy in a Dutch community sample (n = 1418). Data collected as part of the Dutch Grand National Numeracy Survey.



who had a more accurate subjective sense of their judgment performance (i.e., estimated confidence) and those who tended to spend more time deliberating during decision making tended to perform better. While there are likely many other important metacognitive and numeracy-related skills at work (Peters, 2012; Peters, Meilleur, & Tompkins, in press; Reyna & Farley, 2006; Reyna et al., 2009), the current data accord with previous research suggesting that deliberation and accurate self-monitoring often play central roles in domain-general superior judgment and decision making.[17]

4.2 Confidence and deliberation

The relationship between confidence and superior judgment and decision making is well-established (Bruine de Bruin et al., 2007), as are the relations between confidence, numeracy, and intelligence (Stankov, 2000). Research indicates that subjective estimates of confidence tends to derive from two factors—i.e., self-consistency (e.g., how reliably and quickly a judgment comes to mind)

[17]There are many theories about the causal mechanisms that give rise to the link between domain-general abilities and superior performance, as well as many compelling critiques of those theories (Baron, 1985; Kahneman, 2003, 2011; Reyna at al., 2009; Stanovich & West, 2000, 2008).

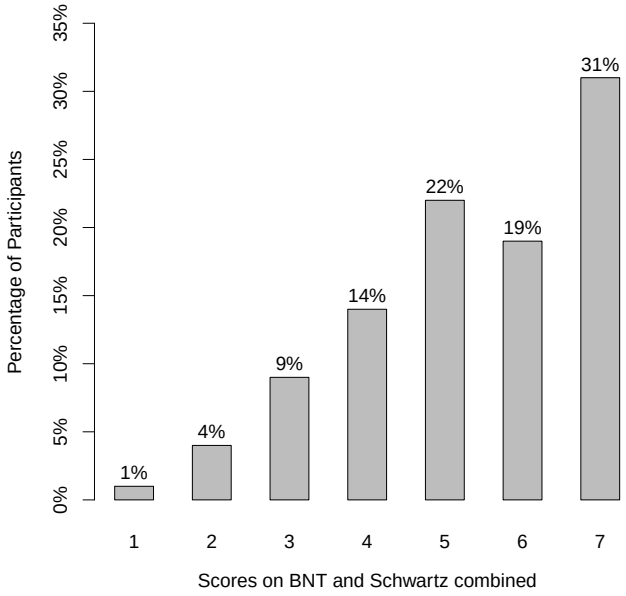
https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Table 12: Model comparison using BNT and Schwartz et al.'s (1997) measures as predictors.

	β	R	R^2	R^2 change	F change
Financial decisions					
Model 1		0.24	0.06	0.06	83.9**
BNT	0.24**				
Model 2		0.27	0.07	0.015	22.8**
BNT	0.19**				
Schwartz	0.13**				
Medical decisions					
Model 1		0.22	0.05	0.05	72.2**
BNT	0.22**				
Model 2		0.24	0.06	0.01	13.7**
BNT	0.19**				
Schwartz	0.10**				

Note: ** $p < .001$.

Figure 6: Levels of numeracy in a Dutch community sample using Schwartz et al.'s (1997) and BNT measures combined (n = 1418).



and the breadth of information that comes to mind (Koriat, 2012; see also Pleskac & Busemeyer, 2010).

Interestingly, although many studies treat confidence as a linear variable, here the relationship between confidence and performance was found to be curvilinear, resulting in an "unskilled and unaware" effect (Ehrlinger & Dunning, 2003; Ehrlinger et al., 2008).

Our confidence results accord with a variety of factor-analytic studies indicating that confidence self-assessment can operate as a domain-general skill that will be correlated with but also an independent predictor of general abilities, personality traits, and cognitive performance (Baker, 2010; Schraw, 2010; Stankov, 2000; Stankov & Lee, 2008). These results also accord with metacognitive theory suggesting confidence tends to be useful specifically because it is instrumental in self-regulation—i.e., the monitoring and control of cognition (Nelson & Narens, 1990; see also Metcalfe & Finn, 2008). For example, Koriat and Goldsmith (1996) describe how confidence accumulates and then is checked against a criterion in order to decide what type of information will be output in a memory task. Related studies of factors like "feeling of correctness" show that confidence-type judgments predict differences in information search and elaboration. In addition to predicting judgments about the correctness of one's answer, one's feeling of correctness tends to be related to "rethinking" times and the likelihood of changing one's initial answer during reasoning (Thompson, Prowse Turner, & Pennycook, 2011). These studies and others suggest that factors related to how one uses and assesses confidence may often be essential components determining the extent to which one deliberates during judgment and decision making (e.g., how much evidence does one require in order to feel confident in one's decision?).

The links between deliberation and various types of superior cognitive performance are also well-established. Deliberation is related to and can even cause differences in domain general cognitive abilities, such as intelligence and attentional control (Baron, 1978, 1985; Cokely, Kelley, & Gilchrist, 2006; Hertzog & Robinson, 2005; Stanovich, 2012). Consistent with the current results, deliberation is thought to be an essential component of rational thinking (e.g., reflectiveness and active open-minded thinking; Baron, 1985, 2008; Baron et al., 1986). Unfortunately, the current data do not provide process-level details about the content of deliberation in the current study.

Previous cognitive process tracing studies suggest that the observed differences in deliberation are not likely to result from differences in normative decision strategies.[18] Consider the protocol analysis conducted by Cokely and Kelley (2009) examining deliberative processes in simple risky lotteries. Although a pilot study indicated that most college students could perform the required math (e.g., "what is 3% of 7000"), less than 5% of their sample calculated expected value during decision making. Anal-

[18] For related experimental evidence see the study by Peters et al. (2006) showing that, while numeracy is related to superior performance, it is also predictably related to biases, reflecting the influence of heuristic processes (e.g., influenced by affective precision).

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

yses of formal decision models, reaction times, and retrospective memory reports indicated that the ability-to-performance relationship was fully mediated by large differences in heuristic-based deliberation and elaborative processing (Cokely & Kelley, 2009; see also Pachur & Galesic, 2013). Better risky decision making followed from differences in *how participants thought* about the decision (e.g., meaning-oriented elaborative processes such as imagining how the changes in wealth could affect one's life and how that might feel in contrast to others who treated the task as if it was just a game of chance; see also Reyna et al., 2009). Better risky decision making also followed from differences in *how much participants thought* about the decision (e.g., elaborating multiple reasons for each decision, transforming probabilities, and reframing outcomes). Similar results have been found in other protocol analyses, eye-tracking studies, and memory analyses used to examine some medical and economic judgments and decisions (Barton et al., 2009; Woller-Carter et al., 2012). Protocol analysis also suggests that, during move selection in chess, the systematic use of more deliberation tends to be associated with large performance advantages for both novices and experts (Moxley, Ericsson, Charness, & Krampe, 2012; see also Ericsson, Prietula, & Cokely, 2007).

We suggest that links between deliberation, confidence, and performance likely reflect a host of early selection metacognitive processes (Cokely & Kelley, 2009). Research shows that individuals who score higher on domain-general cognitive ability measures often spend more time preparing for tasks and also more elaborately process information, deliberatively building richer cognitive representations in long-term memory in order to provide better monitoring and control during subsequent task performance (Baron, 1978, 1985; Cokely & Kelley, 2009; Cokely et al., 2006; Ericsson & Kintsch, 1995; Hertzog & Robinson, 2005; Sternberg, 1977; Vigneau, Caissie, & Bors, 2005). As an analogy, in manufacturing one can improve the quality of goods sent to market by (a) improving inputs (e.g., higher quality materials and plans), (b) improving outputs (e.g., careful inspection and repair), or (c) doing both. In the metacognition literature these quality control efforts are referred to in terms of (a) early selection versus (b) late correction processes (Jacoby, Kelley, & McElree, 1999; Jacoby, Shimizu, Daniels, & Rhodes, 2005). Late correction processes attempt to detect and repair (e.g., System 2) the output of faulty automatic processes (e.g., System 1), such as biased intuitions. In contrast, early selection uses controlled processing (e.g., System 2) to generate goals, strategies, and mental contexts that qualitatively alter the output of automatic processes (e.g., System 1) before biased intuitions are generated (e.g., approaching the task more carefully).

To the extent that early selection metacognitive processes are recruited, they involve deliberation and elaborative encoding (e.g., contextualizing the problem by deeply thinking about the various aspects of the problem and their potential implications). This elaborative encoding causes information in working memory to be more robustly stored and represented in long-term memory, freeing-up limited attentional resources and creating more enduring and detailed problem representations (Cokely et al., 2006; Craik & Lockhart, 1972). Such representations may be similar in some important respects to those described in Fuzzy-Trace Theory as gist-based representations (e.g., one may use elaborative processing to build a more comprehensive intuitive representation). Ultimately, confidence calibration can be improved because some biased intuitions are never experienced and because more detailed representations provide more diagnostic cues for accurate cognitive monitoring (i.e., better quality evidence for monitoring; see Mitchum & Kelley, 2010). Note, however, that mere deliberation does not guarantee improved performance. Performance incentives that increase deliberation often fail to improve calibration or performance because participants tend to search for evidence that confirms their current beliefs (Koriat, Lichtenstein, & Fischhoff, 1980; see also Nickerson, 1998).[19] Improving calibration typically requires either changing task structures or training with individualized feedback. This type of training can lead to nearly perfect calibration. However, confidence will tend to be highly domain specific unless training also focuses on transferable metacognitive skills (e.g., practice using metacognitive heuristics such as searching for disconfirming evidence; Arkes, 1991).

4.3 Conclusions

Cognitive skills and abilities generalize only to the extent that similar elements of the skills are present on *training* and *transfer* tasks. Transfer requires shared elements (Thorndick & Woodworth, 1901; see also Blume, Ford, Baldwin, Huang, 2010). Many skills are highly domain-specific and so they are unrelated to performance outside a narrow band of expertise (e.g., surgical skill is not related to managerial decision making; Ericsson, Charness, Feltovich, & Hoffman, 2006; Ericsson et al., 2007). Numeracy is different. In the modern world, mathematical concepts are ubiquitous: Numeracy is an essential component of risk literacy and scientific thinking (Bruine de Bruin & Bostrom, in press; Cokely et al., 2012; Gigerenzer 2002; 2012). However, consistent with a large body of data, the current results suggest that numeracy tests don't simply predict use of abstract mathematics or normative decision strategies. Beyond the essential contributions of one's

[19] See Cokely and Kelley (2009) for a more detailed discussion of deliberative early selection versus late correction cognitive control processes.

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

mathematical competence, numeracy tests predict superior judgment and decision making because they assess (i) heuristic-based deliberation and metacognition (Cokely & Kelley, 2009; Cokely et al., 2012; see also Stanovich, 2012; reflectiveness, Baron, 1985), (ii) affective numerical intuition (Peters, 2012; Peters et al., 2006; Slovic, Finucane, Peters, MacGregor, 2002), and (iii) meaningful intuitive understanding (e.g., Reyna & Brainerd, 1991, 2005a; Reyna et al., 2009). More research is needed to investigate and model training and transfer across numeracy, metacognition, and decision tasks. For example: When does training numeracy improve metacognition? Why does training metacognition improve numeracy? What types of numeracy and metacognitive training improve decision making? To the extent that we develop a higher-fidelity understanding of underlying shared elements, we may be able to more efficiently reduce and anticipate non-adaptive judgment and decision making biases (e.g., intelligent tutoring systems, interactive risk communications).

References

Arkes, H. R. (1991). The costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*, 480–498.

Baker, S. F. (2010). *Calibration analysis within the cognitive and personality domains: Individual differences in confidence, accuracy, and bias.* (Unpublished doctoral dissertation). University of Southern Queensland, Australia.

Baron, J. (1978). Intelligence and general strategies. In G. Underwood (Eds.), *Strategies of information processing* (pp. 403–450). London: Academic Press.

Baron, J. (1985) *Rationality and intelligence*. New York, NY: Cambridge University Press.

Baron, J. (2008). *Thinking and deciding*. New York, NY: Cambridge University Press.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, vol. 3 (pp. 173–220). Hillsdale, NJ: Erlbaum.

Barton, A., Cokely, E. T., Galesic, M., Koehler, A., & Haas, M. (2009). Comparing risk reductions: On the dynamic interplay of cognitive strategies, numeracy, complexity, and format. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2347–2352). Austin, TX: Cognitive Science Society.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*, 1065–1105.

Bruine de Bruin, W., & Bostrom, A. (in press). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences. doi/10.1073/pnas.1212729110*

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

Cokely, E. T., Ghazal, S., Galesic, M., Garcia-Retamero, R., & Schulz, E. (2013). How to measure risk comprehension in educated samples. In R. Garcia-Retamero & M. Galesic (Eds.), T*ransparent communication of health risks: Overcoming cultural differences* (pp. 29–52). New York: Springer.

Cokely, E. T., Ghazal, S., & Garcia-Retamero, R. (in press). Measuring numeracy. In B. L. Anderson & J. Schulkin (Eds.), *Numerical reasoning in judgments and decision making about health.* Cambridge, UK: Cambridge University Press.

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 & Review*, *13*, 991–997.

Cokely, E. T., Parpart, P., & Schooler, L. J. (2009). On the link between cognitive control and heuristic processes. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2926–2931). Austin, TX: Cognitive Science Society.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making*, *25*, 331–351.

Del Missier, F., Mäntylä, T., Hansson, P., Bruine de Bruin, W., & Parker, A. M. (2013). The multifold relationship between memory and decision making: An individual differences study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1344–1364.

Di Girolamo, A., Harrison, G. W., Lau, M. I., & Swarthout, J. T. (2014). *Characterizing financial and statistical literacy* (No. 2013-04). Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University. Working paper.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Beverly Hills, CA: SAGE.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education and the workplace. *Psychological Science in the Public Interest*, *5*, 69–106.

Edwards, A., & Elwyn, G. (2009). *Shared decision-making in health care: Achieving evidence-based patient choice*. Oxford, Oxford University Press.

Ehrlinger, J., & Dunning, D. A. (2003). How chronic self-views influence (and mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*, 5–17.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*, 98–121.

Ericsson, K. A., Charness, N., Feltovich, P., & Hoffman, R. R. (2006). *Cambridge handbook of expertise and expert performance*. Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

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. 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*, 672–680.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911.

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

Garcia-Retamero, R., & Cokely, E. T. (2011). Effective communication of risks to young adults: Using message framing and visual aids to increase condom use and STD screening. *Journal of Experimental Psychology: Applied*, *17*, 270–287.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*, 392–399.

Garcia-Retamero, R., & Cokely, E. (in press). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases. *Journal of Behavioral Decision Making*. http://dx.doi.org/10.1002/bdm.1797.

Garcia-Retamero, R., Cokely, E. T., & Ghazal, S. (2014a). *Comparing risk literacy in 31 countries: New results from the Berlin Numeracy Test*. Manuscript in preparation.

Garcia-Retamero, R., Cokely, E.T., & Ghazal, S. (2014b). *Measuring subjective graph literacy. Manuscript in preparation*. Manuscript in preparation.

Garcia-Retamero, R., Cokely, E. T., Wicki, B., & Hanson, B. (2014). *Improving surgeons' risk literacy with visual aids*. Manuscript submitted for publication.

Garcia-Retamero, R., & Galesic, M. (2009). Communicating treatment risk reduction to people with low numeracy skills: A cross-cultural comparison. *American Journal of Public Health*, *99*, 2196–2202.

Garcia-Retamero, R., Wicki, B., Cokely, E. T., & Hanson, B. (in press). Factors predicting surgeons' preferred and actual roles in Interactions with their patients. *Health Psychology*. http://dx.doi.org/10.1037/hea0000061.

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education*, *16*, 163–176.

Gigerenzer, G. (2002). Calculated risks: How to know when numbers deceive you. New York: Simon & Schuster.

Gigerenzer, G. (2012). Risk literacy. In J. Brockman (Ed.), This will make you smarter: New scientific concepts to improve your thinking (pp. 259-261). New York: Harper Perennial.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients to make sense of health statistics. *Psychological Science in the Public Interest*, *8,* 53–96.

Ginsburg, L., Manly, M., & Schmitt, M. J. (2006). *The components of numeracy* [NCSALL Occasional Paper]. Cambridge, MA: National Center for Study of Adult Learning and Literacy (NCSALL).

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

Halpern, D. (1998). Teaching critical thinking for transfer across domains. *American Psychologist*, *53*, 449–455.

Hayes, A. F., & Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, *45*, 627–660.

Hertzog, C., & Robinson, A. E. (2005). Metacognition and intelligence. In O. Wilhelm & R. W. Engle, (Eds.), *Handbook of understanding and measuring intelligence* (pp. 101–121). CA: Sage publishers.

Huff, D., & Geis, I. (1954). *How to lie with statistics*. New York: Norton.

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope, (Eds.), *Dual process theories in social psychology* (pp. 383–402). New York, US: The Gilford Press.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*, 852–857.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NJ.: Macmillan.

Koriat, A. (2012). The self-consistence model of subjective confidence. *Psychological Review, 119,* 80–113.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of American adults: Results from the 2003 National Assessment of Adult Literacy* (NCES 2006- 483). Washington, DC: National Center for Education Statistics.

Låg, T., Bauger, L., Lindberg, M., & Friborg, O. (2013). The role of numeracy and intelligence in health-risk estimation and medical data interpretation. *Journal of Behavioral Decision Making*. http://dx.doi.org/10.1002/bdm.1788.

Levy, H., Ubel, P. A., Dillard, A. J., Weir, D. R., & Fagerlin, A. (2014). Health numeracy: The importance of domain in assessing numeracy. *Medical Decision Making*, *34*, 107–115.

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*, *25*, 361–381.

Lindskog, M., Winman, A., & Juslin, P. (2013). Are there rapid feedback effects on approximate number system acuity? *Frontiers in Human Neuroscience*, *7*, 270.

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.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lucangeli, D., & Cornoldi, C. (1997). Mathematics and metacognition: What is the nature of relationship? *Mathematical Cognition, 3*, 121–139.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review*, *15*, 174–179.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategy can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 699–710.

Moxley, J. H., Ericsson, K. A., Charness, N., & Krampe, R. T. (2012). The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition*, *124*, 72–78.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, vol. 26* (pp. 125–141). New York: Academic Press.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine*, *35*, 261–274.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.

OECD (2012). PISA 2009 Technical Report. Paris: OECD Publishing. Retrieved May 18, 2012 from http://dx.doi.org/10.1787/9789264167872-en.

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*, *25*, 390–401.

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*, 260–271.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.

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

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

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*, 435–448.

Peters, E., Meilleur, L., & Tompkins, M. K. (in press). Numeracy and the affordable care act: Opportunities and challenges. Chapter prepared for the Roundtable on Health Literacy, Institute of Medicine.

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

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of confidence, choice, and response time. *Psychological Review*, *117*, 864–901.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*, 717–731.

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science, 13,* 60–66.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making*, *7*, 332–359.

Reyna V. F., & Brainerd C. J. (1991). Fuzzy-trace theory and children's acquisition of mathematical and scientific concepts. *Learning and Individual Differences, 3*, 27–59.

Reyna V. F., & Brainerd, C. J. (2005a). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1–75.

Reyna V. F., & Brainerd, C. J. (2005b). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107.

Reyna V. F., & Farley F. (2006). Risk and rationality in adolescent decision-making: Implications for theory, practice, and public policy. *Psychological Science in the Public Interest*, *7*, 1–44.

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.

Riege, A. H., & Teigen, K. H. (2013). Corrigendum to "Additivity neglect in probability estimates: Effects of numeracy and response format." *Organizational Behavior and Human Decision Processes*, *121*, 41–52.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345.

Schapira, M. M., Walker, C. M., Cappaert, K. J., Ganschow, P. S., Fletcher, K. E., McGinley, E. L., Del Pozo, S., Schauer, C., Tarima, S., & Jacobs, E. A. (2012). The numeracy understanding in medicine instrument: A measure of health numeracy developed using item response theory. *Medical Decision Making*, *32*, 851–865.

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

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grows (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York: Macmillan.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*, 113–125.

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist, 45*, 258–266.

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.

Sheridan, S. L., Pignone, M. P., & Lewis, C. L. (2003). A randomized comparison of patients' understanding of number needed to treat and other common risk reduction formats. *Journal of General Internal Medicine, 18*, 884–892.

Sherrod, P. H. (2003). *DTREG: Predictive modeling software*. Software available at http://www.dtreg.com.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York: Cambridge University Press.

Stankov, L. (2000). Structural extension of a hierarchical view on human cognitive abilities. *Learning and Individual Differences*, *12*, 35–51.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, *100*, 961–976.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *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.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). Intelligence and rationality. In R. J. Sternberg & S. B.

Kaufman (Eds.), *Cambridge handbook of intelligence* (pp. 784–826). New York: Cambridge University Press.
Steen L. A. (1990). *On the shoulders of giants: New approaches to numeracy.* Washington, DC, US: National Academy Press.
Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84,* 353–378.
Toplak, M. E., West, R. F., & Stanovich, K. E. (in press a). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking and Reasoning.*
Toplak, M. E., West, R. F., & Stanovich, K. E. (in press b). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology.*
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107–140.
Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficacy of other functions. *Psychological Review*, *8*, 247–261.
Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York, NY: Springer.
Vigneau, F., Caissie, A. F., & Bors, D. A. (2005). Eye movement analysis demonstrates strategic influence on intelligence. *Intelligence, 34,* 261–272.
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*, 198–212.
Woller-Carter, M. M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2012). Communicating and distorting risks with graphs: An eye-tracking study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 1723–1727.
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.

Appendix A: Material

Financial decision tasks

Which option do you prefer?
a) €3400 This month b) €3800 next month

Which option do you prefer?
a) €100 for sure b) 60% chance of €250

Which option do you prefer?
a) 75% chance to lose €200 b) €100 surely lose

Medical and metacognitive judgment task

The new drug BENOFRENO, the risk of death from a heart attack reduced for people with high cholesterol. A study with 900 with high cholesterol showed that 80 of the 800 people who have not taken the drug deceased after a heart attack, compared with 16 of the 100 people who have taken the drug.

1.How beneficial was the Benofreno?
Not beneficial 1 2 3 4 5 6 7 **very beneficial**

2. How confident are you about your decision?
Not sure 1 2 3 4 5 6 7 **very sure**

4.3.1 Berlin Numeracy Test (BNT) four questions (used in adaptive format)

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the 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.

2. 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)?

3. 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?

4. 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?

4.3.2 Schwartz three numeracy questions

1. Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips?

2. In the Big Bucks Lottery, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1,000 people each buy a single ticket to Big Bucks?

3. In ACME Publishing Sweepstakes, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME Publishing Sweepstakes win a car?

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press

Appendix B: Screen shots of the decision tasks as presented in the experiment

Welke optie heeft uw voorkeur?

◯ € 3400 deze maand ◯ € 3800 volge de maand

Welke optie heeft uw voorkeur?

◯ € 100 met zekerheid ◯ 60% kans op € 250

Welke optie heeft uw voorkeur?

◯ 75% kans € 200 te verliezen ◯ € 100 zeker verliezen

voortzetten

Het nieuwe medicijn BENOFRENO moet de kans op overlijden aan een hartaanval verminderen voor mensen met een hoog cholesterolgehalte. Uit een studie met 900 mensen met een hoog cholesterolgehalte is gebleken dat 80 van de 800 mensen die het medicijn niet hebben ingenomen overleden na een hartaanval, vergeleken met 16 van de 100 mensen die het medicijn hebben ingenomen.

a) Hoe heilzaam was Benofreno?

	1	2	3	4	5	6	7	
helemaal niet heilzaam	◯	◯	◯	◯	◯	◯	◯	zeer heilzaam

b) Hoe zeker bent u van uw vorige antwoord?

	1	2	3	4	5	6	7	
helemaal niet zeker	◯	◯	◯	◯	◯	◯	◯	zeer zeker

voortzetten

https://doi.org/10.1017/S1930297500004952 Published online by Cambridge University Press