CHAPTER SEVEN

# What Makes Everyday Scientific Reasoning So Challenging?

**Priti Shah[*,1], Audrey Michal[§], Amira Ibrahim[*], Rebecca Rhodes[¶] and Fernando Rodriguez[||]**

[*]University of Michigan, Ann Arbor, MI, United States
[§]Northwestern University, Evanston, IL, United States
[¶]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States
[||]University of California, Irvine, Irvine, CA, United States
[1]Corresponding author: E-mail: priti@umich.edu

## Contents

*Psychology of Learning and Motivation,* Volume 66
ISSN 0079-7421
http://dx.doi.org/10.1016/bs.plm.2016.11.006

## Abstract

Informed citizens are expected to use science-based evidence to make decisions about health, behavior and public policy. To do so, they must judge whether the evidence is consistent with the claims presented (theory-evidence coordination). Unfortunately, most individuals make numerous errors in theory-evidence coordination. In this chapter, we provide an overview of research on science evidence evaluation, drawing from research in cognitive and developmental psychology, science and statistics education, decision sciences, political science and science communication. Given the breadth of this research area, we highlight some influential studies and reviews across these different topics. This body of research provides several clues about: (1) why science evidence evaluation is challenging, (2) the influence of the content and context of the evidence and (3) how the characteristics of the individual examining the evidence impact the quality of the evaluations. Finally, we suggest some possible directions for empirical research on improving evidence evaluation and point to the responsibility of scientists, especially social and behavioral scientists, in communicating their findings to the public. Overall, our goal is to give readers an interdisciplinary view of science evidence evaluation research and to integrate research from different scientific communities that address similar questions.

## 1. INTRODUCTION

People are regular consumers of science claims presented in newspapers, advertisements, scientific articles and word of mouth (Baram-Tsabari & Osborne, 2015; Bromme & Goldman, 2014). Consider the following headlines:

*Lifting Lighter Weights Can Be Just as Effective as Heavy Ones.*
***NY Times (July 20, 2016)***

*Dose of nature is just what the doctor ordered.*
***Sciencedaily.com (June 23, 2016)***

*Why scientists think your social media posts can help prevent suicide.*

**Mashable (June 26, 2016)** [1]

Informed citizens are expected to use this information to make decisions about health, behavior, and public policy (Kolstø et al., 2006; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). However, media reports of research often overstate the implications of scientific evidence, overlook methodological or statistical flaws or even present pseudoscience (Bromme & Goldman, 2014). People need to learn to distinguish between high quality science, low quality science and pseudoscience (Anelli, 2011; Miller, 1996, pp. 185–204; Sagan, 1996a, 1996b; Trefil, 2008). Unfortunately, students and sometimes even trained scientists make errors when reasoning about evidence (Halpern, 1998). In fact, more than 70% of American adults report to believe in paranormal phenomena (Halpern, 1998), and a similar percent of Americans acknowledge holding at least one pseudoscientific belief (Moore, 2005). One-third of Americans think evolution is "absolutely false" and another 21% is not sure (Miller, Scott, & Okamoto, 2006). Only 43% understand that the earth is billions of years old (Bishop, Thomas, Wood, & Gwon, 2010), and only about 30% of American adults can read and understand the science section of the New York Times (Anelli, 2011).

Poor evidence evaluation skills are especially problematic for issues related to public policy and personal choices. About one-third of the population in the United States does not believe that climate change is caused by human activity, and of six different countries, US individuals are least concerned (Gallup, 2016; Shi, Visschers, Siegrist, & Arvai, 2016). In addition, over half of Americans believe that genetically modified foods are unsafe despite fairly strong evidence that they are safe (Langer, 2015). The ability to understand scientific evidence, including evaluating the statistical properties of evidence, is associated with difficulty in risk perception and medical decision making (Galesic & Garcia-Retamero, 2011; E. Peters, 2012; J.D. Peters, 2012; Reyna, Han, Deickmann, 2009).

The current state of science, combined with the nature of scientific writing in the media, adds to the confusion. Consider a question of relevance to cognitive scientists, the extent to which "brain training" might have a positive—real life—impact on cognitive abilities in individuals. There

[1] Reynolds (2016), Sciencedaily.com (2016) and Ruiz (2016).

have been a recent barrage of media articles on this topic, and they all seem to provide contradictory recommendations and conclusions. A quick search on Google News on July 28, 2016 found the following headlines, all published within the last several days: ones that suggest "brain training" is not effective ("Brain Training Does Not Improve Academic Outcomes in Kids", ""Brain training" boost might just be a "placebo" effect, study finds"), ones that purport that brain training is promising ("Could "Brain Training" Games Actually Work? New Study Surprises Scientists", "Brain training game for troops tackles effects of combat", ""Brain training" cut dementia risk in healthy adults"), and ones that suggest a more complex picture ("All brain training protocols do not return equal benefits, study reveals".). How can even educated readers make sense out of these headlines? Is it so unfair for readers to simply assume that scientists do not know what they are talking about, especially when there are headline grabbing articles suggesting that much of psychological science cannot be replicated (Open Science Collaboration, 2015; Pashler & Harris, 2012)? Thousands of fMRI studies have errors in data analysis (Eklund, Nichols, & Knutsson, 2016), scientists are accused of "p-hacking" or fishing for publishable results (Ioannidis, 2008; Simmons, Nelson, & Simonsohn, 2011), and for many scientific concepts (like brain training) there is no scientific consensus (Katz & Shah, 2016a). Throughout this paper, we refer to the context of brain training to illustrate the challenges of evidence evaluation to the lay reader.

One might ask, is it not the goal of our educational system to teach students the inquiry skills necessary to critically evaluate scientific evidence and to assess whether or not evidence is consistent with claims or theories (Lehrer & Schauble, 2006; Next Generation Science Standards, 2013; Kolstø et al., 2006; Kuhn, 2001)? It is clear that teaching science content alone does not help students reason about science (Crowell & Schunn, 2016). As an example, though Chinese students learn a great deal more science content than US students, Chinese and US students perform equivalently on a standard measure of scientific reasoning (Bao et al., 2009; Lawson, 1978). At least one study found that people who had taken eight or more college science courses did not do much better on some scientific reasoning tasks than high school students (Norris & Phillips, 1994; Norris, Phillips, & Korpan, 2003), though many other studies find that college education and college training in scientific reasoning does at least correlate with better performance on science evidence evaluation tasks (Amsel et al., 2008; Burrage, 2008; Huber & Kuncel, 2015; Kosonen & Winne, 1995; Norcross, Gerrity, & Hogan, 1993). In another example, knowledge about the physical characteristics of climate change is actually associated with *less* concern about climate change (Kahan et al., 2012).

Several factors may contribute to the difficulty of applying scientific inquiry skills to everyday contexts, especially those that involve making personal or political decisions about health, behavior and social science data. Everyday contexts tend to evoke experiential thinking rather than analytic thinking (Kahneman, 2011). This tendency to think about one's own prior experiences is exacerbated by features of media writing, such as the inclusion of personal anecdotes, which increases tendencies towards experiential thinking and significantly decreases deeper analytic thinking (Rodriguez, Ng, & Shah, 2016; Rodriguez, Rhodes, Miller, & Shah, 2016). Science writers themselves may, perhaps inadvertently, take advantage of people's motivations and biases. In one analysis of headlines in the *New Scientist*, for example, one of the most frequent noncontent words used was "your" (Molek-Kozakowska, 2014), presumably to emphasize the relevance to the reader but also likely to activate personal beliefs. Although the reliance on heuristic thinking may explain many poor evidence evaluation outcomes, it is clear that relying on analytic thinking is not always sufficient for high-quality evidence evaluation, and sometimes individuals who are actually good at evidence evaluation in neutral contexts are even more polarized when evaluating evidence relevant to their own identities (Kahan, Peters, Dawson, & Slovic, 2013).

As the discussion above suggests, people are highly influenced by their own prior beliefs when evaluating evidence (see Evans & Curtis-Holmes, 2005; Klaczynski, 2000; Sá, West, & Stanovich, 1999, for some examples). Likewise, people are *motivated reasoners* and are influenced by their hopes and emotions in addition to their prior beliefs, especially when the context of the evidence is relevant to decisions in their own lives (Klaczynski, 2000; Kunda, 1990; Lord, Ross, & Lepper, 1979; Sinatra, Kienhues, & Hofer, 2014). When evaluating evidence that is congruent with prior beliefs, there is a tendency to rely on heuristic thinking ("that makes sense to me") and not engage in analytic thinking. Thus, beliefs and tendency towards heuristic thinking often go hand in hand. In contrast, belief-incongruent evidence often triggers analytic thinking.

Finally, studies of everyday health and behavior contexts, by practical necessity, often incorporate potential threats to scientific validity, and thus require significant attention to these issues. A large portion of work in epidemiology, economics and public health involves correlational data, and even when the data are "big" and many factors are statistically controlled, it is hard to evaluate whether or not the conclusions are sound; such evaluations are made by scientists in the context of knowledge about other studies and mechanistic models (e.g, Shiffrin, 2016; Vandenbroucke, Broadbent, & Pearce, 2016). Scientists themselves are in large part responsible for overstating the implications of their own findings. In one recent analysis of 462 health science press releases, Sumner et al. (2014) found that 40% contained advice that overstated the implications of the findings, 33% exaggerated the causal implications of correlational findings and 36% made inappropriate inferences regarding animal research. The authors of this study also analyzed the media articles that were written based on these press releases and found that the media articles are not adding to the exaggeration problem—rather, the exaggeration seems to arise from press releases themselves, which are presumably vetted by the scientists who conducted the research.

Whatever the reason, college students and laypersons rarely notice common inferential reasoning errors in everyday science contexts spontaneously, especially when they have beliefs or behaviors consistent with those claims (e.g., Rhodes, Rodriguez, & Shah, 2014; Rhodes & Shah, 2016a, 2016b; Rodriguez, Ng, et al., 2016; Rodriguez, Rhodes, et al., 2016). Specifically, people often accept correlational data as providing evidence of causality, fail to notice poorly controlled studies such as those that include obvious sampling bias, do not recognize low quality measurement or operationalization of variables, are insensitive to effect sizes, and often do not pay attention to important features of data such as sample size and variance (see for example Fong & Nisbett, 1991; Rhodes et al., 2014; Rodriguez, Ng, et al., 2016). They are also swayed by factors such as anecdotes (Rodriguez, Ng, et al., 2016), irrelevant neuroscience (Fernandez-Duque, Evans, Colton, & Hodges, 2015; Hopkins, Weisberg, & Taylor, 2016; McCabe & Castel, 2008; Rhodes et al., 2014; Rhodes & Shah, 2016a; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008), irrelevant mathematical equations (Eriksson, 2012), chemical formulas and graphs (Tal & Wansink, 2016).

Despite performing rather poorly across a range of everyday scientific evidence evaluation and statistical reasoning tasks, people often report learning about scientific reasoning flaws in science classrooms and indeed often apply their knowledge when motivated to do so. For example, most students in one of our studies report that they learned the methodological issue at hand (e.g., participants should be randomly assigned to conditions rather than selecting conditions); however, they rarely spontaneously refer to selection bias when evaluating evidence even when a study is clearly biased (see Rhodes et al., 2014). In other words, providing individuals with knowledge about how to critically evaluate scientific evidence is not enough to ensure that they will do so in everyday contexts. Furthermore, as we argue later in the paper, it is not clear whether or not the advice given to clinicians and public policy experts regarding the hierarchy of evidence [i.e, meta-analysis, systematic reviews, multiple randomized control trials (RCT), a single RCT trial and on down the line; see for example Concato, Shah, & Horwitz, 2000; D. Evans, 2003; J.S.B.T. Evans, 2003] or even statistical reasoning (i.e., large sample sizes are good) is appropriately interpreted.

The primary goal of this chapter is to review research on everyday science evidence evaluation. Evidence evaluation is only one component of scientific inquiry, but one of the more important and somewhat domain general skills (Schunn & Anderson, 1999). This review does not focus on other components of scientific inquiry such as hypothesis generation or experimentation skills (Klahr & Dunbar, 1988). We begin with providing a (purposely narrow) definition of quality evidence evaluation. We then provide a necessarily limited review of the long history of research on why people are not good at evidence evaluation, which documents the problem without proposing an adequate solution. We outline how the characteristics of information communication influence evidence evaluation quality, and how individual differences also impact evidence evaluation. Finally, we suggest some possible directions for empirical research on improving evidence evaluation and point to the responsibility of scientists, especially social and behavioral scientists, in communicating their findings to the public.

## 2. DEFINING QUALITY EVIDENCE EVALUATION AS THEORY-EVIDENCE COORDINATION

Theory-evidence coordination is a cornerstone of scientific reasoning and, in particular, evidence evaluation (Kuhn, 2001). Theory-evidence coordination involves judging whether or not evidence is consistent with a particular theory or interpretation, and whether or not the evidence provides adequate support for that theory. In everyday science, evidence evaluation involves judging whether a particular study finding is consistent with a claim or theories (usually a causal claim), is not consistent with any other claims or theories (alternative explanations), and often, whether it warrants making a recommended behavior or policy change.

In this section we consider two of the media articles about brain training cited above to illustrate theory-evidence coordination. The first, "Working Memory Training Shows No Benefit for Academics in Children" (http://www.neurologyadvisor.com) refers to a recent study "Academic outcomes 2 years after working memory training for children with low working memory: a randomized clinical trial" (Roberts et al., 2016). The second is Hurley (2016) "Can brain training prevent dementia?" based on an as yet unpublished study presented at the Alzheimer's Association annual meeting. These articles were not selected at random; the first author of this chapter wrote a brief letter in response to the Roberts article (Katz & Shah, 2016b), and her brother asked her, after reading a media report of the dementia study whether he should purchase the training program for their parents. Although our take on both of these articles is somewhat critical, they are probably just as good or better than many studies on both sides of the scientific debate on "brain training", and our own work in this area too has significant limitations. For either media report, the reader's goal might be to decide whether to purchase the software in question for his or her child or aging parent. To do so, the reader would have to judge first whether or not the research design and outcome supported the conclusions that working memory training has no benefit on academics in children, or in the aging parent scenario, that speed of processing training reduces the risks of dementia.

In research methods, scientific reasoning errors are characterized as "threats to validity" (Picardi & Masick, 2013; Reis & Judd, 2000). To what extent are nonscientists able to identify threats to validity in the context of theory-evidence coordination? The ability to recognize threats to scientific validity is only minimally taught in k-12 science classrooms; it is most often taught after students have taken courses in statistics or concurrently with statistics during postsecondary education.

To generate some baseline data on the ability of college undergraduates with differing experience and majors to recognize threats to validity, Burrage (2008) asked 268 University of Michigan undergraduates to discuss and to critically evaluate eight short vignettes describing scientific studies. Each vignette contained one of four threats to validity (further described below): causality bias, selection bias, poor construct validity and overgeneralization of small effect sizes. An example vignette from her study is:

> *A study of 77 children, aged 3 to 5, found that those with the most body fat had the most "controlling" mothers when it came to the amount of food eaten. "The*

> *more control the mother reported using over her child's eating, the less self-regulation the child displayed."*
>
> ***Dr. Johnson and her coauthor said***

The aforementioned vignette contains a threat to internal validity: inappropriately drawing a causal conclusion from correlational data. When asked to critically evaluate these vignettes, participants provided critique of the methodology or quality of the evidence less than 60% of the time and even fewer noticed target errors (e.g., interpreting correlational data as causal). College seniors were significantly more critical and were more likely to notice target flaws than freshmen. In addition, individual levels of actively open-minded thinking (AOT) (Stanovich & West, 1997) and need for cognition (NFC) (Cacioppo & Petty, 1982) predicted how frequently participants noticed flaws. We refer to this and other studies from our laboratory as we discuss individual threats to validity later.

## 2.1 Threats to Internal Validity

A *threat to internal validity* refers to a problem that makes it unclear whether or not a dependent variable is affected by a treatment or independent variable (e.g., an experiment where there is a confounding variable that varies along with the independent variable).

### *2.1.1 Causality Bias*

A common theory-evidence coordination error and problem of internal validity (Reis & Judd, 2000) is drawing strong causal conclusions based on correlational data or judging that those who do so are correct (Burrage, 2008; Hatfield, Faunce, & Job, 2006; Rodriguez, Ng, et al., 2016; Rodriguez, Rhodes, et al., 2016). If two variables are correlated there are several possible reasons: variable a causes variable b; variable b causes variable a; variable c causes both a and b; there is an interaction such that a causes b and in turn b causes a or the correlation is spurious or coincidental.

Nonetheless, people often interpret correlational data as supporting causal claims. In fact, such reasoning is not necessarily incorrect and is appropriately prevalent amongst scientists and nonscientists, who often make inferences about likely causality based on a combination of correlated variables and a strong theory of mechanism to explain a causal link (Koslowski, 1996; Murphy & Medin, 1985; Shaklee & Elek, 1988). Even nonscientists are more likely to believe that two correlated variables are causally related when there is a plausible mechanism (Koslowski, 1996). In fact, people are very likely to seek out information about possible mechanisms when asked to draw conclusions about causality (Ahn, Kalish, Medin, & Gelman, 1995). Furthermore, people are more likely to believe causal theories when they were able to integrate multiple pieces of evidence into a coherent framework and less likely to believe them when some of the evidence could not be explained by that framework (e.g., Koslowski, Marasia, Chelenza, & Dublin, 2008). Adults also take into account other information, such as temporal or physical contiguity, amount of data/sample size and other factors in evaluating explanations for evidence (Ahn & Kalish, 2000; Ahn et al., 1995; Koslowski et al., 2008). People are also sensitive to the existence of alternative explanations (Sloman, 1994) and to the coherence of explanations (Lien & Cheng, 2000). Young children, on the other hand, seem unable to incorporate both covariation and plausible mechanisms in judgments of likely causality. For example, sixth graders in one study were insensitive to the existence or nonexistence of a causal mechanism and primarily made causality judgments based on covariation (Koslowski, Okagaki, Lorenz, & Umbach, 1989).

Scientists recognize that the inference of causation based on covariation plus a mechanistic model is not solid proof and that covariation data leave open alternative explanations. Unfortunately, for many everyday contexts individuals can easily identify potential mechanisms for many relationships and thus may readily accept causal models of correlational data even though the presumed mechanisms have little validity or are merely assumed. In other words,

> *because explanations embody prior beliefs, they have an undisputed danger: when generated from true beliefs, explanations provide an invaluable source of constraint; when generated from false beliefs, explanations can perpetuate inaccuracy.*
>
> **Lombrozo (2006, p. 466)**

Consider, for example, the assertion that "people who regularly attend religious services are healthier and live longer than people who do not attend religious services, perhaps because of the social support people receive from attending church." When asked to generate "alternative" explanations for these types of descriptions, we found that individuals tend to identify additional causal mechanisms that are consistent with the framing of the assertion (e.g., "and also maybe people have more meaning in life leading to motivation to be healthy") or experiences that are consistent with the mechanism proposed (e.g., "my mother goes to church, and when she's sick people bring her casseroles"; Durante, 2015). Unfortunately, individuals rarely generate mechanisms for other causal patterns despite explicit instructions to do so (e.g., “healthier people are more likely to be able to attend religious services regularly” or “people who are conscientious are more likely to attend religious services regularly as well as maintain good health habits”). One reason everyday scientific reasoning may be especially difficult is that causal mechanisms are easy to generate for familiar contexts (Ahn & Kalish, 2000), and once individuals generate an explanation, they are more likely to believe a claim (Glassner et al., 2005). Furthermore, the combination of a model and a causal mechanism can lead to a false sense of understanding or “illusion of explanatory depth” (Rozenblit & Keil, 2002).

In several studies, we have found that adult college students rarely notice when people inaccurately assume causality from correlational data (Burrage, 2008; Rodriguez, Ng, et al., 2016; Rodriguez, Rhodes, et al., 2016). In one study, participants read a fictional article that contained correlation/causation errors such as “…experimenters found a positive relationship between achievement motivation and job status. This study showed that people stayed in low status positions because they lacked the personal motivation to achieve…”. They rated studies that contained such errors as identical in quality to studies that had no causal interpretation error (Rodriguez, Ng, et al., 2016; Rodriguez, Rhodes, et al., 2016). Even when explicitly asked to critically evaluate such evidence, Burrage (2008) found that University of Michigan college students noted problems regarding causal inference less than 15% of the time.

### *2.1.2 Control of Variables*

Another threat to internal validity is the existence of uncontrolled explanatory variables (Popper, 1959). It is critical for people to consider whether experimental variables have been adequately controlled when evaluating scientific studies. The core concept of controlling variables is that a causal claim is only valid if a single contrast has been made between two experimental conditions. The control of variables strategy is considered a domain-general skill that relies on both the ability to create unconfounded experiments and the ability to distinguish between confounded and unconfounded experiments (Chen & Klahr, 1999). Furthermore, people should use the control of variables strategy when making inferences about an experiment; for instance, they should be able to recognize the limitations of causal claims from a confounded experiment. However, people rarely take into account whether experimental variables have been accurately controlled when evaluating evidence; even though they can do so when their prior knowledge supports the alternative interpretation, and this is particularly true for young children. In one study, even after 20 sessions of exploring the effects of multiple variables on different outcomes, adults made fewer than 75% valid causal inferences, and children made fewer than 25% valid causal inferences (Kuhn et al., 1995). There is some evidence that young children can improve their ability to use the control of variables strategy, but only after extensive training that includes both explicit instruction and probe questions (Chen & Klahr, 1999; Klahr & Nigam, 2004; Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016).

### *2.1.3 Selection Bias*

Selection bias is closely related to the control of variables, but focuses on the assignment of individuals to condition. Consider the following vignette, variations of which we have used in different studies (e.g., Rhodes et al., 2014; Rhodes & Shah, 2016a, 2016b)

> *New research is shedding light on whether meditating is effective for improving academic performance. In a recent study conducted in a large classroom, students volunteered to be part of a meditation group or a control group. Participants in the meditation group were instructed to meditate for 30 minutes every day for three months. Forty percent of the class volunteered to be in this group. The control group was required to avoid meditating during this time period. Researchers found that students in the meditation group showed a greater increase in their academic performance at the end of the semester than the control group. Researchers concluded that meditation can improve academic performance.*

Because participants self-selected to be in either the meditation group or the control group, it cannot be determined if differences in those two groups are due to the meditation manipulation or preexisting differences that lead them to select a particular condition. As a result, the evidence here does not adequately support the idea that meditation helps students study. Across numerous studies, even when asked to critically evaluate or generate alternative explanations for data, University of Michigan students and adults explicitly note these errors only about half the time (e.g. Burrage, 2008). Although our studies suggest that individuals are actually better at noticing selection bias than other errors, it is clear that there is room for improvement.

### *2.1.4 Other Threats to Internal Validity*

There are numerous additional threats to internal validity; however, we have not collected data on the ability of laypersons to identify these threats, and so we describe these in less detail. They include changes in a dependent variable as a function of development or time rather than as a function of an intervention. For example, in a cognitive training study, children might improve more from pretest to posttest on an outcome variable because they are older rather than because of the cognitive training intervention per se. Another threat to internal validity is repeated testing or measurement being responsible for change rather than an intervention. Experimenter and (in human studies) participant expectations may also pose threats to validity, especially when there is subjectivity in reporting. And other factors can also reduce confidence that an independent variable was responsible for an effect on a dependent variable, such as a historical event (e.g., it rained on the day the control group was tested). It is often impossible to discern whether or not some of these factors played a role in a study; for example, unless the experimenter had a suspicion that something like the rain made a difference, she may not even mention that it rained.

In summary, people rarely notice threats to internal validity when they are present (except selection bias, which is noticed nearly half the time, and even more in some of our studies depending on the context). Perhaps more disturbing is that many studies with other kinds of threats to validity are lauded as being of high quality; Hurley (2016) in his *New Yorker* article, for example, describes the study in question as boasting "excellent bona fides" in part because it is an RCT. By noting that a study does not contain the most salient threat to validity (and thus legitimizes it), a science writer might inadvertently suggest a more general positive evaluation of the study, even if it is not warranted.

## 2.2 Threats to Construct Validity

A *threat to construct validity* refers to whether the constructs being measured are appropriate. Another common flaw is when a claim is based on data in which the measurement of one or more variables is invalid or unreliable (Picardi & Masick, 2013). For example, a study might claim to measure creativity but uses a single self-reported question "Are you creative?" Critically evaluating evidence involves evaluating the quality of the variables and measurement. In general, people rarely attend to this issue. For example, Burrage (2008) included two vignettes that were written with no information regarding how key variables were defined. One vignette read:

> *One in four adolescents said they were abused within the past year, according to a new survey. The telephone survey of 2,000 children ages 10 to 16, suggests, "We're*

> *not doing a very good job of counting and tracking the problem," said David Finkelhor, a sociologist at the University of New Hampshire and coauthor of the study.*

In this case, the word "abuse" is not defined in any way, so it should be difficult to draw conclusions from this study. However, fewer than 20% of participants mentioned the unclear definition when asked to critically evaluate the vignettes. One of our critiques of the Roberts et al. (2016) paper about working memory interventions in children was the test used to measure academic achievement (Katz & Shah, 2016a, 2016b), which we argued may not necessarily show benefits of working memory training. This example points to the fact that expertize may be required to recognize threats to construct validity. However, one study in Burrage (2008), in which she asked graduate students in several disciplines to describe psychological data in graphs, found a somewhat contradictory effect; graduate students who were not familiar with psychology (specifically, history and engineering graduate students) were often more concerned with measurement of psychological constructs such as "creativity" than psychology graduate students who were more likely to trust that the construct was measured appropriately. One possibility is that the graduate students in our study, knowing that these constructs are often measured well, trusted the fictional "researchers" who generated the graphs (only a very brief study description was provided and participants were told that the data were fictional). In the case of the working memory training study, however, we read the entire paper and were highly familiar with the tests used to measure academic outcomes. These contradictory data suggest, however, that a little familiarity but not substantial expertise might lead to a false sense of understanding.

## 2.3 Threats to Statistical Validity

Another aspect of science evidence evaluation is understanding something about the statistical properties of the evidence and whether or not the statistical conclusions are valid. In the threat to internal and external validity examples aforementioned, the results are implied to be statistically significant and there is minimal discussion of possible problems with sample size, variance and so forth. However, it is also possible that the statistical inferences are not correct (Beaudry & Miller, 2016; Cook, Campbell, & Day, 1979). Beaudry and Miller (2016) describe three types of statistical validity concerns: (1) conducting multiple analyses until statistical significance is found, (2) using unreliable measures and (3) not having a large enough or representative sample. Scientists themselves often fall prey to the first statistical validity concern, conducting multiple analyses until statistical significance is found. Although in experimental sciences there is currently a greater awareness of such issues, very few studies as yet follow recommended procedures (such as preregistration of hypotheses and methods). Furthermore, it might actually be informative, for exploratory purposes, to test multiple posthoc hypotheses. However, it is important to keep in mind that those exploratory analyses must be replicated with studies designed explicitly to test these hypotheses. This is one possible concern with the brain training/dementia study reported here; though the sample size is large and there was random assignment, there have been numerous hypotheses tested and published for over a decade (e.g., Ball et al., 2002). These papers ought to be peer reviewed and, if appropriate, published, but it may be too early to report these findings to the general public, even if the results are deemed "preliminary". As we discuss below, people (students at least) rarely pay attention to "hedging" in reports of science evidence.

## 2.4 Threats to External Validity

Another type of error in drawing conclusions from scientific experiments is a threat to external validity—to what extent are the results meaningful in the real world, and will they apply in different contexts?

Overgeneralizing relevance of a conclusion to other related contexts or situations comprises one type of external validity error. A study of the effectiveness of physical exercise for diabetes patients, for example, will often involve a certain kind of exercise (like treadmill walking or yoga), a specific dosage (twice a week for six months), participants with some range of characteristics (low SES sedentary adults with Type 2 diabetes) and be conducted in a particular environment (the local YMCA). It is not clear whether or not the results are applicable to other kinds of exercise, dosages, participants or environments. Though we have not systematically coded our participants' responses to assess whether they ever mention these external validity issues, it is clear that such comments are generally rare. In the case of the Roberts et al. (2016) and other studies that find that cognitive training is or is not effective, it is always important to consider the potential breadth of transfer (for example, is it just Cogmed that is not effective, or all working memory interventions?). Part of the answer is dependent on results of other studies of cognitive training. The evaluation of threats to external validity, then, is also heavily dependent on prior content knowledge and also understanding that claims from single studies are not themselves conclusive unless replicable.

Another external validity error occurs when effect sizes are small even when based on a large sample; thus, although the results may be statistically significant, making changes based on the finding are not warranted because their impact will be small and doing so may be extremely costly (Domhoff & Schneider, 1999; Rosenthal, Rosnow, & Rubin, 2000). Burrage (2008) found that students rarely commented on very tiny effect sizes for which large and costly societal changes were proposed. The overgeneralization of small effect sizes is not the same as statistical validity, because the statistical evidence is not in dispute; rather, it involves the interpretation or application of the statistical evidence. One common approach to exaggerating effect sizes in science communication is to report effects of intervention in terms of proportion change (Baron, 1997). For example, in the Hurley (2016) *New Yorker* article, a primary statistic that is provided is a 48% reduction in risk (a change from 12.1% to 8.2%, also reported in the article) but people tend to consider a "48% reduction" as a bigger change than 12.1% to 8.2% (Baron, 1997). Many media reports (e.g., this article in the *Wall Street Journal* http://www.wsj.com/articles/this-brain-exercise-puts-off-dementia-1469469493) only refer to the 48% figure. Presenting risk reductions graphically also leads to people increasing their assessment of the importance of those risks (for better or worse; Chua, Yates, & Shah, 2006; Stone et al., 2003; Stone, Yates, & Parker, 1997)

A related error, that we have called "number absolutism", occurs when a specific numerical data point is believed to represent more precision than is possible (e.g., more than three cups a day of "ichemas tea" increases cancer risk, or a temperature of 100.3 degrees F or higher indicates risk of pneumonia) (Pan & Shah, 2016). Individuals who do not consider variance and measurement error are likely to make decisions that are overly reliant on specific numbers (e.g., deciding that drinking precisely 2.99 cups of ichemas tea is okay, assuming a patient must not have pneumonia because their temperature is just 100.2). In preliminary studies, we asked workers on mechanical turk to answer questions about numerical differences (paraphrased for brevity) such as "John scored a 97 on a math test, and Bob scored a 96; what is the likelihood that John is better at math than Bob"; or, "A physician prescribed 2½ pills but the patient took 2⅔; what is the likelihood that she suffered from symptoms of overdose?" We found that there were individual differences in "number absolutist" attitudes such that people who focused on specific numbers and ignored variance were more likely to do so in other contexts (though there was some separability of items that involved comparisons between two numbers as in the John versus Bob example and comparisons between a measurement and a referent as in the pill example). Taking specific numbers too seriously in contexts with a referent was associated with decisions participants reported they might make (e.g., paying a dollar more for a small bag of potato chips with five fewer calories if they were on a diet). Fortunately, a brief intervention in which participants were asked to consider issues of variance and measurement error (e.g., answering questions like "what is the likelihood that if Bob took the test again tomorrow, he'd have the same score?") led not only to improvements on transfer absolutism items, but also on a standardized test of numeracy.

Above, we characterize good science evidence evaluation in terms of the ability to identify threats to scientific validity and point to several studies that suggest that, though individuals are often capable of identifying threats to validity they frequently ignore these threats in everyday contexts. Much of the research base focuses on people's ability to understand and recognize threats to internal validity, whereas we were able to find much less research that focuses on people's attention to external validity concerns. In the next sections, we discuss why readers of scientific evidence often fail to detect threats to scientific validity, the factors that affect whether or not they detect these threats, and potential methods to increase people's ability to do so.

## 3. HEURISTIC (SYSTEM 1) THINKING VERSUS ANALYTIC (SYSTEM 2) THINKING

Why do readers of scientific evidence fail to detect threats to validity? One proposal is that people rely on *heuristic thinking* (or "System 1") rather than *analytic thinking* ("System 2") in the context of evidence evaluation (Amsel et al., 2008). Heuristic thinking is fast, frugal, automatic, emotional and unconscious (Kahneman, 2011); when costs of error are low and correction is easy, saving effort through heuristic thinking is appropriate. Analytic, or System 2 thinking, requires much slower, substantial effort, and due to limited cognitive resources, it occurs less frequently. Relying on heuristic thinking can lead people to "speed" through processing and make errors in reasoning. However, when people reason analytically, they are more likely to identify threats to scientific validity. Unfortunately, k-12 students, college students and the laypublic all tend to rely on heuristic thinking when reading media articles about scientific studies (e.g., Norris & Phillips, 1994). For example, when Rodriguez, Ng, et al. (2016) and Rodriguez, Rhodes, et al. (2016) asked college students to discuss scientific evidence about everyday topics, thinking analytically was uncommon, and most participants discussed personal opinions and experiences. When explicitly asked to *critically* evaluate scientific evidence, they did so more frequently. In a similar study (Kosonen & Winne, 1995), college students performed no better than 7th or 10th grade students when asked to evaluate the validity of an experimental study. It was only when the researchers explicitly prompted college students to think critically that they became better at evaluating experiments compared to middle and high school students. In other words, people often have the capability of analytically evaluating evidence, but do not typically do so without explicit instruction. Rather, they rely on a set of heuristics, outlined below, that often lead to reasoning errors.

## 3.1 Appeal to Authority

When people do not have sufficient background knowledge about scientific issues, they may judge scientific claims in part by deferring to scientific authority (Bromme & Goldman, 2014). For instance, Brossard and Nisbet (2007) found that trust in scientific authority was the strongest predictor of support for agricultural biotechnology, stronger even than knowledge about the science behind the issue. In particular, schools in the United States emphasize the objectivity and neutrality of science, which may further promote the authoritative status of scientists (Brossard & Nisbet, 2007). Although it is good for students to trust scientists in general, deference to scientific authority can lead to an oversimplified view of the scientific process. For instance, not understanding that novel findings are tentative and that scientific experts often disagree about tentative findings. Additionally, because it is heuristic, people may rely on deference to authority as a way to avoid analytical thinking, particularly when a description of a study is perceived as difficult to comprehend (Scharrer, Bromme, Britt, & Stadtler, 2012).

## 3.2 Bias Towards Certainty

In another example, both high school students (Norris & Phillips, 1994) and college students (Norris et al., 2003) overestimated the certainty of findings from science news reports and were biased to interpret statements from the reports as true versus false. Students from both studies found it particularly difficult to interpret hedging statements, such as "X is likely to be true", "uncertain of the truth status of X" or "X is unlikely to be true". The authors speculated that students overestimated the certainty of scientific findings for several reasons. First, people are generally biased to seek out certainty/avoid uncertainty; thus, students may have perceived nuanced statements as more certain than they really were to feel a sense of closure, resolution, etc. Second, textbooks typically present scientific information in black and white terms and with a limited historical context, perhaps giving the impression that science is itself black and white.

## 3.3 Relying on Fluency

In addition to overestimating the certainty of findings from scientific news articles, students also overestimated their own understanding of the studies (e.g., Norris et al., 2003). In general, people tend to equate the comprehension difficulty of a text with their ability to recognize the words in the text (Pressley & Wharton-McDonald, 1997). Because the articles used in the Norris and Phillips (1994) and Norris et al. (2003) studies were written in layman's terms using simple wording, students may have underestimated the complexity of the studies described in the articles. Additionally, despite experiencing difficulty interpreting the studies, students were very good at locating relevant information from the article. Due to the high readability and ease of finding specific text in the articles, students may have been overly confident in their ability to understand the articles. Thus, overconfidence in one's ability to comprehend a scientific news report may reduce critical thinking.

## 3.4 Avoiding Conflicting Information

Many journal articles present science news in an oversimplified, black and white manner; for instance, they may only promote one viewpoint or finding. Thus, students may not consider alternative explanations for findings because conflicting information is either not available or must be integrated across multiple sources (Bromme & Goldman, 2014; Stadtler, Scharrer, Brummernhenrich, & Bromme, 2013). However, evidence suggests that students are more likely to critically evaluate scientific evidence and generate explanations for phenomena when they process conflicting information at a deep level (Mason, 2000; Stadtler et al., 2013). For instance, Mason (2000) found that middle school students who acknowledged and understood a piece of information that conflicted with their belief in a school-taught theory (e.g., that dinosaur extinction was caused by a collision between Earth and an asteroid) were more likely to accept an alternative theory or decrease their belief in the original theory.

### 3.5 Reasoning From Prior Experience and Gut Feelings

Finally, people's prior experience and emotional response to content also influence their evaluations of evidence; when evidence "feels right" (often when it is belief congruent) or makes one feel good (because it is what people desire), they are likely to judge the evidence as being of high quality without further consideration. This issue is discussed further in the section below on the content and communication of evidence.

In summary, one reason that people tend to ignore threats to scientific validity is that they rely on relatively low-effort heuristic strategies. Below, we discuss the conditions under which individuals are more or less likely to rely on heuristic versus analytic processes when evaluating science evidence.

## 4. CONTENT AND COMMUNICATION OF EVIDENCE

The type of information presented as well as the way in which evidence is presented has an impact on evidence evaluation. Below, we outline several factors that influence people's reasoning.

### 4.1 Belief Consistency

When individuals are faced with evidence that is incongruent with their existing beliefs, they are more likely to activate an analytic mode of thinking than when information is congruent with their beliefs (D. Evans, 2003; J.S.B.T. Evans, 2003; Evans & Curtis-Holmes, 2005; Klaczynski, 2000; Kunda, 1990; Sá et al., 1999; Sinatra et al., 2014; see Nickerson, 1998 for an early review). This phenomenon was demonstrated in a classic study by Lord et al. (1979). Participants were asked to evaluate empirical studies about controversial topics that were consistent or inconsistent with their personal beliefs. For example, some participants read a report summarizing a study about the effectiveness of the death penalty. In one report, the empirical finding described was that in 11 out of 14 states with capital punishment, murder rates dropped between the year that it was adopted and the year following. People who supported capital punishment tended to rate that study as being of high quality, whereas people who were against capital punishment were far more critical of the findings, and were more likely to generate alternative explanations for the data. People tend to use more sophisticated reasoning strategies (Ditto & Lopez, 1992), appropriate statistical principles such as base rates, and the law of large numbers (Ginossar & Trope, 1987; Sanitioso & Kunda, 1991) when they have a desire to disprove a claim. Returning to the issue of brain training, it is virtually certain that the first author's personal hope for at least some kind of cognitive training intervention to be effective for the many populations who might benefit (aging adults, ADHD children, brain-injured individuals, fighter pilots who need to stay sharp) affects his/her evaluation of studies with negative outcomes with more scrutiny than those with positive outcomes.

## 4.2 Presence of Anecdotes

In general, people tend to pay more attention to anecdotal information compared to statistical information in the context of decision making (Betsch, Ulshöfer, Renkewitz, & Betsch, 2011; Fagerlin, Wang, & Ubel, 2005; Sanders Thompson, 2013; Slater & Rouner, 1996; Ubel, Jepson, & Baron, 2001). For example, individuals are more likely to select classes based on the recommendation of an individual face-to-face encounter than based on a representative sample of student ratings (Borgida & Nisbett, 1977). Likewise, women who watched narrative videos reported fewer barriers to mammography than women who watched didactic informational videos (Kreuter et al., 2010). In a recent study, Rodriguez, Ng, et al. (2016) and Rodriguez, Rhodes, et al. (2016) addressed the extent to which anecdotal information affected evaluations of scientific evidence. Across two studies, college students read and evaluated a set of fictional scientific news articles. These articles provided summaries of research studies in psychology. However, all of the articles made unwarranted interpretations of the evidence, such as making strong conclusions from weak evidence or implying causality from correlational results. Students were randomly assigned to the anecdote or control conditions. For the anecdote condition, each news article contained personal narrative that corroborated the results of the research study. The control condition news articles only contained the summaries of the research studies (Study 1) or a descriptive text alongside these summaries (Study 2). Even after controlling for important variables, such as level of college training, knowledge of scientific concepts and prior beliefs, the presence of anecdotal stories significantly decreased students' ability to provide scientific evaluation of the studies.

## 4.3 Microlevel Evidence

Readers interested in the social and behavioral sciences have likely noticed that over the past two decades, there has been a steady increase in the number of news headlines that feature neuroimaging results. This increase not only reflects the growth of fields like cognitive neuroscience, but also suggests that neuroscience information may be appealing to readers. To test this hypothesis, researchers have examined how the presence of neuroscience information in a behavioral science report influences scientific evaluations. When neuroscience is mentioned in an explanation or study description, people tend to rate the explanation/description as being higher in quality (Beck, 2010; Fernandez-Duque et al., 2015; McCabe & Castel, 2008; Weisberg et al., 2008). Reports that include neuroimaging evidence often include brain pictures as well, perhaps showing results of fMRI analyses, which could also have an effect on scientific evaluations. However, researchers have found that the presence of brain images alone does not seem to have a strong effect on reasoning (Farah & Hook, 2013; Hook & Farah, 2013); rather, textual neuroscience information appears to be most influential. One reason this may be the case is that textual neuroscience information is often incorporated into the explanation of a phenomenon and may appear to provide stronger support for a claim. The catch, however, is that researchers find that *irrelevant* neuroscience information, which has no bearing on the study being described, also improves ratings of explanations and study quality (Weisberg et al., 2008). In this way, neuroscience information can have a "seductive" appeal, such that people will judge studies containing neuroscience information more favorably regardless of whether they actually understand it.

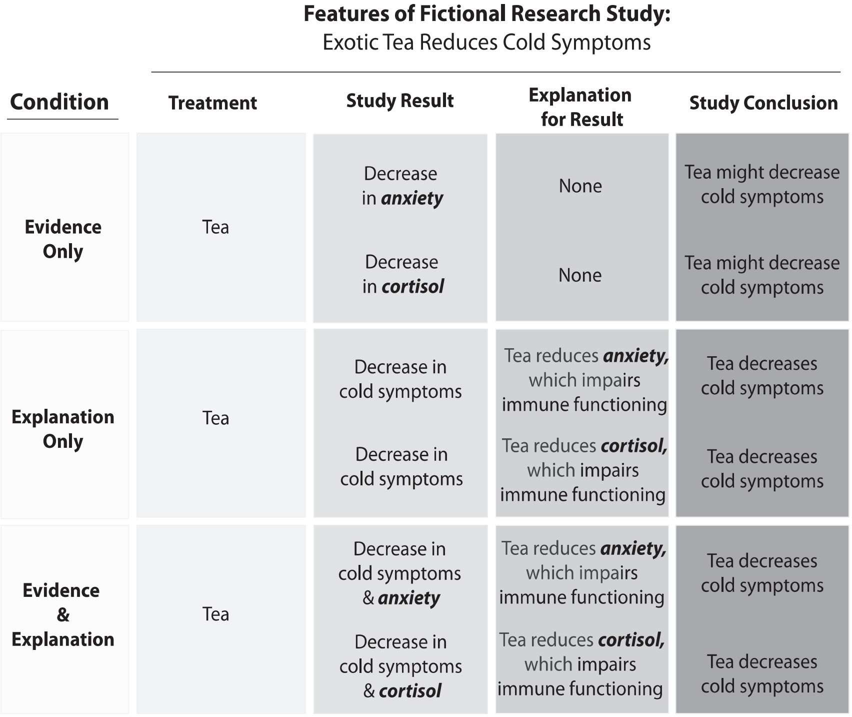
In recent studies we extended findings associated with the influence of neuroscience on evidence evaluation by taking into account people's prior beliefs about the claims. Adult participants, recruited via Amazon's Mechanical Turk ($N = 400$), were first asked to indicate whether they believed that listening to music while studying was beneficial. Next, they read a fictional news article describing a research study that found positive effects of listening to music while studying. All participants read some introductory text followed by a study description, and the introductory text was manipulated to either contain neuroscience information or not. For half of the participants, the news article began with the following neuroscience text: "Years of neuroscience research have made it clear that listening to music is associated with distinct neural processes. Functional MRI scans reveal that listening to music engages cortical areas involved in music and sound perception, and this activation is thought to be present even while doing other tasks, such as studying or learning new information". For the other half of the participants, the introductory text contained the same number of words but no neuroscience: "Although some people prefer to work in silence, many people opt to listen to music while working or studying. In fact, due to the increased mobile access to music, a brief glimpse into a library or coffee shop will reveal dozens of individuals poring over their laptops and books with earphones wedged into their ears". Following the introductory text, a research study was described that contained a methodological flaw; specifically, participants in the study had self- selected the condition (music listening or no music) they were to be in, so the study was not properly controlled and contained a "sampling bias" threat to validity.

Overall, participants rated their understanding of the mechanisms underlying the music phenomenon higher in the neuroscience condition than in the control condition, despite the fact that the neuroscience information did not actually provide any concrete underlying mechanisms for the effect of music on study performance. Furthermore, participants also rated the article as being of higher quality when the neuroscience information was present, at least when participants had neutral prior beliefs.

Recent research suggests that neuroscience is not alone in its seductive qualities. People seem to be attracted to genetic explanations for phenomenon over cultural or behavioral explanations (Dar-Nimrod & Heine, 2011). People also demonstrate a preference for data when they are presented in graphic form or when there are meaningless chemical formulas (Tal & Wansink, 2016). People tend to evaluate studies that contain meaningless mathematics equations in abstracts as being of higher quality than when the same abstract is not accompanied with a mathematical formula (Eriksson, 2012). One explanation for these findings could be that neuroscience information, genetic information, chemical formulas and mathematical equations are all consistent with our representations of "science", and so the presence of these things may inspire more faith in the results. Another possibility could be that, when judging explanation quality, the more reductive the explanation is, the more explanatory it appears. Hopkins et al. (2016) demonstrated that, across a range of phenomena from a variety of science domains—including social science, psychology, neuroscience, biology, chemistry and physics—people preferred reductive explanations, regardless of whether those explanations were logically relevant to the phenomena being described.

In a recent study (Rhodes, 2015; Rhodes & Shah, 2016b), we further examined the extent to which people prefer reductionist, microlevel explanations for data more generally compared to more holistic, macrolevel explanations. Specifically, we examined whether the influence of reductionist information depends on the level of causal information provided. Participants were asked to read a series of brief study descriptions from a range of sciences and rate how supported the conclusion was. Participants (n = 330) were in one of three conditions, (1) Evidence Only, in which no causal explanation was given, (2) Explanation Only, in which a causal explanation was provided in the explanation but lacked supporting evidence or (3) Evidence and Explanation, in which the causal chain for a phenomenon was explicitly spelled out through the evidence and explanation provided (An example of each of these conditions can be seen in Fig. 1).

We asked all participants to compare two sets of studies—those that contained microlevel evidence versus those that contained macrolevel evidence (e.g., tea decreases anxiety and in turn decreases cold symptoms) versus microlevel evidence (e.g., tea decreases cortisol and in turn decreases cold symptoms). We found that participants tended to prefer microlevel evidence over macrolevel evidence. Importantly, this effect was pronounced when the sample contained human participants and there was only an implicit mechanism. These results suggest that reductionist information



**Features of Fictional Research Study:**
Exotic Tea Reduces Cold Symptoms

| Condition | Treatment | Study Result | Explanation for Result | Study Conclusion |
|---|---|---|---|---|
| **Evidence Only** | Tea | Decrease in ***anxiety*** | None | Tea might decrease cold symptoms |
| | | Decrease in ***cortisol*** | None | Tea might decrease cold symptoms |
| **Explanation Only** | Tea | Decrease in cold symptoms | Tea reduces ***anxiety,*** which impairs immune functioning | Tea decreases cold symptoms |
| | | Decrease in cold symptoms | Tea reduces ***cortisol,*** which impairs immune functioning | Tea decreases cold symptoms |
| **Evidence & Explanation** | Tea | Decrease in cold symptoms & ***anxiety*** | Tea reduces ***anxiety,*** which impairs immune functioning | Tea decreases cold symptoms |
| | | Decrease in cold symptoms & ***cortisol*** | Tea reduces ***cortisol,*** which impairs immune functioning | Tea decreases cold symptoms |

**Figure 1** Mechanistic manipulations for example research scenario. Location of mechanistic information (microlevel vs. macrolevel) in each condition is bolded and italicized. *From Rhodes, R. E. (2015).* The influence of reductionist information on perceptions of scientific validity *(Doctoral dissertation). Ann Arbor, MI: University of Michigan.*

may be most likely to influence scientific reasoning when there is a lack of other causal information available. An example of such a situation is newspaper headlines which, due to space constraints and the goal of attracting readers, typically highlight a surprising finding with no causal explanation. In these cases, people may be at higher risk for being "seduced" by reductionist details, regardless of their relevance or logical relation to the study being reported.

## 4.4 Presence of Graphics

In addition to the influence of neuroscience information, our group has evidence to suggest that the mere presence of a data visualization (i.e., scatterplot) can shape the conclusions people draw from scientific information (Ibrahim, Seifert, Adar, & Shah, 2016). Based on decades of research, the World Health Organization (WHO) has deemed GMOs as safe for human consumption; however as mentioned earlier, a large proportion of the public still dispute this claim. A total of 186 adult participants, recruited through Amazon's Mechanical Turk, were presented with a fictional popular science article that discussed the research on the safety of GMOs. Participants reported their previous beliefs on the safety of GMOs before being presented with any materials. Participants were randomly presented with one of two visualization conditions (scatterplot or no scatterplot) and one of two text conditions (consistent versus inconsistent evidence). In the consistent evidence text, participants were informed of the decades of research and the WHO designation of GMOs as safe then presented with the results of one *fictional* study that found no correlation between GMO consumption and annual doctor visits. In the inconsistent evidence text, participants were first informed of the safety of GMOs in the same manner as stated in the consistent text condition, but in this condition the *fictional* study found a positive correlation between GMO consumption and annual doctor visits (that is the greater the GMO consumption, the higher incidence of doctor visits per year). For those presented with the inconsistent evidence text, we found those presented with a data visualization were more likely to make causal inferences based on the results of the *new fictional* study than those who were not presented with a data visualization. To elaborate, participants presented with *mixed evidence and an associated data visualization* were more likely to ignore the previous wealth of evidence deeming GMOs as safe, and more likely to make a causal inference based on the one *fictional* study presented (i.e., that GMO consumption causes illness). In addition, those whose previous beliefs were that "GMOs are unsafe" made more causal inferences between GMOs and health than those whose previous beliefs were either "neutral" or "GMOs are safe". Regardless of the influence of previous beliefs, the presentation of a data visualization affected all three groups in a similar manner (leading to more causal inferences in the inconsistent evidence text and data visualization condition). These results demonstrate that the presentation of a visualization can easily influence individuals to ignore past evidence and rely on new evidence under this circumstance. It is likely that the presentation of the visualization, adds credibility to the results of the new study, especially if individuals are relying on System 1 thinking. This pattern of results is especially worrisome since it demonstrates how easily people can be convinced by new data, regardless of the actual scientific merit of the result. Our findings also support findings of previous research mentioned earlier that has demonstrated a similar effect of bar graphs (Tal & Wansink, 2016), scientific formulas (Tal & Wansink, 2016), and neuroscience information (Rhodes et al., 2014).

As the discussion above suggests, several features of how science evidence is presented in the media affect whether or not individuals are able to provide high quality evidence evaluations. In addition to the characteristics of media articles, characteristics of the reader also influence evidence evaluation. These factors are discussed in the next section.

## 5. INDIVIDUAL DIFFERENCES

Several individual difference factors are associated with evidence evaluation skills. Note that these are primarily dispositional, and that everyday evidence evaluation is less related to cognitive abilities (i.e., IQ) than these dispositional factors. Furthermore, while any individual evidence evaluation task is influenced by domain specific knowledge, domain general factors do seem to play a role across a wide variety of circumstances as well.

### 5.1 Cognitive Flexibility

Although prior beliefs tend to affect the depth of reasoning one uses to evaluate evidence, they do not affect everyone. Some individuals are very good at reasoning in a more objective way, independent of their own personal beliefs. The tendency to do this can be measured by the AOT scale (Stanovich & West, 1997). This 41-item scale asks about people's ability to think flexibly and be open to new information, regardless of what they personally believe. A high score on the AOT scale reflects more sophisticated thinking dispositions; specifically, it indicates a motivation to have accurate beliefs, even if that means changing one's current beliefs. Although

AOT performance is correlated with cognitive ability, the two constructs are separable; performance on the AOT predicts data-driven thinking during argument evaluation tasks, even after partialling out the variance associated with cognitive ability (Stanovich & West, 1997). People who score highly on the AOT scale are more likely to reason in a data-driven, as opposed to a belief-driven way (Stanovich & West, 1997).

## 5.2 Cognitive Reflection

The cognitive reflection test (CRT; Frederick, 2005) is a widely used measure of one's ability to suppress an intuitive response, resulting from heuristic processing, in favor of a more deliberate response. This test consists of three items that tend to elicit automatic, but incorrect answers. The correct answer requires more thinking than it initially seems. The CRT is correlated both with cognitive ability (Frederick, 2005; Toplak, West, & Stanovich, 2011) and rational thinking measures, such as syllogistic reasoning problems (Toplak et al., 2011). CRT performance also predicts performance on many heuristics and biases tasks (Cokely & Kelley, 2009; Frederick, 2005; Toplak et al., 2011). Most importantly, Toplak et al. (2011) found that the CRT predicts rational thinking and performance on heuristics and biases tasks after partialling out the variance associated with assessments of intelligence, thinking dispositions, executive functions and cognitive skills. Thus, people who score highly on the CRT can be categorized as people who are more likely to engage in rational, analytic thinking. Participants who score higher on the CRT are less likely to be religious or believe in paranormal phenomena (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012).

## 5.3 Need for Cognition

NFC (Cacioppo & Petty, 1982) is defined as a disposition towards thinking; people high in NFC report enjoying difficult or effortful cognitive activities. NFC is measured by an 18-item scale, a sample item is "I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought". NFC is associated with deliberate thinking and more effortful higher quality evaluations of evidence. In a recent study (Minahan & Siedlecki, 2016), for example, individuals who were low in NFC were less able to notice circular scientific explanations.

## 5.4 Faith in Intuition

The work on faith in intuition finds that some individuals are more willing to trust their hunches or intuitions than others (Epstein, Pacini, Denes-Raj,

& Heier, 1996). Faith in intuition is typically measured via a 15-item survey with questions such as "I hardly ever go wrong when I listen to my deepest feelings to find an answer" or "Using logic usually works well for me in figuring out problems in my life". Faith in intuition is associated with heuristic rather than statistical judgments on quantitative judgment tasks (Shiloh, Salton, & Sharabi, 2002).

## 5.5 Epistemic Beliefs

Epistemic beliefs refer to people's knowledge about knowledge: what do you know about the world, what are good sources of information about the world and how certain do you feel about what you know (Hofer & Pintrich, 1997; King & Kitchener, 1994; Sandoval, 2005; Schraw, Bendixen, Dunkle, Hofer, & Pintrich, 2002). People with sophisticated epistemic beliefs have an attitude in which they try to ensure that their beliefs and values represent an accurate reflection of what is known about the world (Baron, 2008; Stanovich, 2009; Stanovich & Stanovich, 2010). Thus, individuals with sophisticated epistemic beliefs tend to be more critical about scientific evidence (Bromme & Goldman, 2014). A commonly used scale for measuring epistemic beliefs is the Epistemic Beliefs Inventory (Schraw et al., 2002). It includes questions that assess the extent to which individuals defer to authority, belief that there are often complex rather than simple answers to questions and so forth.

## 5.6 Numeracy and Statistical Reasoning Skills

Numeracy or statistical literacy refers to the ability to understand mathematical concepts such as probabilities and percentages (and not necessarily being able to perform computations); statistical reasoning typically involves being able to reason about statistical concepts such as the law of large numbers or the effect of outliers. There are several well-known measures of numeracy, most of which were developed in the context of medical decision making research. Some common measures include the Berlin Numeracy scale (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012), a scale developed by Schwartz, Woloshin, Black, and Welch (1997) which was then expanded by Lipkus, Samsa, and Rimer (2001). The Subjective Numeracy Scale developed by Fagerlin et al. (2007) asks individuals about their own numerical competency. There are also several measures of statistical reasoning, including the Statistical Reasoning Assessment (Garfield, 2003); this 20-item assessment measures understanding of basic statistical concepts like probability, sampling variability, as well as rejection of misconceptions, such as about the law of large numbers. The base-rate conflict task trades off stereotypes and base rates with items like the classic Kahneman and Tversky "Linda the bank teller" problem (De Neys & Glumicic, 2008). Individuals with poor statistical reasoning skills may not have a good understanding of the law of large numbers, the idea that one can be more confident about observations the larger the sample size (Fong, Krantz, & Nisbett, 1986; Kahneman & Tversky, 1972; Sedlmeier & Gigerenzer, 1997), or understand the notions of variability/measurement error/randomness (Garfield, 2003; Nisbett, Fong, Lehman, & Cheng, 1987).

### 5.7 Domain Knowledge

Knowledge about the content of a study helps readers evaluate scientific evidence in a sophisticated manner and, in general, domain-specific factors have a substantial impact on scientific reasoning (Schunn & Anderson, 1999). No one can be an expert in every scientific topic, but everyone can at least learn to apply their general skills to evaluate the validity of a research design and the extent to which conclusions are supported by evidence (Bromme & Goldman, 2014).

One highlight of this review of individual differences is that most are dispositional, focusing on people's flexibility or open-mindedness, their motivations towards thinking and reflection and their epistemological beliefs; as with many aspects of human reasoning, general cognitive abilities seem to not play as big a role as these dispositional factors (Stanovich, 2009).

## 6. BEYOND HEURISTIC VERSUS ANALYTIC THINKING: SPECIFIC EVIDENCE EVALUATION SKILLS

Activating analytic thinking and even skepticism may not be adequate for understanding threats to validity when evaluating scientific evidence in everyday contexts. When asked to be critical of evidence, college students tend to provide relatively general "knee-jerk" criticisms that could apply to virtually any study rather than identifying specific threats to validity. For example, participants frequently refer to superficial methodological concerns, such as "outliers could also affect the results", or the sample could be larger or "more diverse" (Durante, 2015). In fact, the Hurley (2016) article about dementia and cognitive training discussed earlier includes the very quote, oft-repeated in undergraduates' critical evaluations of scientific evidence, "We need to see it confirmed and replicated in a larger and more diverse population". Though these responses require somewhat more analysis than noncritiques (e.g., "the results make sense"), they do not actually involve reference to key threats to validity, and they are often so generic that they could be applied to virtually any study. What other factors limit reasoning about evidence?

## 6.1 Poor Analysis of Covariance Reasoning

One specific skill that people need to develop is referred to by Klaczynski, Gordon, and Fauth (1997) as "analysis of covariance (ANCOVA) reasoning". ANCOVA reasoning involves considering possible third variables and possible confounds, that might be responsible for a relationship between two variables. As an example, Klaczynski et al. (1997) write:

> *Consider now a second person who notes that business professors have generally higher life satisfaction than history professors, but who also ignores the discrepant salaries and teaching loads of the two groups and draws the conclusion that business is an inherently more enjoyable field than history.*

People tend not to notice this reasoning error (Schaller, 1992); however, they are more likely to notice it when they spend more time thinking about covariance (Schaller & O'Brien, 1992). ANCOVA reasoning is relevant in "control of variables" contexts (Klahr & Nigam, 2004), and more generally, ANCOVA reasoning is associated with generating alternative explanations for any empirical evidence. As discussed in the section on causation bias earlier, people have difficulty generating alternative explanations for data (Durante, 2015).

## 6.2 Identity-Protective Cognition

Kahan et al. (2013) point out that merely invoking System 2, analytic thinking does not explain why people often have vastly different interpretations of evidence. They found that when people were given a belief-neutral study (about a new skin rash ointment), those with greater numeracy were better able to evaluate the results of a study (presented in a 2 × 2 contingency table and requiring conversion to percentages). However, when people were given a politically polarizing study about gun control, greater numeracy resulted in *worse* performance. When data were in disagreement with their beliefs (crime increased with greater gun control for liberals and disagreement with greater gun control for conservatives) people took advantage of their greater numeracy to essentially "slam" the findings; when the data were consistent with their beliefs, they were generous. Thus, greater numeracy resulted in more polarization, a phenomenon they refer to as "identity-protective cognition".

This situation probably explains polarization of cognitive science researchers regarding certain topics, like cognitive training (or, say, the effect of bilingual experience on cognitive development). Scientists are even better able to use their content knowledge and quantitative skills to critically evaluate some subsets of the studies in a particular field and give a pass to other studies. Compounding this is that in systematic reviews and even meta-analysis, studies that are included differ substantially. Katz and Shah (2016a), for example, compared five reviews and meta-analyses. Though meta-analyses putatively use "objective" means for identifying studies, small differences in keywords, or date range, turned out to have a fairly substantial impact on the studies included in them. Systematic reviews, which do not often use objective means for collecting studies, include even a more diverse set of studies. We compared five reviews and meta-analysis as an example and found that of the 110 total studies that were included in at least one of the five studies, the numbers of those studies in each analysis were 21, 22, 28, 46 and 55. Not surprisingly, each analysis drew different conclusions regarding the impact of cognitive training.

## 7. IMPROVING EVIDENCE EVALUATION

The majority of this paper focuses on characterizing a problem; people have difficulty evaluating science evidence in everyday contexts. Here, we discuss some research on improving evidence evaluation. In general, two approaches are taken to improve evidence evaluation. Debiasing attempts to improve people's evaluations of an individual piece of evidence or set of evidence (e.g., helping people understand that vaccines are safe; Horne, Powell, Hummel, & Holyoak, 2015), whereas education involves teaching people strategies for improving evidence evaluation skills more generally.

### 7.1 Debiasing

When people are initially given misinformation and then that misinformation is corrected, they do not update their beliefs unless the new information is consistent with their own prior beliefs. For example, when the erroneous claim that Iraq possessed weapons of mass destruction was corrected with new information, only those whose political beliefs were supported by the new information changed their minds; others became even more entrenched in their original beliefs (Nyhan & Reifler, 2010). Decision scientists have thus been very interested in finding out how to convince people to take into account new evidence and correct misinformation (Lewandowsky et al., 2012; Schwarz, Newman, & Leach, 2016).

To replace misinformation with new information, it is central that people can generate alternate causal models (Johnson & Seifert, 1994;

Seifert, 2002). However, simply asking people to generate alternative explanations on their own is not always an effective technique, because these alternative explanations are not yet integrated into the causal model and therefore they are less accessible (Schwarz, Sanna, Skurnik, & Yoon, 2007). In fact, our own work (see causality bias aforementioned) finds that people are not very good at generating alternative explanations for data (Durante, 2015). One way to make people consider alternatives is tell them that others are skeptical of the evidence, or to tell them that the individual who presented the original evidence is somehow biased (Schul, 1993; Schul, Mayo, & Burnstein, 2008). Another approach that is somewhat successful is having people go through "challenge interviews" in which they think out loud as they reason about evidence, leading to better ability to come up with counterarguments (Prasad et al., 2009). Similarly, asking people to take the position of a skeptic or opponent and identify arguments and counterarguments is also effective (Kuhn & Udell, 2007). Unfortunately, these types of debiasing techniques have limited success. In one study, for example, Munro (2010) found that reading belief-disconfirming evidence, in a scientific abstract, resulted in increased rejection of science itself.

One approach that has been somewhat successful is changing the content or framing of the argument. For example, Horne et al. (2015) found that presenting evidence surrounding the benefits of vaccines and the risk of communicable diseases was much more effective for refuting myths about vaccinations compared to simply trying to present evidence that vaccines are safe. In general, appealing to personal values (e.g., protecting one's own children) is more effective than highlighting the quality of the counterargument. Of course, this approach addresses the question of how to change people's minds more so than the question of improving their evidence valuation per se.

We have attempted to help students reason about correlational data by presenting data in the context of an animated scatterplot (Gao, Seifert, Shah, & Adar, 2016) in that individual data points start at the origin and then move towards their correct positions. Participants viewed one of four conditions: scatterplot presented statically, both x and y animated simultaneously (i.e., so it looks like points are moving diagonally), x animated first then y (i.e., points move along the x-axis then upwards), or y animated first and then x. We found that animating both variables together reduced the percentage of causal interpretations compared to the other three conditions. In other words, the animated visualization itself highlighted the correlational nature of the data. Although our earlier discussion pointed to the possible risks of presenting data visualizations along with misleading information (Ibrahim et al., 2016), it is possible that when used appropriately, data visualizations can be used to increase attention to new information as well as support the development of more correct conclusions.

Visual representations can also help students understand and evaluate causal models and generate alternative explanations (Oestermeier & Hesse, 2000). In an example closely related to scientific evidence evaluation, Easterday, Aleven, Scheines, and Carver (2008) developed a policy deliberation tutor. College students were asked to evaluate a policy conclusion and relevant evidence (e.g., should junk food advertising be limited; people who watch more advertisements for junk foods are more likely to be obese). Participants first read through some example scenarios (just text with causal models highlighted, with diagrams that represented accurate causal models, or with a tool for diagramming causal models). Next, participants were given a transfer problem with a new scenario. Students who had received the diagramming tool (even though they did not have it for the transfer problem) were most likely to generate alternative causal models (e.g., maybe it is TV watching that causes reduction in exercise and that is the real cause of obesity) compared to those presented with other materials. Thus, working with the causal modeling tool served as a scaffold for helping students learn to generate alternative models and even led to improvements on transfer problems.

## 7.2 Teaching Scientific Reasoning

There are numerous current approaches to teaching scientific reasoning, most of which take place in the science classroom. Given that the focus here is on adult, layperson scientific reasoning, we only briefly discuss some approaches to teaching k-12 students to develop scientific inquiry skills with the hope of application to everyday evidence evaluation.

Kuhn has argued for the importance of teaching and learning science via argument (Kuhn, 2010). In one-year-long curriculum, she had students examine several topics for approximately 7—8 weeks. For each topic, students first form groups and generated arguments to support a position on one side of a controversial topic and predict what some possible counter-arguments might be. Students then discussed the topic with a student on the opposing side via a computer interface (so that the arguments remain visible) and reflected on their arguments in writing. After several sessions, they met together with their group to plan a formal debate. The topics were social and political rather than traditional science topics (for example, one topic posed "Should a misbehaving student be expelled from school?"). Of course, these topics require searching for and evaluating evidence. She has used the same approach with more traditional science topics (such as the extinction of dinosaurs; Kuhn, 2010). Both curricula led to improvements in reasoning about evidence more generally, though the science context applied more to social science than vice versa (Kuhn, 2010). In another study, she tested a smaller scale social science 3-week project-based inquiry curriculum with low-SES children as they addressed a real-life social science question about figuring out what is effective for reducing teen crime. Students who participated in the curriculum improved in their understanding of control of variables on both standardized measures, as well as in an interview context, whereas students who merely observed the curriculum in practice did not show improvement (Jewett & Kuhn, 2016).

Numerous studies have focused specifically on teaching "control of variables", one of the key necessities for scientific validity (Klahr & Nigam, 2004; Kuhn & Dean, 2005; Lorch et al., 2014). Schwichow et al. (2016) published a meta-analysis that includes 72 studies whose goal is to teach control of variables. Overall, control of variables instruction is fairly effective, with an average effect size of 0.62. This study did find several features of control of variables instruction helped ensure students' understanding and transfer of this concept. Specifically, effect sizes were larger if the instruction involved presenting students with demonstrations of good experiments. In addition effect sizes were also larger when instruction involved activating "cognitive conflict" by presenting flawed experiments with obvious alternative explanations and thus highlighting the importance of control of variables. Control of variables training has successfully been shown to transfer to some everyday science reasoning tasks—specifically, children who received control of variables training were better able to critically evaluate science fair posters and noticed when studies did not appropriately control for variables (Klahr & Nigam, 2004).

Nisbett et al. also focused on teaching individual reasoning strategies or concepts (e.g., law of large numbers). In a classic study, Fong et al. (1986) found that a combination of explicit, rule-based training plus examples was the most effective for teaching the law of large numbers in a manner that transferred to everyday science reasoning contexts. Furthermore, students who had formal statistics training were best able to apply statistical reasoning principles to everyday contexts. In fact, students who majored in a traditional science domain (chemistry) did not do as well on everyday science reasoning contexts compared to students who majored in social sciences with statistics training (psychology). These results suggest the promise of teaching science reasoning for transfer to everyday contexts by using a combination of rule- and example-based approaches; ideally, these examples should include everyday contexts (see also Fong & Nisbett, 1991). This approach towards teaching reasoning has been successful in several other content areas. For example, Schoenfeld (1979) showed that explicitly teaching five problem solving heuristics and practice problems was more successful than merely having students solving the same set of practice problems without heuristics training.

Another important approach to teaching scientific reasoning is to focus on integrating science activities with statistical literacy. Lehrer et al. had students do authentic measurement activities to learn about things like precision of measurement, error, variability and so forth. Students who experienced these activities essentially reinvented important statistical principles (Lehrer, Kim, & Schauble, 2007; Lehrer & Romberg, 1996; Lehrer & Schauble, 2004). A similar approach to slightly more sophisticated/higher-level statistical reasoning activities has been developed by Sedlmeier for adults (e.g., Sedlmeier, 2002).

Teaching students "model-based thinking" via rich simulated scenarios in which they can manipulate variables and evaluate the effects of these changes on other variables are excellent for teaching students about complex systems (e.g., Marx, Blumenfeldt, Krajcik, & Soloway, 1997; Schwarz et al., 2009). Such rich inquiry activities give students a much deeper understanding of notions of explanation and complex interactions between variables; students who experience these activities think much less simply about evidence and complex systems (Windschitl, Thompson, & Braaten, 2008). However, it is not clear how they apply these inquiry processes to everyday science contexts such as social science, health and behavior.

As a whole, the science and statistics education studies outlined here suggest some promising strategies: identifying and teaching specific evidence evaluation concepts (such as control of variables) in the context of rich inquiry. To our best knowledge, many interventions focus on one or a small number of evidence evaluation concepts like the law of large numbers, regression to the mean or control of variables. In contrast, there are numerous threats to validity that do not seem to garner as much attention. One reason may be that many evidence evaluation skills are taught within traditional and relatively basic physical science contexts. As discussed in the introduction, social, behavioral and health science content may be more susceptible to threats to validity, many of which are not the focus of k-12 curricula. In our brain training example studies, the potential problems were threats to construct validity, external validity and a threat to statistical validity (that of multiple posthoc statistical tests). Even an educated reader may not have the science evidence evaluation training or content knowledge to evaluate those studies.

## 8. CONCLUDING THOUGHTS

Good everyday scientific reasoning involves judging whether or not evidence is consistent with a claim or theory, or whether there are threats to validity in the evidence that render a claim invalid. Poor scientific reasoning is almost overdetermined, in that numerous factors negatively impact performance including reliance on fast and frugal heuristics, influence of prior beliefs and motivations, poor numeracy and statistical reasoning and misleading science communication. The problem is clear, but there is less consensus regarding potential solutions.

One potential solution involves bridging the disconnect between how students are taught to interpret science in k-12 and how people interpret science in reality. At least in the United States, students learn that science is objective and black and white, and there is an emphasis on scientists as authority figures (Brossard & Nisbet, 2007). Reasoning about scientific evidence is not “supposed” to involve feelings or be subject to personal beliefs or values. However, the reality is that dispositional factors have a powerful influence on how people interpret scientific evidence, specifically on the tendency to think critically about evidence. Thus science instruction as early as k-12 should not only focus on scientific inquiry skills, but also critical thinking skills in the context of science evidence evaluation. Students can be taught to challenge evidence without losing trust in science as an enterprise. They can be taught both to trust scientific authority in general and to take novel scientific findings with a grain of salt. If all students are taught to approach scientific evidence with an analytical eye, perhaps dispositional and contextual factors will eventually play a more limited role in determining whether people think critically about scientific evidence.

Given the importance of dispositional factors (not only for science evidence evaluation but for reasoning in general; Stanovich, 1999), more research should address the extent to which these dispositions are malleable and the type of experiences that might help students gain cognitive flexibility/open-mindedness, need for cognition or tendency towards analytic thinking. Some of the educational curricula and interventions discussed in our final section likely have the potential for improving dispositions more generally (such as the argument-based science activities developed by Kuhn (2010)), and indeed there is some evidence that they do at least help students understand multiple sides of arguments and avoid "myside bias", which is the tendency to focus on arguments supporting one's position rather than considering opposition arguments (Stanovich, West, & Toplak, 2013). Another proposal for improving dispositions towards thinking involves infusing such attitudes throughout the curriculum such that the dispositions are "enculturated" (Perkins, Jay, & Tishman, 1993). Perkins and Grotzer (1997) review some earlier studies whose goal is to teach thinking and reasoning more generally such as Project Intelligence (Herrnstein, Nickerson, Sanchez, & Swets, 1986) and Philosophy for Children (Lipman, 1976), both of which focus on depth of thinking; however, most of these studies did not use standard assessments of thinking dispositions. In a very preliminary study, we explored whether or not self-reported early life activities were associated with thinking dispositions. We found that engaging in arts and reading in childhood was correlated with flexible thinking, need for cognition and critical thinking (whereas game playing and athletics were not; Katz et al., 2015).

In addition to dispositional and contextual factors, knowledge about the scientific process—or the *nature of science*—also strongly predicts whether people critically evaluate scientific evidence. This includes epistemic knowledge; for instance, understanding that novel findings are tentative, as well as knowledge about the research process and the many ways in which scientific studies can be flawed (i.e., due to threats to internal or external validity). However, k-12 education does not focus on nature of science issues as much as they probably should. For example, while there is a substantial body of literature on teaching control of variables (Schwichow et al., 2016), perhaps because of its historical import in the field of developmental psychology (Inhelder & Piaget, 1958), much less research has focused directly on other specific threats to scientific validity. The Schwichow et al. (2016) paper points out some key features of successful control of variables interventions, and testing these in the contexts of threats to validity might be valuable.

By focusing on control of variables and random assignment as key factors to consider in the context of evaluating evidence, it is possible that when a study does involve random assignment, individuals might assume that the study is of high quality even if other threats to validity are present. More research on what the public values as evidence, and whether mentioning some characteristics of research (such as random assignment or meta-analysis or, more recently, "big data") leads to ignoring other possible factors (like construct validity) is of concern. A possible consequence of the "hierarchy of evidence" is too much trust in studies that may contain threats to validity. In applied contexts, RCTs and even meta-analyses are often statistically sound and avoid threats to internal validity; the source of potential concern may be construct validity or external validity.

Here, we consider two examples of studies reported in the media for which the subsequent possible policy impacts were quite high. On 13th February, 2014, The Diane Rehm Show included a discussion of a meta-analysis of mammography that found that while use of routine mammograms increased the diagnosis of breast cancer, they did not reduce death from breast cancer. As experts began discussing the study, several possible threats to validity were identified, including that the longitudinal study followed women whose mammograms (by necessity) are less advanced than those with modern technology (external validity). Attention (or not) to this problem may influence future policy decisions regarding mammogram use. In another famously controversial example, Levitt and Dubner argued in their 2005 book *Freakonomics* that there was no benefit of booster seats for older children (compared to having children in seat belts). They based their conclusion on an analysis of crash data. Many in the transportation safety research community objected to their conclusions for several reasons (Flannagan, personal communication, 2016). The study used as its sample a database that consisted only of crashes with one or more fatalities (sampling bias, a threat to statistical validity). Such accidents are unique because they tend to be very severe or have vulnerable passengers (e.g., older adults or anyone who is unbelted); as such, generalizations from that sample cannot be applied to the population of children in car crashes as a whole, nor can standard analysis methods be used. Furthermore, many other analyses of different crash datasets have found benefits of booster seats, as have studies with different methodologies such as crash dummy tests, belt fit studies and dynamic simulations (Durbin & Winston, 2005).

The potential serious consequences of the mammogram and car seat examples highlight not only the importance of evidence evaluation skills, but also that scientists have a responsibility to communicate their results with caution. Scientists, *including ourselves* and the other psychological and cognitive scientists who might read this chapter, need to reflect carefully before communicating the results of an individual study to the general public. Science is largely incremental, but there seems to be an increasing tendency towards writing short articles with little methodological detail in high-profile journals like *Psychological Science* and the *Proceedings of the National Academy of Sciences*. To be published in these forums, it seems helpful to exaggerate the novelty of one's findings, pay insufficient attention to relevant prior research and perpetuate false dichotomies (Meyer, 2016). In addition, no single study (or even meta-analysis) with a simple yes or no question fully resolves questions in the social and behavioral sciences, despite the implications of many press releases. Asking 20 questions of nature (Newell, 1973) is bound to fail.

A final note: this chapter, perhaps more so than other review articles with keener focus, takes an idiosyncratic view of the literature on science evidence evaluation. We hope that by presenting research from several different disciplines, the reader might find some valuable pointers, at least, to some novel findings.

## ACKNOWLEDGMENTS

We wish to thank the many individuals who have discussed some of these ideas with us and/or collaborated on some of the research herein. They include, in alphabetical order, Eytan Adar, Marie Burrage, Benjamin Katz, David Meyer, Xingyu Pan, Stephanie Preston, Colleen Seifert and Frank Yates. We thank Jessica Glazier for her editorial help.

## REFERENCES

Ahn, W. K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. *Explanation and Cognition*, 199–225.

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*(3), 299–352. http://dx.doi.org/10.1016/0010-0277(94)00640-7.

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development, 23*(4), 452–471. http://dx.doi.org/10.1016/j.cogdev.2008.09.002.

Anelli, C. M. (2011). Scientific literacy: What is it, are we teaching it, and does it matter? *American Entomologist, 57*(4), 235–244.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., … Unverzagt, F. W. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. *JAMA, 288*(18), 2271–2281. Chicago.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., … Wang, Y. (2009). Learning and scientific reasoning. *Science, 323*(5914), 586–587.

Baram-Tsabari, A., & Osborne, J. (2015). Bridging science education and science communication research. *Journal of Research in Science Teaching, 52*(2), 135–144. http://dx.doi.org/10.1002/tea.21202.

Baron, J. (1997). Confusion of relative and absolute risk in valuation. *Journal of Risk and Uncertainty, 14*(3), 301–309. http://dx.doi.org/10.1023/A:1007796310463.

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge: Cambridge University Press.

Beaudry, J. S., & Miller, L. (2016). *Research literacy: A primer for understanding and using research.* New York City: The Guildford Press.

Beck, D. (2010). The appeal of the brain in the popular press. *Perspectives in Psychological Science, 5*(6), 762–766. http://dx.doi.org/10.1177/1745691610388779.

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. http://dx.doi.org/10.1177/0272989X11400419.

Bishop, B., Thomas, R. K., Wood, J. A., & Gwon, M. (2010). Americans' scientific knowledge and beliefs about human evolution in the year of Darwin. *Reports of the National Center for Science Education, 30*(3), 16–18. Retrieved from http://ncse.com/rncse/30/3/americans-scientific-knowledge-beliefs-human-evolution-year.

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs. Concrete information on decisions. *Journal of Applied Social Psychology,* 7(3), 258–271. http://dx.doi.org/10.1111/j.1559-1816.1977.tb00750.x.

Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*(2), 59–69. http://dx.doi.org/10.1080/00461520.2014.921572.

Brossard, D., & Nisbet, M. C. (2007). Deference to scientific authority among a low information public: Understanding U.S. opinion on agricultural biotechnology. *International Journal of Public Opinion Research, 19*, 24–52.

Burrage, M. (2008). *"That's an interesting finding, but…": Postsecondary students' interpretations of research findings* (Doctoral dissertation).

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120. http://dx.doi.org/10.1111/1467-8624.00081.

Chua, H. F., Yates, J. F., & Shah, P. (2006). Risk avoidance: Graphs versus numbers. *Memory & Cognition, 34*(2), 399–410. http://dx.doi.org/10.3758/BF03193417.

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

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*(1), 20–33.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized control trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*, 1887–1892. http://dx.doi.org/10.1056/NEJM200006223422507.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Crowell, A., & Schunn, C. (2016). Unpacking the relationship between science education and applied scientific literacy. *Research in Science Education, 46*(1), 129–140.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin, 137*(5), 800–818. http://dx.doi.org/10.1037/a0021860.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*(3), 1248–1299. http://dx.doi.org/10.1016/j.cognition.2007.06.002.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*(4), 568. http://dx.doi.org/10.1037/0022-3514.63.4.568.

Domhoff, G. W., & Schneider, A. (1999). Much ado about very little: The small effect sizes when home and laboratory collected dreams are compared. *Dreaming, 9*(2–3), 139–151. Retrieved from http://psycnet.apa.org/doi/10.1023/A:1021389615347.

Dose of nature is just what the doctor ordered. *Science Daily*, (June 23, 2016). Retrieved from https://www.sciencedaily.com/releases/2016/06/160623095252.htm.

Durante, M. (2015). *Everyday scientific reasoning: Critical approaches outside the classroom* (Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Arts with Honors in Psychology (or BCN) from the University of Michigan 2015).

Durbin, D., & Winston, F. (July 24, 2005). *The seat-belt solution*. The New York Times. Retrieved from http://www.nytimes.com/2005/07/24/opinion/magazine/the-seatbelt-solution-507326.html?_r=0.

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2008). Constructing causal diagrams to learn deliberation. *International Journal of Artificial Intelligence in Education, 18*(4), 1–3.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, 113*(28), 7900–7905. http://dx.doi.org/10.1073/pnas.1602413113.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology, 71*(2), 390. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-3514.71.2.390.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making, 7*(6), 746–749.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing, 12*(1), 77–84. http://dx.doi.org/10.1046/j.1365-2702.2003.00662.x.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science,* 7(10), 454–469. http://dx.doi.org/10.1016/j.tics.2003.08.012.

Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382–389. http://dx.doi.org/10.1080/13546780542000005.

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25*(4), 398–405. http://dx.doi.org/10.1177/0272989X05278931.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672–680. http://dx.doi.org/10.1177/0272989X07304449.

Farah, M. J., & Hook, C. J. (2013). The seductive allure of "seductive allure". *Perspectives on Psychological Science, 8*(1), 88–90. http://dx.doi.org/10.1177/1745691612469035.

Fernandez-Duque, D., Evans, J., Colton, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience, 27*(5), 926–944. http://dx.doi.org/10.1162/jocn_a_00750.

Flannagan, C. (August 1, 2016). Personal communication.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*(3), 253–292. http://dx.doi.org/10.1016/0010-0285(86)90001-0.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*(1), 34. http://dx.doi.org/10.1037/0096-3445.120.1.34.

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

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy a cross-cultural comparison. *Medical Decision Making, 31*(3), 444–457. http://dx.doi.org/10.1177/0272989X10373805.

Gallup. (March 2–6, 2016). U.S. concern about global warming at eight-year high. *Gallup*. Retrieved from http://www.gallup.com/poll/190010/concern-global-warming-eight-year-high.aspx.

Gao, T., Seifert, C., Shah, P., & Adar, E. (2016). *Animation affects causality interpretation of scatterplots* (in preparation).

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38.

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology, 52*(3), 464–474. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-3514.52.3.464.

Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology, 75*, 105–118. http://dx.doi.org/10.1348/000709904X22278.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455. Retrieved from http://psycnet.apa.org/doi/10.1037/0003-066X.53.4.449.

Hatfield, J., Faunce, G. J., & Job, R. S. (2006). Avoiding confusion surrounding the phrase "correlation does not imply causation". *Teaching of Psychology, 33*(1), 49–51.

Herrnstein, R. J., Nickerson, R. S., Sanchez, M., & Swets, J. A. (1986). Teaching thinking skills. *American Psychologist, 41*(11), 1279–1289. http://dx.doi.org/10.1037/0003-066X.41.11.1279.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88–140. http://dx.doi.org/10.3102/00346543067001088.

Hook, C. J., & Farah, M. J. (2013). Look again: Effects of brain images and mind-brain dualism on lay evaluations of research. *Journal of Cognitive Neuroscience, 25*(9), 1397–1405. http://dx.doi.org/10.1162/jocn_a_00407.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition, 155*, 67–76. http://dx.doi.org/10.1016/j.cognition.2016.06.011.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences, 112*(33), 10321–10324. http://dx.doi.org/10.1073/pnas.1504019112.

Huber, C. R., & Kuncel, N. R. (2015). Does college teach critical thinking? A meta-analysis. *Review of Educational Research, 1987*, 1–38. http://dx.doi.org/10.3102/0034654315605917.

Hurley, D. (July 24, 2016). *Could brain training prevent dementia?* New Yorker. Retrieved from http://www.newyorker.com/tech/elements/could-brain-training-prevent-dementia retrieved July 28, 2016.

Ibrahim, A., Seifert, C., Adar, E., & Shah, P. (2016). *Using graphs to debias misinformation* (in preparation).

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–646. http://dx.doi.org/10.1097/EDE.0b013e31818131e7.

Jewett, E., & Kuhn, D. (2016). Social science as a tool in developing scientific thinking skills in underserved, low-achieving urban students. *Journal of Experimental Child Psychology, 143*, 154–161. http://dx.doi.org/10.1016/j.jecp.2015.10.019.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420. http://dx.doi.org/10.1037/0278-7393.20.6.1420.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013). *Motivated numeracy and enlightened self-government* (Vol. 307). Yale Law School, Public Law Working Paper.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732—735. http://dx.doi.org/10.1038/nclimate1547.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430—454. http://dx.doi.org/10.1016/0010-0285(72)90016-3.

Katz, B., Park, I., Jantz, T., Brink, K., Buchanan, A., & Shah, P. (April 2015). *Specific childhood leisure activities predict cognitive abilities and dispositions towards thinking.* Philadelphia, PA: Poster presented at the Biannual Meeting of the Society for Research in Child Development.

Katz, B., & Shah, P. (2016a). Logical and methodological considerations in cognitive training research. In M. Bunting, J. Novick, M. Dougherty, & R. Engle (Eds.), *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development.* Oxford University Press (in press).

Katz, B., & Shah, P. (2016b). The jury is still out on working memory training. *JAMA Pediatrics*.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults.* San Francisco, CA: Jossey-Bass Higher and Adult Education Series and Jossey-Bass Social and Behavioral Science Series.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71*(5), 1347—1366. http://dx.doi.org/10.1111/1467-8624.00232.

Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1997). Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology, 89*(3), 470. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-0663.89.3.470.

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

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661—667. http://dx.doi.org/10.1111/j.0956-7976.2004.00737.x.

Kolstø, S. D., Bungum, B., Arnesen, E., Isnes, A., Kristensen, T., Mathiassen, K., … Ulvik, M. (2006). Science students' critical examination of scientific information related to socioscientific issues. *Science Education, 90*(4), 632—655. http://dx.doi.org/10.1002/sce.20133.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development, 23*(4), 472—487. http://dx.doi.org/10.1016/j.cogdev.2008.09.007.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60*(6), 1316—1327. http://dx.doi.org/10.2307/1130923.

Kosonen, P., & Winne, P. H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Educational Psychology, 87*(1), 33. http://dx.doi.org/10.1037/0022-0663.87.1.33.

Kreuter, M. W., Holmes, K., Alcaraz, K., Kalesan, B., Rath, S., Richert, M., … Clark, E. M. (2010). Comparing narrative and informational videos to increase mammography in low-income African American women. *Patient Education and Counseling, 81*(1), S6–S14. http://dx.doi.org/10.1016/j.pec.2010.09.008.

Kuhn, S. (2001). How do people know? *Psychological Science, 12*(1), 1–8. http://dx.doi.org/10.1111/1467-9280.00302.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94*(5), 810–824. http://dx.doi.org/10.1002/sce.20395.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866–870. http://dx.doi.org/10.1111/j.1467-9280.2005.01628.x.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the society for research in child development, 60*(4), i-157. http://dx.doi.org/10.2307/1166059.

Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking & Reasoning, 13*(2), 90–104. http://dx.doi.org/10.1080/13546780600625447.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480. http://dx.doi.org/10.1037/0033-2909.108.3.480.

Langer, G. (June 19, 2015). Poll: Skepticism of genetically modified foods. *ABC News.*

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11–24. http://dx.doi.org/10.1037/0033-2909.108.3.480.

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195–216. http://dx.doi.org/10.1007/s10758-007-9122-2.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*(1), 69–108. http://dx.doi.org/10.1207/s1532690xci1401_3.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635–679. http://dx.doi.org/10.3102/00028312041003635.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In W. Damon, & R. Lerner (Series Eds.) & K. A. Renninger, & I. E. Sigel (Vol. Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed.). New York: John Wiley and Sons. http://dx.doi.org/10.1002/9780470147658.chpsy0405.

Levitt, S., & Dubner, S. (2005). *Freakonomics: A rogue environment explores the hidden side of everything*. New York, NY: William Morrow.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. http://dx.doi.org/10.1177/1529100612451018.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*(2), 87–137. http://dx.doi.org/10.1006/cogp.1999.0724.

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. http://dx.doi.org/10.1177/0272989X0102100105.

Lipman, M. (1976). Philosophy for children. *Metaphilosophy,* 7(3/4), 17–39. http://dx.doi.org/10.5840/thinking1982339.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470. http://dx.doi.org/10.1016/j.tics.2006.08.004.

Lorch, R. F., Jr., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology, 106*(1), 18. http://dx.doi.org/10.1037/a0034375.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109. http://dx.doi.org/10.1037/0022-3514.37.11.2098.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., & Soloway, E. (1997). Enacting project-based science. *The Elementary School Journal, 97*(4), 341–358.

Mason, L. (2000). Role of anomalous data and epistemological beliefs in middle school students' theory change about two controversial topics. *European Journal of Psychology of Education, 15*(3), 329–346. http://dx.doi.org/10.1007/BF03173183.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*, 343–352. http://dx.doi.org/10.1016/j.cognition.2007.07.017.

Meyer, D. E. (2016). The essential Dave Meyer: Some musings on scholarly 'eminence' and important scientific contributions. In R. Sternberg, S. Fiske, & D. Foss (Eds.), *Scientists making a difference: The greatest living behavioral and brain scientists talk about their most important contributions* (pp. 93–98). New York: Cambridge University Press.

Miller, J. D. (1996). *Scientific literacy for effective citizenship*. Science/Technology/Society as Reform in Science Education.

Miller, J. D., Scott, E. C., & Okamoto, S. (2006). Science communication: Public acceptance of evolution. *Science, 313*, 765–766.

Minahan, J., & Siedlecki, K. L. (2016). Individual differences in Need for Cognition influence the evaluation of circular scientific explanations. *Personality and Individual Differences, 99*, 113–117. http://dx.doi.org/10.1016/j.paid.2016.04.074.

Molek-Kozakowska, K. (December 2014). Hybrid styles in popular reporting on science: A study of new scientist's headlines. In *Electronic proceedings* (p. 135).

Moore, D. W. (June 16, 2005). *Three in four Americans believe in paranormal*. Gallup Poll News Service. Retrieved from http://www.gallup.com/poll/16915/three-four- americans-believe-paranormal.aspx.

Munro, G. D. (2010). The scientific impotence excuse: Discounting belief-threatening scientific abstracts. *Journal of Applied Social Psychology, 40*(3), 579–600. http://dx.doi.org/10.1111/j.1559-1816.2010.00588.x.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289. http://dx.doi.org/10.1037/0033-295X.92.3.289.

Newell, A. (1973). In *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium*.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. http://dx.doi.org/10.1037/1089-2680.2.2.175.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238*(4827), 625–631. http://dx.doi.org/10.1126/science.3672116.

Norcross, J. C., Gerrity, D. M., & Hogan, E. M. (1993). Some outcomes and lessons from a cross-sectional evaluation of psychology undergraduates. *Teaching of Psychology, 20*(2), 93–96. http://dx.doi.org/10.1207/s15328023top2002_6.

Norris, S. P., & Phillips, L. M. (1994). Interpreting pragmatic meaning when reading popular reports of science. *Journal of Research in Science Teaching, 31*(9), 947–967. http://dx.doi.org/10.1002/tea.3660310909.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*(2), 123–145.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330. http://dx.doi.org/10.1007/s11109-010-9112-2.

Oestermeier, U., & Hesse, F. W. (2000). Verbal and visual causal arguments. *Cognition, 75*(1), 65–104. http://dx.doi.org/10.1016/S0010-0277(00)00060-3.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Pan, X., & Shah, P. (2016, in preparation). *Number absolutism.*

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science,* 7(6), 531–536. http://dx.doi.org/10.1177/1745691612463401.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*(3), 335–346. http://dx.doi.org/10.1016/j.cognition.2012.03.003.

Perkins, D. N., & Grotzer, T. A. (1997). Teaching intelligence. *American Psychologist, 52*(10), 1125–1133. http://dx.doi.org/10.1037/0003-066X.52.10.1125.

Perkins, D. N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly, 39*(1), 1–21.

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science, 21*(1), 31–35. http://dx.doi.org/10.1177/0963721411429960.

Peters, J. D. (2012). *Speaking into the air: A history of the idea of communication*. University of Chicago Press.

Picardi, C. A., & Masick, K. D. (2013). *Research methods: Designing and conducting research with a real-world focus*. SAGE Publications.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Prasad, M., Perrin, A. J., Bezila, K., Hoffman, S. G., Kindleberger, K., Manturuk, K., & Powers, A. S. (2009). "There must be a reason": Osama, Saddam, and inferred justification. *Sociological Inquiry, 79*(2), 142–162. http://dx.doi.org/10.1111/j.1475-682X.2009.00280.x.

Pressley, M., & Wharton-McDonald, R. (1997). Skilled comprehension and its development through instruction. *School Psychology Review, 26*(3), 448–566.

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

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

Reynolds, G. (July 20, 2016). *Lifting lighter weights can be just as beneficial as lifting heavier ones.* The New York Times. Retrieved from http://well.blogs.nytimes.com/2016/07/20/lifting-lighter-weights-can-be-just-as-effective-as-heavy-ones/.

Rhodes, R. E. (2015). *The influence of reductionist information on perceptions of scientific validity* (Doctoral dissertation). Ann Arbor, MI: University of Michigan.

Rhodes, R. E., Rodriguez, F., & Shah, P. (2014). Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1432–1440. http://dx.doi.org/10.1037/a0036844.

Rhodes, R. E., & Shah, P. (2016a). *Seeing behavior through the brain: Evidence of neurorealism* (Under review).

Rhodes, R. E., & Shah, P. (2016b). *Preference for micro-level science* (Under review).

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., … Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics, 170*(5), e154568. http://dx.doi.org/10.1001/jamapediatrics.2015.4568.

Rodriguez, F., Ng, A., & Shah, P. (2016). Do college students notice errors in evidence when critically evaluating research findings? *Journal on Excellence in College Teaching, 27*(3), 63–78.

Rodriguez, F., Rhodes, R. E., Miller, K., & Shah, P. (2016). Examining the influence of anecdotal stories and the interplay of individual differences on reasoning. *Thinking & Reasoning, 22*(3), 274–296. http://dx.doi.org/10.1080/13546783.2016.1139506.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562.

Ruiz, R. (June 26, 2016). *Why scientists think your social media posts can help prevent suicide*. Mashable. Retrieved from http://mashable.com/2016/06/26/suicide-prevention-social-media/#yNz82Hv_Daqg.

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*(3), 497–510. http://dx.doi.org/10.1037/0022-0663.91.3.497.

Sagan, C. (1996a). Does truth matter? Science, pseudoscience, and civilization. *Skeptical Inquirer, 20*, 28–33.

Sagan, C. (1996b). *The demon-haunted world*. New York: Ballantine Books.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*(4), 634–656. http://dx.doi.org/10.1002/sce.20065.

Sanitioso, R., & Kunda, Z. (1991). Ducking the collection of costly evidence: Motivated use of statistical heuristics. *Journal of Behavioral Decision Making, 4*(3), 161–176. http://dx.doi.org/10.1002/bdm.3960040302.

Schaller, M. (1992). In-group favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes. *Journal of Personality and Social Psychology, 63*(1), 61. http://dx.doi.org/10.1037/0022-3514.63.1.61.

Schaller, M., & O'Brien, M. (1992). "Intuitive analysis of covariance" and group stereotype formation. *Personality & Social Psychology Bulletin, 18*(6), 776–785. http://dx.doi.org/10.1177/0146167292186014.

Scharrer, L., Bromme, R., Britt, M. A., & Stadtler, M. (2012). The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learning and Instruction, 22*, 231–243. http://dx.doi.org/10.1016/j.learninstruc.2011.11.004.

Schoenfeld, A. H. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education, 10*(3), 173–187. http://dx.doi.org/10.2307/748805.

Schraw, G., Bendixen, L. D., Dunkle, M. E., Hofer, B. K., & Pintrich, P. R. (2002). Development and validation of the epistemic belief inventory (EBI). In *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 261–275). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology, 29*(1), 42–62. http://dx.doi.org/10.1006/jesp.1993.1003.

Schul, Y., Mayo, R., & Burnstein, E. (2008). The value of distrust. *Journal of Experimental Social Psychology, 44*(5), 1293–1302. http://dx.doi.org/10.1016/j.jesp.2008.05.003.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), 337–370. http://dx.doi.org/10.1016/S0364-0213(99)00006-3.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, G. H. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*, 966–971.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., … Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632—654. http://dx.doi.org/10.1002/tea.20311.

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science and Policy, 2*, 85—95.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology, 39*, 127—161. http://dx.doi.org/10.1016/S0065-2601(06)39003-X.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37—63. http://dx.doi.org/10.1016/j.dr.2015.12.001.

Sedlmeier, P. (2002). Associative learning and frequency judgments: The PASS model. *Etc.: Frequency processing and cognition*, 137—152.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*(1), 33—51.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *Psychology of Learning and Motivation, 41*, 265—292. http://dx.doi.org/10.1016/S0079-7421(02)80009-3.

Shaklee, H., & Elek, S. (1988). Cause and covariate: Development of two related concepts. *Cognitive Development, 3*(1), 1—13. http://dx.doi.org/10.1016/0885-2014(88)90027-5.

Shi, J., Visschers, V. H., Siegrist, M., & Arvai, J. (2016). Knowledge as a driver of public perceptions about climate change reassessed. *Nature Climate Change, 6*, 759—762. http://dx.doi.org/10.1038/nclimate2997.

Shiffrin, R. M. (2016). Drawing causal inference from big data. *Proceedings of the National Academy of Sciences, 113*(27), 7308—7309. http://dx.doi.org/10.1073/pnas.1608845113.

Shiloh, S., Salton, E., & Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences, 32*(3), 415—429. http://dx.doi.org/10.1016/S0191-8869(01)00034-4.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359—1366. http://dx.doi.org/10.1177/0956797611417632.

Sinatra, G. M., Kienhues, D., & Hofer, B. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist, 49*(2), 123—138. http://dx.doi.org/10.1080/00461520.2014.916216.

Slater, M. D., & Rouner, D. (1996). How message evaluation and source attributes may influence credibility assessment and belief change. *Journalism & Mass Communication Quarterly, 73*(4), 974—991. http://dx.doi.org/10.1177/107769909607300415.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*(1), 1—21. http://dx.doi.org/10.1016/0010-0277(94)90002-7.

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cognition and Instruction, 31*(2), 130—150. http://dx.doi.org/10.1080/07370008.2013.769996.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

Stanovich, K. E., & Stanovich, P. J. (2010). A framework for critical thinking, rational thinking, and intelligence. In *Innovations in educational psychology: Perspectives on learning, teaching and human development* (pp. 195—237).

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. http://dx.doi.org/10.1037/0022-0663.89.2.342.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22*(4), 259—264. http://dx.doi.org/10.1177/0963721413480174.

Stone, E. R., Sieck, W. R., Bull, B. E., Yates, J. F., Parks, S. C., & Rush, C. J. (2003). Foreground: Background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes, 90*(1), 19—36. http://dx.doi.org/10.1016/S0749-5978(03)00003-7.

Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied, 3*(4), 243—256. http://dx.doi.org/10.1037/1076-898X.3.4.243.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., … Boy, F. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ: British Medical Journal, 349*, 1—8. http://dx.doi.org/10.1136/bmj.g7015.

Tal, A., & Wansink, B. (2016). Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science, 25*(1), 117—125. http://dx.doi.org/10.1177/0963662514549688.

Thompson, V. L. S. (2013). Making decisions in a complex information environment: Evidential preference and information we trust. *BMC Medical Informatics and Decision Making, 13*(3), 1. http://dx.doi.org/10.1186/1472-6947-13-S3-S7.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275—1289. http://dx.doi.org/10.3758/s13421-011-0104-1.

Trefil, J. (2008). Science education for everyone: Why and what? *Liberal Education, 94*(2), 6—11.

Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decision aids effects on treatment choices. *Medical Decision Making, 21*(1), 60—68. http://dx.doi.org/10.1177/0272989X0102100108.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, 1—11. http://dx.doi.org/10.1093/ije/dyv341.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470—477. http://dx.doi.org/10.1162/jocn.2008.20040.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education, 92*(5), 941—967. http://dx.doi.org/10.1002/sce.20259.