Journal of Behavioral Decision Making, *J. Behav. Dec. Making*, **30**: 26–38 (2017)
Published online 12 October 2015 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1906

Development and Validation of the Scientific Reasoning Scale

CAITLIN DRUMMOND[1]* and BARUCH FISCHHOFF[2]

[1]*Departmental of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA*

[2]*Department of Social and Decision Sciences, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA*

ABSTRACT

Scientific findings and innovations play an important role in a range of decisions faced by nonscientists, yet little is known about the skills that nonscientists need in order to read and evaluate scientific evidence. Drawing on research in public understanding of science, cognitive developmental psychology, and behavioral decision research, we develop an individual difference measure of *scientific reasoning skills*, defined as the skills needed to evaluate scientific findings in terms of the factors that determine their quality. We present the results of three studies assessing its psychometric validity. Our results indicate that the Scientific Reasoning Scale (SRS) is internally consistent and distinct from extant measures of scientific literacy. Participants with higher SRS scores are more likely to have beliefs consistent with the scientific consensus on potentially contentious issues, above and beyond education, political and religious beliefs, and scores on two widely used measures of scientific literacy. Participants with higher SRS scores also had better performance on a task requiring them to analyze scientific information. Our results suggest that the SRS provides a theoretically informed contribution to decoding lay responses to scientific results and controversies. Copyright © 2015 John Wiley & Sons, Ltd.

KEY WORDS individual differences; decision making; judgment; scientific reasoning; science communication

Scientific findings and innovations play an important role in everyday life, from the technologies we use to the evidence we cite in political debates. We exercise our ability to assess scientific evidence whenever we consult the package insert of a pharmaceutical, debate the risks of climate change, or read a news story on the economic impact of new healthcare laws. However, despite the importance of assessing scientific evidence in everyday life, it is unclear to what extent lay people possess and apply this ability. Surveys of the American population reveal wide variation in scientific knowledge (National Science Board, 2014), and sizeable fractions of the population hold beliefs at odds with scientific evidence on topics such as global warming and human evolution (Funk & Rainie, 2015).

Differences in how lay people interpret scientific evidence have been studied with various tests of *scientific literacy*, defined as, the "capacity to use scientific knowledge to…draw evidence-based conclusions" (OECD, 2003, pp. 132–133). These differences have been attributed to differences in general and technical education (Miller, 1998, 2002), information sources (Taber & Lodge, 2006), sociocultural background (Kahan et al., 2012), and cognitive styles such as actively open-minded thinking (Stanovich & West, 1997). Here, we focus on a largely unexplored factor potentially affecting lay interpretations of scientific evidence: *scientific reasoning skills*. To that end, we develop an individual difference measure meant to capture the ability to evaluate the quality of scientific evidence.

Such a measure could help to clarify whether people who reject the scientific consensus on an issue cannot assess the quality of scientific evidence or whether their interpretation of that evidence is biased by their belief in claims that most scientists would consider false or incomplete. To illustrate, consider a study by Downs, Bruine de Bruin, and Fischhoff (2008) assessing parents' mental models of childhood vaccination. They found that some skeptics' mental models of the processes determining vaccine safety included concerns about the quality of reporting for potential side effects. An assessment of pro-vaccination and anti-vaccination communications found that opponents of vaccination stirred these doubts, whereas proponents failed to address them. If skeptics have scientific reasoning skills, then their views might change if they receive this missing information regarding the reporting of vaccine side effects in a credible, comprehensible, and respectful way (Fischhoff, 2013). On the other hand, if skeptics do not have scientic reasoning skills, then lack of this knowledge is incidental to their opposition. In this case, the evidence will not speak for itself, and increasing vaccination rates will require more assertive means than providing additional information.

The measure developed here, the Scientific Reasoning Scale (SRS), draws on three areas of research. The philosophy and methodology of science informed our normative analysis of the skills needed to evaluate scientific evidence; public understanding of science research informed our descriptive analysis of the knowledge and skills measured by existing tests of scientific literacy; and cognitive developmental psychology informed our prescriptive analysis of how individuals develop the ability to think like a scientist. Our normative analysis defines scientific reasoning skills as those needed to evaluate scientific findings. After reviewing prior research, we describe the development of the SRS, followed by three studies assessing its psychometric validity.

Scientific literacy and scientific thinking

In 1978, the National Science Foundation invited Jon Miller and Kenneth Prewitt to develop a survey instrument

*Correspondence to: Caitlin Drummond, Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15206, USA. E-mail: cdrummond@cmu.edu

measuring public understanding of and attitudes toward science and technology (Miller, 2004). These questions were initially included on the National Center for Science and Engineering Statistics' Survey of Public Attitudes Toward and Understanding of Science and Technology. Since 2006, they have constituted a module in the General Social Survey. Responses are reported in the National Science Board's (NSB) biennial report *Science and Engineering Indicators*.

The survey measures two dimensions of scientific literacy, based on Miller's research. The Trend Factual Knowledge of Science Scale (TFKSS) assesses *knowledge of scientific concepts*, using true–false or multiple-choice questions, such as "True or False? The center of the Earth is very hot." (NSB, 2014). The items on this scale include both contemporary topics (e.g., lasers and antibiotics) and fundamental concepts, under the assumption that "the individual who does not comprehend basic terms like atom, molecule, cell, gravity, or radiation will find it nearly impossible to follow the public discussion of scientific results" (Miller, 1983, p. 38). The Understanding of Scientific Inquiry Scale (USIS) assesses *knowledge of the scientific method*, including experimental design, probability, and the scientific method, using multiple-choice and open-ended questions such as "Can you tell me, in your own words, what it means to study something scientifically?" (NSB, 2014). These two scales have been widely used in the field of public understanding of science (e.g., Allum, Sturgis, Tabourazi, & Brunton-Smith, 2008; Sturgis & Allum, 2004).

The SRS developed here extends this previous work in two ways. First, it derives its items from a normative analysis of the skills needed to demonstrate competence in evaluating scientific evidence. Some of these skills are measured by the USIS. However, our normative analysis identified topics addressing not just the agreed-upon process of conducting science, which is measured by the USIS, but also the potential weaknesses of scientific research arising from shortcomings in its processes or results. These weaknesses are essential to the critical reasoning needed to evaluate imperfect evidence and controversies and are not addressed by current tests.

Second, the SRS integrates research from cognitive developmental psychology on scientific reasoning skills. Cognitive developmental psychologists have identified two key dimensions underlying individuals' ability to perform scientific inference, based on experiments requiring individuals to act as amateur scientists, generating and testing hypotheses (Zimmerman, 2000). The first involves knowledge of scientific facts, of the sort taught in school (and tested on the TFKSS). For example, Vosniadou and Brewer (1992) examined children's understanding of the shape of the earth, with questions such as "Can you fall off the edge?" The second dimension concerns the "reasoning processes that permeate science: induction, deduction, experimental design, causal reasoning, concept formation, hypothesis testing, and so on" (Dunbar & Klahr, 2012, p. 611), using the "general skills implicated in experimental design and evidence evaluation" (Zimmerman, 2000, p. 104; see also Klahr & Dunbar, 1988; Wason & Johnson-Laird, 1972; Toplak & Stanovich, 2003).

These two dimensions parallel those identified by public understanding of science researchers. Nonetheless, the two literatures rarely cite one another, perhaps because cognitive researchers are concerned with the developmental process, whereas public understanding of science researchers are concerned with its products. Presumably, both knowledge and reasoning skills are needed to evaluate scientific evidence, especially when scientists report weak evidence or when strong evidence is attacked (e.g., for political reasons). Without basic knowledge, individuals cannot grasp the topics; without inferential skills, individuals cannot establish a critical perspective where scientists fail to realize or report the imperfections with their work (e.g., experiments with confounds and improper statistical methods). One must be able to think like a scientist in order to evaluate scientific research, and the SRS is designed to measure this ability.

The scientific reasoning scale

We develop and validate a theoretically informed individual difference measure of scientific reasoning skills. Our measure requires participants to apply their reasoning skills to analyze evidence (as in cognitive developmental research), but in the form of a survey allowing ready calculation of individual scores (as in public understanding of science research). After defining the domain of scientific reasoning skills, we develop items to represent it and then assess the convergent, divergent, and predictive validity of the resultant scale in terms of correlation of scores on it with nomologically related constructs including education, numeracy, cognitive reflection ability, and actively open-minded thinking. We also compare SRS scores with ones on two widely used scientific literacy scales, in order to assess the extent to which the SRS draws upon skills not measured by previous tests. Finally, we examine whether SRS scores predict the following: (1) beliefs consistent with the scientific consensus and (2) performance on a task requiring the evaluation of scientific information.

STUDY 1: A DEVELOPMENTAL STUDY

Studies 1a–1d used quantitative and qualitative data to develop test items for the SRS, using American adults drawn from Amazon.com's Mechanical Turk (MTurk). Study 1a had participants ($N = 397$) respond to an initial set of test items, in order to assess their reliability and validity. Study 1b refined those draft test items by eliciting qualitative feedback from small samples and iteratively updating test items. Study 1c replicated Study 1a with the revised items ($N = 395$) to test whether the revision improved the items. Study 1d replicated Study 1b, eliciting qualitative feedback on a small number of test items that exhibited poor psychometric properties in Study 1c.

STUDY 1A

Study 1a defined the domain of scientific reasoning, developed items to measure it, and administered these preliminary test items to a sample of American adults. We predicted that respondents with higher scores on these items would have high self-reported education, numeracy, cognitive reflection

DOI: 10.1002/bdm

ability, and scores on the existing scientific literacy measures. We also predicted that those individuals would be more likely to hold beliefs consistent with the scientific consensus.

Method

Development of scientific reasoning scale test items

We defined the domain of scientific reasoning in terms of concepts emphasized in research method textbooks and guidelines for assessing research quality. We first identified concepts from textbooks for research method courses (Hoyle, Harris, & Judd, 2002; Reis & Judd, 2000; Trochim & Donnelly, 2007), looking for concepts relevant to multiple disciplines. We sought ones related to both internal validity (i.e., quality) and external validity (i.e., relevance and generalizability). We grouped similar concepts and sought general versions of specific concepts. We then added concepts from two prominent methodologies for assessing the quality and validity of scientific research, those published by the Cochrane Group (and endorsed by the Agency for Healthcare Research and Quality) (Barkhordarian et al., 2013) and the Numeral Unit Spread Assessment Pedigree criteria for evaluating the strength of sciences (Funtowicz & Ravetz, 1990; see also Fischhoff & Davis, 2014). The resulting list, displayed in Table 1, had 20 concepts.

We next wrote test items for each concept, comprised of one-sentence to three-sentence scientific scenarios followed by a statement that respondents must evaluate as either true or false. Items were written so as to avoid specialized terminology. Each item had a unique scenario containing subject matter from a specific area of science (e.g., psychology, environmental science, marketing) to reflect the concrete circumstances in which scientific reasoning is required, but we avoided politicized domains, such as climate science.

Table 1. Factor analysis of the Scientific Reasoning Scale

Concept	Factor loading	Item–total *r*
Retained items		
Blind/double blind	0.46	0.47
Causality	0.59	0.53
Confounding variables	0.52	0.48
Construct validity	0.39	0.43
Control group	0.46	0.45
Ecological validity	0.63	0.55
History	0.60	0.54
Maturation	0.63	0.55
Random assignment to condition	0.60	0.53
Reliability	0.53	0.50
Response bias	0.51	0.48
Items omitted from final scale		
Attrition		
Directness[a]		
Empirical basis[a]		
Measurement error		
Methodological rigor[a]		
Natural variation		
Selection bias		
Statistical power		
Validation[a]		

Note: [a]Concepts drawn from Fischhoff and Davis (2014).

Participants

Four hundred and one American adults were recruited online via MTurk and paid $2.50. Four of them failed two data quality assurance questions drawn from Meade and Craig (2012) and were excluded from subsequent analyses. Based on self-reports, among the remaining 397 adults, 51% were male, 77% Caucasian, and 45.4% had at least a Bachelor's degree. Age was elicited in ranges: 17% reported being between 18 and 24; 44% between 25 and 34; 20% between 35 and 44; 11% between 45 and 54; 7% 55 or older.

Materials and procedure

Participants completed an online survey that included the 20 preliminary SRS items, administered in randomized order to each participant, followed by (in randomized order) the three-item Cognitive Reflection Test (Frederick, 2005), a two-item numeracy test (drawn from Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007), and the TFKSS (eight items; NSB, 2014) and USIS (three items; NSB, 2014). Descriptive statistics for these measures are reported in Table 2. Participants then answered questions about their political identity (on a 5-point scale from *very conservative* to *very liberal*) and their beliefs on five controversial scientific issues (Table 2). Beliefs about climate change were assessed with a one-item measure drawn from the Global Warming Six Americas scale (Leiserowitz, Maibach, Roser-Renouf, & Smith, 2011), using a 9-point scale from *extremely sure that global warming is occurring* to *extremely sure that global warming is not occurring*. Beliefs about the safety of genetically modified foods and vaccines were assessed on a 5-point scale from *very safe* to *very unsafe*. Beliefs about the existence of human evolution and the Big Bang were elicited with *true/false* questions.

Results

Scientific reasoning scale preliminary item analysis

On average, participants answered 14.7 of the 20 preliminary SRS questions correctly, $SD = 2.7$. On individual items, the percent correct ranged from 29% (random assignment to condition) to 95.7% (response bias). Tetrachoric correlations revealed that two items, reliability and directness, were negatively or weakly correlated with most of the other 18 questions. Further consideration suggested that these two items were poorly formulated, and they were removed from subsequent analyses. On average, participants answered 13.3 of the remaining 18 questions correctly, $SD = 2.7$ (Table 2).

Inter-item reliability

Cronbach's alpha for the preliminary 18-item scale was $\alpha = 0.65$. By way of perspective on this value and others reported later, alpha values tend be lower for scenario-based measures such as the SRS, because scenarios can present only one instance of the category that they seek to represent (Tangney, 1996), and for measures with heterogeneous domains, such as the present one. As a result, John and Benet-Martinez (2000) recommend that "if one wants to

DOI: 10.1002/bdm

Table 2. Descriptive statistics and internal consistency for the SRS, convergent measures, scientific consensus beliefs, and demographic characteristics

	Preliminary scale		Revised scale		Final scale			
	Study 1a (n = 397)		Study 1c (n = 395)		Study 2 (n = 270)		Study 3 (n = 294)	
Measure	*M* (*SD*)	α	*M* (*SD*)	α	*M* (*SD*)	α	*M* (*SD*)	α
Scale								
SRS	13.3 (2.7)	0.65	13.1 (3.0)	0.66	6.7 (2.6)	0.70	7.0 (2.6)	0.71
CRT	1.8 (1.2)	0.80	1.7 (1.2)	0.73	1.5 (1.3)	0.80	—	—
Numeracy	1.7 (0.54)	0.23	4.1 (1.2)	0.55	4.0 (1.2)	0.52	—	—
AOT	—	—	—	—	—	—	38.2 (6.2)	0.80
TFKSS	6.9 (1.2)	0.37	6.9 (1.2)	0.42	6.9 (1.2)	0.44	—	—
USIS	1.9 (0.85)	0.34	1.9 (0.90)	0.38	2.0 (0.91)	0.41	—	—
Scientific consensus belief								
Global warming	7.2 (1.8)	—	7.3 (1.8)	—	7.1 (1.9)	—	6.7 (2.0)	—
Genetically modified foods	3.0 (1.3)	—	3.0 (1.3)	—	3.1 (1.3)	—	3.2 (1.2)	—
Vaccines	4.2 (1.0)	—	4.1 (1.0)	—	4.0 (1.0)	—	4.2 (1.0)	—
Human evolution	0.78 (0.42)	—	3.7 (1.5)	—	3.8 (1.4)	—	3.6 (1.4)	—
The Big Bang	0.78 (0.42)	—	3.1 (1.7)	—	3.3 (1.7)	—	3.0 (1.7)	—
Demographic characteristic								
Political liberalism	3.4 (1.2)	—	3.4 (1.1)	—	3.4 (1.2)	—	3.4 (1.2)	—
Religiosity	—	—	2.4 (1.3)	—	2.5 (1.5)	—	2.3 (1.6)	—

Note: The preliminary SRS in Study 1a contained 18 items, the revised SRS in Study 1c contained 17 items, and the final SRS contains 11 items. All scientific consensus beliefs were elicited on a 5-point scale, with a 5 indicating strong belief in the consensus, except for global warming (9-point scale) and human evolution and the Big Bang in Study 1a (true/false). Political liberalism was elicited on a 5-point scale from *very conservative* to *very liberal*; religiosity was elicited on a 5-point scale from *strongly disagree* to *strongly agree* to the statement, "I consider myself to be a religious person."
SRS, Scientific Reasoning Scale; CRT, Cognitive Reflection Test; AOT, Actively Open-minded Thinking; TFKSS, Trend Factual Knowledge of Science Scale; USIS, Understanding of Scientific Inquiry Scale.

measure broader constructs, one should probably include a larger number of items to compensate for the greater content heterogeneity" (p. 347). On the other hand, alpha values tend to increase with both the number of items in a scale and with the average item intercorrelation (Cortina, 1993). We report both alpha values and number of items throughout.

Construct validity

Correlations of the preliminary 18-item scale with convergent measures confirmed our predictions[1] (Table 3). Individuals who scored higher on the preliminary SRS had higher scores on the Cognitive Reflection Test ($r = 0.36$, $p < 0.001$) and on the numeracy measure ($r = 0.28$, $p < 0.001$). Participants with higher SRS scores also had higher scores on the two scientific literacy tests (TFKSS: $r = 0.39$, $p < 0.001$; USIS: $r = 0.36$, $p < 0.001$).[2] Individuals with higher SRS scores reported higher levels of education ($r = 0.30, p < 0.001$). Among those who indicated completing at least some college ($N = 344$), individuals with higher SRS scores reported having taken more science classes ($r = 0.29$, $p < 0.001$).

Demographic correlates

Preliminary SRS scores were positively correlated with age ($r = 0.14$, $p < 0.01$), and with being more politically liberal ($r = 0.12$, $p = 0.02$) (Table 3). No gender difference was observed ($t(395) = 0.51$, $p = 0.61$).

Predictive validity

We used partial correlations to assess the relationship between preliminary SRS scores and scientific consensus beliefs, controlling for self-reported education and political liberalism, in order to reduce the risk that our results are driven by sample demographics. Table 4 presents these results. We find that individuals with higher preliminary SRS scores had higher scores on a composite belief measure, constructed by standardizing responses on each of the five beliefs and summing the standardized responses, with higher scores indicating a greater tendency to hold consensus beliefs ($r = 0.17, p < 0.001$). That pattern varied some across the specific beliefs. Individuals with higher SRS scores were more likely to believe that vaccines ($r = 0.22$, $p < 0.001$) were safe, consistent with the scientific consensus, but were no more likely to believe in global warming ($r = 0.06$, $p = 0.21$) or in the safety of genetically modified foods ($r = 0.08, p = 0.08$). Logistic regressions revealed that, controlling for self-reported education and political liberalism, higher SRS scores were associated with a greater likelihood of believing in the Big Bang ($B = 0.10$, $p = 0.04$), but not in human evolution ($B = 0.10, p = 0.07$).

Incremental predictive validity

To assess how the preliminary SRS differs from the two measures of scientific literacy, we regressed TFKSS and USIS

[1] Where our statistical tests involve a priori predictions, we report p-values uncorrected for Type I error in the text. However, because there were many of them, Tables 3 and 4 also report p-values corrected for Type I error using a Bonferroni correction.

[2] The small size of these correlations can be partly explained by the low reliabilities of the TFKSS and USIS (Table 2). The expected upper limit of the correlation for a measure is the square root of its reliability (John & Benet-Martinez, 2000). All reported correlations between the measures and others are below these expected upper limits.

DOI: 10.1002/bdm

Table 3. Bivariate correlations of the SRS with convergent measures and demographic characteristics

Variable	Preliminary scale Study 1a ($n = 397$)	Revised scale Study 1c ($n = 395$)	Final scale Study 2 ($n = 270$)	Final scale Study 3 ($n = 294$)
Convergent measure				
CRT	0.36***	0.38***	0.45***	—
Numeracy	0.28***	0.36***	0.40***	—
Education	0.30***	0.31***	0.30***	0.30***
AOT	—	—	—	0.41***
TFKSS	0.39***	0.40***	0.42***	—
USIS	0.36***	0.43***	0.43***	—
Demographic characteristic				
Age	0.14**	0.09	0.13*	0.18**
Political liberalism	0.12*	0.14**	0.11	0.05
Religiosity	—	−0.02	0.11	−0.08

Note: The preliminary SRS in Study 1a contained 18 items, the revised SRS in Study 1c contained 17 items, and the final SRS contains 11 items. We report uncorrected p-values in this table. We also applied a Bonferroni correction to maintain a family-wise error rate (within study) of 0.05. We observe that all correlations reported in this table with uncorrected p-values that are significant at the $p < 0.01$ level are also significant when we apply the Bonferroni correction, using a Bonferroni-adjusted alpha level of 0.007 per test in Study 1a (0.05/7), 0.006 in Studies 1c and 2 (0.05/8), and 0.01 in Study 3 (0.05/5).
SRS, Scientific Reasoning Scale; CRT, Cognitive Reflection Test; AOT, Actively Open-minded Thinking; TFKSS, Trend Factual Knowledge of Science Scale; USIS, Understanding of Scientific Inquiry Scale.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 4. Partial correlations of the Scientific Reasoning Scale (SRS) with scientific consensus beliefs

Belief	Preliminary scale Study 1a ($n = 397$)	Revised scale Study 1c ($n = 395$)	Final scale Study 2 ($n = 270$)	Final scale Study 3 ($n = 294$)
Composite measure	0.17***	0.22***	0.32***	0.29***
Global warming	0.06	0.06	0.01	0.08
Genetically modified foods	0.08	0.21***	0.32***	0.19**
Vaccines	0.22***	0.15**	0.29***	0.19***
Human evolution	0.10[a]	0.23***	0.21***	0.26***
The Big Bang	0.10*[a]	0.03	0.06	0.16**

Note: The preliminary SRS in Study 1a contained 18 items, the revised SRS in Study 1c contained 17 items, and the final SRS contains 11 items. We report partial correlations controlling for political conservatism and self-reported education (Study 1a) as well as religiosity (Studies 1c, 2, and 3). We report uncorrected p-values in this table. We also applied a Bonferroni correction to maintain a family-wise error rate (within study) of 0.05. We observe that all correlations reported in this table with uncorrected p-values that are significant at the $p < 0.01$ level are also significant when we apply the Bonferroni correction, using a Bonferroni-adjusted alpha level of 0.008 per test (0.05/6).
[a]Because belief in human evolution and the Big Bang were measured as binary variables in Study 1a, betas from logistic regressions are reported rather than partial correlations.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

scores on the composite belief measure, including self-reported education and political liberalism in the model as demographic controls. We then reran that regression, this time including SRS scores as a regressor, and found that including SRS scores improved model fit ($\Delta R^2 = 0.01$, $F(1, 391) = 4.55$, $p = 0.03$) (comparing the first two columns in Table 5). We report the results of including SRS scores on model fit when predicting belief in the safety of vaccines ($\Delta R^2 = 0.027$, $F(1, 391) = 11.8$, $p < 0.001$), the safety of genetically modified foods ($\Delta R^2 = 0.003$, $F(1, 391) = 1.26$, $p = 0.26$), global warming ($\Delta R^2 = 0.003$, $F(1, 391) = 1.27$, $p = 0.26$), human evolution (Δpseudo-$R^2 = 0.003$, $\chi^2(1) = 1.29$, $p = 0.26$), and the Big Bang (Δpseudo-$R^2 = 0.0001$, $\chi^2(1) = 0.04$, $p = 0.84$).

Discussion

Study 1a established the plausibility of the SRS. The preliminary SRS items possessed reasonable internal reliability, as measured by Cronbach's alpha. The preliminary SRS also possessed good construct validity, as indicated by predicted correlations with self-reported education, numeracy, cognitive reflection ability, and existing scientific literacy measures. Additionally, it predicted beliefs consistent with the scientific consensus above and beyond two widely used measures of scientific literacy.

As mentioned, two preliminary SRS items were removed from the analyses because of negative correlations with other items. As a result, the current measure omits two concepts identified as key to scientific reasoning (reliability and directness). Absent evidence on how participants interpreted these items, it is unclear whether they were simply worded poorly, or if those two concepts diverge from the other 18. Study 1b used qualitative methods to refine all the preliminary items.

STUDY 1B

We refined the preliminary SRS test items by eliciting qualitative feedback, in order to improve their clarity and ensure

Table 5. Predictors of scientific consensus beliefs

	Scientific consensus beliefs composite measure					
Variable	Preliminary scale Study 1a		Revised scale Study 1c		Final scale Study 2	
Constant	−7.6***	−8.2***	−7.6***	−7.4***	−9.6***	−9.3***
TFKSS	0.42**	0.34*	0.55***	0.44**	0.43**	0.31*
USIS	0.21	0.12	0.33*	0.20	0.59**	0.42*
Education	−0.06	−0.11	0.12	0.04	0.39**	0.29*
Political liberalism	1.3***	1.3***	0.99***	0.96***	1.1***	1.1***
Religiosity	—	—	−0.31**	−0.34**	−0.14	−0.12
SRS		0.13*		0.21**		0.22**
R^2	0.25	0.26	0.24	0.25	0.36	0.38
F	32.8***	27.4***	24.3***	22.0***	29.1***	27.0***
ΔR^2		0.01*		0.02**		0.03**

Note: $N = 397$ in Study 1a, 395 in Study 1c, and 270 in Study 2. The preliminary SRS in Study 1a contained 18 items, the revised SRS in Study 1c contained 17 items, and the final SRS contains 11 items.
SRS, Scientific Reasoning Scale; TFKSS, Trend Factual Knowledge of Science Scale; USIS, Understanding of Scientific Inquiry Scale.
$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

that they did not lead participants to make unintended assumptions or conclusions. To that end, we conducted 10 rounds of qualitative data collection, starting with the 18 items reported in the Study 1a analyses and revisions of the two problematic items. On the first round of data collection, we recruited 100 participants from MTurk. Participants were paid $0.50. Demographic information was not collected. Each participant answered five randomly selected SRS questions (out of the 20). They were asked to think aloud (in writing) as they interpreted the scenario and reasoned about their answer to the true/false statement, with the instruction, “Why did you select that answer? Please try to describe your thoughts as clearly as you can, using full sentences if possible.”

Responses were pooled by item and then coded by the first author in terms of whether they contained a misinterpretation of the question text.[3] Table 6 has examples of responses with and without such misinterpretation. The question was accepted as worded if an item met the following criteria: (i) most participants who answered correctly provided a reason supported by the question text and (ii) most participants who answered incorrectly provided a reason for their answer that did not reflect misinterpretation of the question. If, however, an item did not meet these criteria, it was revised based on the qualitative data and underwent another round of qualitative testing. Each round of testing involved new participants, with at least 20 responses per question. In each round, participants were paid $0.50. Study 1c evaluated the 20 revised questions.

STUDY 1C

Study 1c repeated Study 1a with the revised test items.

Method

Participants

We recruited from MTurk a sample of 400 American adults who had not participated in Study 1a. Participants were paid $2.50. Out of this initial sample, five were excluded based on their answers to the two data quality assurance questions drawn from Meade and Craig (2012). According to self-reports, in the final sample of 395 adults, 52% were male; 12% were between the ages of 18 and 24, 48% were between 25 and 34, 24% were between 35 and 44, 8% were between 45 and 54, and 8% were 55 or older; 78% were Caucasian; and 49% had received at least a Bachelor’s degree.

Materials and procedure

As in Study 1a, all participants first answered the 20 revised SRS items, with the order determined randomly for each participant. Next, all answered, in random order, the Cognitive Reflection Test, a six-item measure of numeracy (Peters et al., 2007), the TFKSS (NSB, 2014), and the USIS (NSB, 2014).

The questions on political orientation and beliefs in the scientific consensus were the same as in Study 1a, except that beliefs in human evolution and the Big Bang were elicited on a 5-point scale from *very likely* to *very unlikely* (to have occurred), in order to capture greater variation in participant beliefs. In addition, because beliefs on scientific controversies are often related to religious beliefs (Pew Research Center, 2013; Scheufele, Corley, Shih, Dalrymple, & Ho, 2009), we added a question asking participants to what extent they agreed or disagreed with “I consider myself to be a religious person” on a 5-point scale from *strongly agree* to *strongly disagree*. Table 2 presents descriptive statistics on these measures.

Results

Scientific reasoning scale preliminary item analysis

On average, participants answered 13.1 of the 20 revised SRS items correctly, $SD = 3.01$. Tetrachoric correlations revealed that three items, random assignment to condition, natural variation, and methodological basis, were negatively or weakly

[3] We used a single coder because: (1) our qualitative codes are not used as outcome measures, but as guides for the development of test items, and (2) the developed test items are independently tested in a new sample in Study 1c.

DOI: 10.1002/bdm

Table 6. Sample item responses from Study 1b

	Sample responses	
Preliminary SRS item	No misunderstanding	Misunderstanding
A marketing researcher has subjects pick one of two familiar snacks, A and B, then report how happy they are. On average, they report feeling happier after eating Snack A. True or False? This study shows that eating Snack A makes people happier than eating Snack B.	The snacks are a matter of personal preference. Someone who likes Snack B would feel happier eating that snack than the other. Also, if the subjects get to choose, what is the point of the study anyway? They would need to eat both snacks, then report their mood.	People who tried both snacks said they felt happier after eating Snack A.
A researcher develops a new technique for measuring a certain property of liquids. The technique generates different answers for different liquids. Also, when the researcher tests the new technique repeatedly with the same liquid, the technique generates very similar answers each time. True or False? The new technique is an accurate way to measure this property.	The new technique works. It generates similar answers when it is tested on the same liquid.	You should test using different liquids, not the same one over and over again.

correlated with most of the other 17 questions. These items were omitted from further analyses. On the resulting 17-item revised SRS, participants answered an average of 10.9 questions correctly, $SD = 3.0$.

Inter-item reliability
Cronbach's alpha for the revised 17-item scale was 0.66.

Construct validity
Correlations of the revised SRS with convergent measures again showed the predicted patterns: individuals who scored higher on the SRS had higher Cognitive Reflection Test scores, greater numeracy, higher self-reported levels of education, and higher scores on the TFKSS and USIS (Table 3). Among participants who indicated completing at least some college ($N = 338$), those with higher revised SRS scores also reported having taken more science classes ($r = 0.25$, $p < 0.001$).

Demographic correlates
Scores on the revised SRS were positively correlated with age and self-reported liberalism and were unrelated to gender ($t(392) = 0.83$, $p = 0.41$) or self-reported religiosity ($r = -0.02$, $p = 0.7$) (Table 4).

Predictive validity
Partial correlations (Table 4) controlling for self-reported religiosity, education, and political liberalism revealed that individuals with higher scores on the revised SRS again had higher scores on the composite belief measure. They were more likely to believe that genetically modified foods and vaccines were safe, consistent with the scientific consensus, but were no more likely to believe in global warming. Unlike Study 1a, individuals with higher SRS scores were significantly more likely to believe in human evolution and no more likely to believe in the Big Bang.

Incremental predictive validity
Using the same regression procedure as in Study 1a, with religiosity as an additional predictor variable, we found that including SRS scores significantly improved model fit when predicting the composite belief measure ($\Delta R^2 = 0.01$, $F(1, 388) = 7.4$, $p < 0.01$) (Columns 3 and 4 in Table 5). We report the results of including SRS scores on model fit when predicting belief in the safety of vaccines ($\Delta R^2 = 0.01$, $F(1, 388) = 4.41$, $p = 0.04$), the safety of genetically modified foods ($\Delta R^2 = 0.03$, $F(1, 388) = 13.0$, $p < 0.001$), global warming ($\Delta R^2 = 0.0003$, $F(1, 388) = 0.2$, $p = 0.7$), human evolution ($\Delta R^2 = 0.01$, $F(1, 388) = 6.38$, $p = 0.01$), and the Big Bang ($\Delta R^2 = 0.0001$, $F(1, 388) = 0.04$, $p = 0.8$).

Discussion

The general pattern of results paralleled that in Study 1a: scores on the revised version of the SRS had acceptable reliability and good construct validity, predicted scientific consensus beliefs, and added predictive validity to the existing measures of scientific literacy. Overall, these relationships were stronger than in Study 1a, indicating that the qualitative research of Study 1b improved the items. However, three items possessing poor psychometric properties were dropped from our analyses, indicating that these items require additional refinement.

STUDY 1D

Despite the improved overall performance of the SRS test items in Study 1c, three items were removed from the analyses because of negative correlations with the other items. These three items had been completely rewritten as a result of participant responses in Study 1b and tested different concepts than those tested by the two omitted items from Study 1a (the revisions of which were positively correlated with the remaining items in Study 1c). The three omitted items addressed the concepts of random assignment to condition, natural variation, and methodological basis. Given the iterative revisions in Study 1b, it seemed unlikely that these three

DOI: 10.1002/bdm

omitted items were poorly worded. Instead, we hypothesized that the negative correlations reflected some aspect of how these concepts were being tested. In Study 1d, we first rewrote these three items to test a different instantiation of the concepts and then conducted qualitative data collection following the methodology of Study 1b. As in Study 1b, participants were recruited from MTurk and paid $0.50; demographic information was not collected. Each rewritten item went through between one and three rounds of qualitative data collection before it met the criteria of Study 1b.

STUDY 2

Study 2 developed the final version of the SRS, using the items developed in Study 1 and the psychometric tests used in Study 1.

Method

Participants

We recruited from MTurk a new sample of 274 American adults who had not taken part in any of the previous studies. They were paid $2.50. Four were excluded based on their answers to the two data quality assurance questions drawn from Meade and Craig (2012). According to self-reports, in the final sample of 270, 52% were male; 19% were between 18 and 24, 39% were between 25 and 34, 25% were between 35 and 44, 9% were between 45 and 54, and 8% were 55 or older; 77% were Caucasian; and 44% had at least a Bachelor's degree.

Materials and procedure

The procedure was identical to that of Study 1c. Participants answered the 20 SRS test items, including the three items revised in Study 1d. The order of the SRS items was again determined randomly for each participant. Table 2 reports descriptive statistics.

Results

Development of the final Scientific Reasoning Scale

Before conducting an exploratory factor analysis, we computed tetrachoric inter-item correlations to check whether the revisions of Study 1d improved the three test items. We found that all 20 items were positively correlated. We then considered item difficulty, item–total correlations, and internal reliability coefficients to eliminate eight items. These analyses led us to eliminate items corresponding to four concepts drawn from Funtowicz and Ravetz (1990; as interpreted by Fischhoff and Davis, 2014) and designed to test the ability to assess the strength of a body of scientific evidence. These items appeared to be distinct from the other 16 items, displaying low item–total correlations, and low inter-item correlations with the other 16 items. We eliminated two additional items displaying low item–total correlations (r's $< .3$) and two items that were too easy to have variance in responses (M's > 0.90).

We hypothesized that a single latent construct, scientific reasoning ability, drove responses on the remaining 12 items and sought a factor solution that did the following: (i) satisfied Cattell's scree test; (ii) retained items with loadings of at least 0.4; and (iii) made psychological sense given our hypothesized measurement model. In order to determine the number of factors to extract, we conducted Horn's (1965) parallel analysis on the remaining 12 items using the "psych" package in R (Revelle, 2014). Parallel analysis, using the minimum residual factor method and 500 iterations, suggested that we retain one factor, supporting our hypothesis of unidimensionality (Table 7). We conducted an exploratory factor analysis extracting one factor, using the iterated principal factor method on the matrix of tetrachoric inter-item correlations. Exploratory factor analysis confirmed a unidimensional factor structure: after dropping one item with a low factor loading (<0.4), all 11 retained items loaded on one factor. See Table 1 for the factor loadings, and item–total correlations for the retained items.

We conducted a confirmatory factor analysis on a separate sample of MTurk workers ($N = 345$; 52% male, mean age 34 ($SD = 10.6$), 80% Caucasian, and 43% with at least a Bachelor's degree). We report the following goodness-of-fit indices for the one-factor model: $\chi^2(44) = 136$, $p < 0.001$; root mean squared error of approximation = 0.078; standardized root mean square residual = 0.046; comparative fit index = 0.90.

On average, participants in Study 2 answered 6.6 of the 11 SRS questions correctly ($SD = 2.6$). Retained items and their difficulties are provided in the Appendix.

Reliability

The final, 11-item version of the SRS had a Cronbach's α of 0.70.

Construct validity

Correlations of the final SRS with convergent measures showed patterns similar to those in Studies 1a and 1c and are reported in Table 3. Individuals who scored higher on the SRS had higher Cognitive Reflection Test scores, greater numeracy, higher self-reported levels of education, and higher scores on the TFKSS and USIS. Among participants who indicated completing at least some college ($N = 232$), those with higher SRS scores reported having taken more science classes ($r = 0.16$, $p = 0.02$).

Demographic correlates

Scientific Reasoning Scale scores were positively correlated with age and unrelated to gender ($t(268) = -1.46$, $p = 0.15$), political liberalism, and religiosity (Table 3).

Table 7. Comparison of factor analysis and parallel analysis eigenvalues

Factor	Factor analysis eigenvalue	Parallel analysis eigenvalue
1	3.24	0.89
2	0.37	0.44
3	0.26	0.32

Note: Parallel analysis conducted using $N = 500$ iterations.

DOI: 10.1002/bdm

Predictive validity

As reported in Table 4, after controlling for education, religiosity, and political liberalism, individuals with higher scores on the final, 11-item SRS had higher scores on the composite belief measure. Individuals with higher SRS scores were significantly more likely to believe in the safety of genetically modified foods and vaccines, as well as in human evolution. However, they were no more likely to believe in the Big Bang or global warming.

Incremental predictive validity

Using the same regression procedure as in Study 1c, we found that SRS scores displayed incremental predictive validity over the two existing measures of scientific literacy (again controlling for self-reported education, political liberalism, and religiosity), when predicting the composite belief measure ($\Delta R^2 = 0.03$, $F(1, 263) = 11.0$, $p = 0.001$) (Columns 5 and 6 in Table 5). We report the results of including SRS scores on model fit when predicting belief in the safety of genetically modified foods ($\Delta R^2 = 0.04$, $F(1, 263) = 10.9$, $p = 0.001$), the safety of vaccines ($\Delta R^2 = 0.03$, $F(1, 263) = 9.61$, $p = 0.002$), global warming ($\Delta R^2 = 0.0002$, $F(1, 263) = 0.08$, $p = 0.77$), human evolution ($\Delta R^2 = 0.009$, $F(1, 263) = 3.33$, $p = 0.07$), and the Big Bang ($\Delta R^2 = 0.002$, $F(1, 263) = 0.65$, $p = 0.42$).

Discussion

The final version of the SRS has 11 items that load on a single latent factor, meant to represent scientific reasoning ability. Because of the heterogeneous domain of scientific reasoning ability, we were not surprised to find a relatively weak one-factor solution. The final SRS had a Cronbach's α of 0.70, which we consider to be an acceptable level of reliability given the heterogeneity of the domain of scientific reasoning and the scale's reliance on scenarios. Our results indicate that the final SRS possesses good construct validity and incremental predictive validity when predicting belief in the scientific consensus, for the composite measure and three of the five individual topics (vaccine safety, safety of genetically modified foods, and human evolution, but not global warming or the Big Bang).

STUDY 3

One limitation of existing scientific literacy measures is the lack of evidence linking scores on them to performance on tasks requiring the use of scientific information. Study 3 tests whether individuals with higher SRS scores are better able to use scientific information in one concrete context: a drug facts box. We use a validated drug facts box task (Schwartz & Woloshin, 2013; Woloshin & Schwartz, 2011) that requires subjects to interpret numerical information regarding the effectiveness and side effects of certain drugs. We also further assess the construct validity of the SRS by examining how scores relate to a theoretically related cognitive style, the tendency to engage in actively open-minded thinking.

Method

Participants

We recruited 295 American adults from MTurk who had not taken any of the previous studies. Participants were paid \$2.20. One participant was excluded as a result of failing a data quality assurance question from Meade and Craig (2012). In the final sample of 294 adults, the mean reported age was 36 ($SD = 11.4$), 44% were male, 81% were Caucasian, and 50% had at least a Bachelor's degree.

Materials and procedure

All participants first completed the SRS. The order of the items was determined randomly for each participant. Next, they completed the drug facts box test (Woloshin & Schwartz, 2011). In this task, participants received short descriptions of two fake drugs, one for heartburn (PAXCID) and one to prevent heart attacks (QUESTOR). They were also provided with numerical information about patient outcomes (e.g., heart attacks and side effects such as muscle aches) from drug trials comparing each drug with a placebo. Participants then answered questions about these numerical outcomes, such as "True or False? People given QUESTOR were twice as likely to have bothersome muscle aches as people given placebo." Next, participants completed the Actively Open-minded Thinking Scale (Haran, Ritov, & Mellers, 2013). Participants' political identity and beliefs on five controversial scientific issues were elicited as in Study 2. See Table 2 for descriptive statistics.

Results

On average, participants answered 7 of the 11 SRS questions correctly ($SD = 2.6$).

Reliability

The reliability of the final, 11-item SRS in this new sample was $\alpha = 0.71$.

Construct validity

Individuals who scored higher on the SRS also scored higher on the Actively Open-minded Thinking Scale, $r = 0.41$, $p < 0.001$. As before, individuals with higher SRS scores also reported higher levels of education, $r = 0.30$, $p < 0.001$.

Demographic correlates

As before, SRS scores were positively correlated with age and unrelated to gender ($t(292) = -0.81$, $p = 0.42$), religiosity, and political liberalism (Table 3).

DOI: 10.1002/bdm

Predictive validity

Partial correlations controlling for education, religiosity, and political liberalism, displayed in Table 4, revealed that people with higher SRS scores had higher scores on the composite belief measure. As before, they were more likely to believe that vaccines and genetically modified foods are safe and to believe in human evolution. Here, they were more likely to believe in the Big Bang but again were no more likely to believe in global warming.

Drug facts box

Participants answered an average of 17 out of 20 questions correctly on the drug facts box test ($SD = 2.3$). As predicted, individuals who scored higher on the SRS had higher scores on the drug facts box comprehension test, $r = 0.44$, $p < 0.001$.

Discussion

The final, 11-item SRS demonstrated good reliability and ability to predict whether individuals held beliefs consistent with the scientific consensus. SRS scores were related to cognitive styles, as captured by a measure of actively open-minded thinking. Moreover, SRS scores predicted the ability to use the scientific information in a drug facts box.

GENERAL DISCUSSION

The ability to evaluate scientific evidence is important to many decisions in individuals' lives as consumers, patients, and citizens. Here, we develop and validate the SRS, designed to measure individuals' ability to evaluate scientific findings. Using an interdisciplinary approach building on research in behavioral decision research, cognitive developmental psychology, and public understanding of science, we define *scientific reasoning skills* and measure them with an 11-item test that requires participants to apply their reasoning skills to brief scientific scenarios. We find that the resultant scale (the SRS) has good internal consistency and construct validity, as indicated by positive correlations with measures of numeracy, cognitive reflection, education, and actively open-minded thinking. Individuals with higher SRS scores are more likely to have beliefs consistent with the scientific consensus above and beyond two widely used scientific literacy measures, and after controlling for demographic measures. They also perform better on a task requiring analysis of scientific information (Study 3).

One limitation to our research is its use of online samples recruited from Amazon.com's MTurk. Although MTurk participants have been found to be more demographically diverse than typical Internet and American college student samples and to produce good quality data (Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014), caution is warranted when generalizing to the general population. As safeguards to ensure data quality, we excluded participants who failed attention check questions and used filters to ensure that each study had new participants. Nonetheless, relative to the general US population, MTurk samples are biased toward individuals who are young, Caucasian, and have a college education. Although we controlled for key demographics in our analyses, it remains to be seen whether the SRS will display the same properties when administered to a nationally representative sample. More diverse respondents might reveal stronger relationships, or some limit to our measure.

One puzzle in our results is the variation in correlations between SRS scores and the specific scientific consensus beliefs. Across the four samples (and variants of the SRS), respondents with higher SRS scores consistently had higher scores on the composite measure summarizing beliefs on the five issues: global warming, genetically modified foods, vaccination, human evolution, and the Big Bang. However, higher SRS scores consistently predicted beliefs only on three issues (genetically modified foods, vaccines, and human evolution), while being unrelated to beliefs on the other two (global warming and the Big Bang).

One possible explanation lies with how those beliefs were measured. Several strands of recent research reveal sensitivity to how questions on controversial topics are posed. For example, people are far more likely to endorse human evolution and the Big Bang theory when these topics were prefaced with "according to the theory of evolution" and "according to astronomers" (National Science Board, 2014, pp. 7–4). Similarly, using *global warming* rather than *climate change* on a survey generates significantly different responses, with the former evoking stronger endorsements, especially from men and liberals (Leiserowitz, Feinberg, Rosenthal, Smith, Anderson, Roser-Renouf, & Maibach, 2014). Wong-Parodi and Fischhoff (2015) found that differences between believers and nonbelievers in climate change disappeared when participants expressed their beliefs before beginning a task related to sea-level rise. Thus, understanding the role of scientific reasoning skills in beliefs regarding controversial scientific issues depends on how both the skills and the beliefs are measured.

A more substantive explanation lies with differences in the five beliefs. Toplak and Stanovich (2003) asked participants to generate arguments for and against their current position on three controversial (non-scientific) issues. Although participants generated more arguments for rather than against their current positions, the degree of that *belief bias* varied by domain, leading the researchers to conclude that beliefs "differ in how strongly they are structured to repel contradictory ideas" (Toplak & Stanovich, 2003, p. 859). We hope that our measure helps to disentangle the multiple possible sources of disbelief in scientific results, some of which may be unrelated to scientific reasoning ability (Fiske & Dupree, 2014; Medin & Bang, 2014; Taber & Lodge, 2006).

The SRS appears to provide a theoretically informed contribution to decoding lay responses to scientific results and controversies. Future research will seek to identify the conditions under which people reject rather than accept scientific evidence and test strategies to reduce the likelihood of such rejection. It is our hope that our research program will lead to a rich, respectful understanding of lay reasoning, useful to those who comment on the public and those who seek to help it.

DOI: 10.1002/bdm

APPENDIX A: SCIENTIFIC REASONING SCALE ITEMS AND ITEM DIFFICULTIES

			Study 2	Study 3
Concept	Test item	Correct answer	% correct	% correct
Blind/double blind	In a taste test, a researcher puts Brand A coffee in a cup with white tape on it and Brand B coffee in an identical cup with black tape on it. A lab assistant gives tasters one of the cups, while the researcher watches their facial expressions. True or False? The lab assistant should not watch the cups being filled.	True	53	61
Causality	A researcher finds that American states with larger parks have fewer endangered species. True or False? These data show that increasing the size of American state parks will reduce the number of endangered species.	False	54	58
Confounding variables	A researcher has subjects put together a jigsaw puzzle either in a cold room with a loud radio or in a warm room with no radio. Subjects solve the puzzle more quickly in the warm room with no radio. True or False? The scientist cannot tell if the radio caused subjects to solve the puzzle more slowly.	True	76	79
Construct validity	An education researcher wants to measure the general math ability of a sample of high-performing math students. All the students have taken classes in geometry and pre-calculus. True or False? The education researcher can measure general math ability by giving the students a geometry test.	False	56	61
Control group	Two scientists test an anti-acne cream on teenagers with acne. Scientist A wants to give the cream to all the teenagers in the study. Scientist B wants to give the cream to half the teenagers and give a cream without anti-acne ingredients to the other half. True or False? Both ways of testing the cream are equally good.	False	76	77
Ecological validity	A researcher has a group of subjects play a competitive game. Each subject's goal is to make money by buying and selling tokens. Subjects are paid a flat fee for participating in the experiment. True or False? The researcher can confidently state that the behavior in the experiment reflects real-life buying and selling behavior.	False	68	67
History	A randomly selected sample of Americans is surveyed about disease A before and after a 6-month media campaign about the disease. Mid-way through the media campaign, a famous celebrity dies of Disease A. The survey data indicate that knowledge of Disease A is higher after the campaign. True or False? The media campaign may not have increased knowledge of Disease A.	True	69	72
Maturation	Subjects in an experiment must press a button whenever a blue dot flashes on their computer screen. At first, the task is easy for subjects. But as they continue to perform the task, they make more and more errors. True or False? The blue dot must flash more quickly as the task progresses.	False	66	68
Random assignment to condition	Researchers want to see whether a health intervention helps school children to lose weight. School children are sorted into either an intervention or control group. True or False? The researchers should assign the overweight children to the intervention group.	False	64	65
Reliability	A researcher develops a new method for measuring the surface tension of liquids. This method is more consistent than the old method. True or False? The new method must also be more accurate than the old method.	False	49	51
Response bias	Two researchers are developing a survey to measure consumers' feelings about customer service. Researcher A wants customers to rate their agreement with the statement "I am satisfied with customer service" on a 5-point scale, where 1 = *strongly agree* and 5 = *strongly disagree*. Researcher B wants customers to rate customer service on a 5-point scale, where 1 = *not dissatisfied at all* and 5 = *highly dissatisfied*. True or False? These questions are equally good for measuring how consumers feel about customer service.	False	35	45

ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1252522 and NSF grant SES-0949710. This research was supported in part by a small grant from the Center for Behavioral and Decision Research at Carnegie Mellon to the first author. We thank Stephen Broomell for his comments on an early draft.

REFERENCES

Allum, N., Sturgis, P., Tabourazi, D., & Brunton-Smith, I. (2008). Science knowledge and attitudes across cultures: A meta-analysis. *Public Understanding of Science*, *17*(1), 35–54. doi: 10.1177/0963662506070159.

Barkhordarian, A., Pellionisz, P., Dousti, M., Lam, V., Gleason, L., et al. (2013). Assessment of risk of bias in translational science. *Journal of Translational Medicine*, *11*(1), 184. doi: 10.1186/1479-5876-11-184.

DOI: 10.1002/bdm

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. doi: 10.1037/0021-9010.78.1.98.

Downs, J. S., Bruine de Bruin, W., & Fischhoff, B. (2008). Parents' vaccination comprehension and decisions. *Vaccine*, *26*(12), 1595–1607. doi: 10.1016/j.vaccine.2008.01.011.

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701–718). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199734689.013.0035

Fischhoff, B. (2013). The sciences of science communication. *Proceedings of the National Academy of Sciences*, *110*, 14033–14039. doi: 10.1073/pnas.1213273110.

Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences*, *111* (Supplement 4), 13664–13671. doi: 10.1073/pnas.1317504111.

Fiske, S. T., & Dupree, C. (2014). Gaining trust as well as respect in communicating to motivated audiences about science topics. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13593–13597. doi: 10.1073/pnas.1317505111.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi: 10.1257/089533005775196732.

Funk, C., & Rainie, L. (2015). *Public and scientists' views on science and society*. Washington, D.C.: Pew Research Center Retrieved from http://www.pewinternet.org/2015/01/29/public-and-scientists-views-on-science-and-society/, accessed on March 10, 2015.

Funtowicz, S. O., & Ravetz, J. (1990). *Uncertainty and quality in science for policy*. London: Kluwer.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188–201.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hoyle, R. H., Harris, M. J., & Judd, C. M. (2002). *Research methods in social relations*. NY: Wadsworth.

John, O., & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation and scale construction. In Reis, H. T., & Judd, C. M. (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). Cambridge: Cambridge University Press.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., et al. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*, 732–735.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48. doi: 10.1207/s15516709cog1201_1.

Leiserowitz, A., Feinberg, G., Rosenthal, S., Smith, N., Anderson, A., et al. (2014). *What's in a name? Global warming vs. climate change. Yale project on climate change communication*. New Haven, CT: Yale University and George Mason University.

Leiserowitz, A., Maibach, E., Roser-Renouf, C., & Smith, N. (2011). *Global warming's six Americas, May 2011. Yale project on climate change communication*. New Haven, CT: Yale University and George Mason University.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–55. doi: 10.1037/a0028085.

Medin, D. L., & Bang, M. (2014). The cultural side of science communication. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13621–13626. 10.1073/pnas.1317510111.

Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus*, *112*(2), 29–48.

Miller, J. D. (1998). The measurement of civic scientific literacy. *Public Understanding of Science*, *7*, 203–223.

Miller, J. D. (2002). Civil scientific literacy: A necessity in the 21st century. *FAS Public Interest Report: Journal of the Federation of American Scientists*, *55*(1), 3–6.

Miller, J. D. (2004). Public understanding of, and attitudes toward, scientific research: What we know and what we need to know. *Public Understanding of Science*, *13*(3), 273–294. doi: 10.1177/0963662504044908.

National Science Board (2014). *Science and engineering indicators 2014*. Arlington VA: National Science Foundation (NSB 14-01).

Organisation for Economic Co-operation and Development (OECD) (2003). *The PISA 2003 assessment framework—Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi: 10.1177/0963721414531598.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review: MCRR*, *64*(2), 169–90. doi: 10.1177/10775587070640020301.

Pew Research Center (2013). *Public's views on human evolution*. Washington, D.C.: Pew Research Center. Retrieved from http://www.pewforum.org/2013/12/30/publics-views-on-human-evolution/ accessed on March 10, 2015.

Reis, H. T., & Judd, C. M. (Eds) (2000). *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Revelle, W. (2014). Psych: Procedures for personality and psychological research. R package version 1.4.12. http://personality-project.org/r, accessed on March 10, 2015.

Scheufele, D. A., Corley, E. A., Shih, T., Dalrymple, K. E., & Ho, S. S. (2009). Religious beliefs and public attitudes toward nanotechnology in Europe and the United States. *Nature Nanotechnology*, *4*(December), 91–94. doi: 10.1038/nnano.2008.361.

Schwartz, L. M., & Woloshin, S. (2013). The drug facts box: Improving the communication of prescription drug information. *Proceedings of the National Academy of Sciences*, *110*, 14069–14074. doi: 10.1073/pnas.1214646110.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. doi: 10.1037/0022-0663.89.2.342.

Sturgis, P., & Allum, N. (2004). Science in society: Re-evaluating the deficit model of public attitudes. *Public Understanding of Science*, *13*(1), 55–74. doi: 10.1177/0963662504042690.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tangney, J. P. (1996). Conceptual and methodological issues in the assessment of shame and guilt. *Behavioral Research and Therapy*, *34*(9), 741–754.

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, *17*, 851–860.

Trochim, W. M. K., & Donnelly, J. P. (2007). *The research methods knowledge base* (3rd ed.). NY: Atomic Dog Publishing.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: B. T. Batsford.

Woloshin, S., & Schwartz, L. (2011). Communicating data about the benefits and harms of treatment. *Annals of Internal Medicine*, *155*(2), 87–96.

Wong-Parodi, G., & Fischhoff, B. (2015). The impacts of political cues and practical information on climate change decisions. *Environmental Research Letters*, *10*034004. doi: 10.1088/1748-9326/10/3/034004.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149. doi: 10.1006/drev.1999.0497.

Authors' biographies:

Caitlin Drummond. is a fourth year graduate student at Carnegie Mellon University. Her research focuses how nonscientists understand and interpret scientific evidence, with particular emphasis on scientific controversies and the factors that predict acceptance and rejection of science.

Baruch Fischhoff is Howard Heinz University Professor at Carnegie Mellon University. He is co-author (with John Kadvany) of Risk: A Very Short Introduction) and co-editor (with Dietram Scheufele) of two special issues of PNAS on the science of science communications (vol. 110[supplement 3], vol. 111[supplement 4]).

Authors' addresses:

Caitlin Drummond, Departmental of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA.

Baruch Fischhoff, Department of Social and Decision Sciences, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.

DOI: 10.1002/bdm