Journal of Economic Methodology, 2014
Vol. 21, No. 3, 211–231, http://dx.doi.org/10.1080/1350178X.2014.939691

The power of stereotyping and confirmation bias to overwhelm accurate assessment: the case of economics, gender, and risk aversion

Julie A. Nelson*

Department of Economics, University of Massachusetts Boston, 100 Morrissey Blvd, Boston, MA 02125, USA

(*Received 30 May 2013; accepted 7 January 2014*)

Behavioral research has revealed how normal human cognitive processes can tend to lead us astray. But do these affect economic researchers, ourselves? This article explores the consequences of stereotyping and confirmation bias using a sample of published articles from the economics literature on gender and risk aversion. The results demonstrate that the supposedly 'robust' claim that 'women are more risk averse than men' is far less empirically supported than has been claimed. The questions of how these cognitive biases arise and why they have such power are discussed, and methodological practices that may help to attenuate these biases are outlined.

Keywords: stereotyping; confirmation bias; gender; risk aversion; effect size; index of similarity

Introduction

> The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects; in order that by this great and pernicious predetermination the authority of its former conclusions may remain inviolate . . .
>
> Francis Bacon in 1620, quoted in Nickerson (1998, p. 176)

Economists who aspire to do reliable, objective research want our work to be as free as possible from elements of personal and cultural subjectivity and bias. At the same time, however, researchers in the cognitive sciences have demonstrated that human cognition often tends to systematically deviate from norms of context-free impartiality and logic (Kahneman, 2003). The field of Behavioral Economics has recently introduced analysis of these into the study of economic decision-making (Camerer et al., 2003). But to what extent do these biases apply to economic researchers, ourselves? The present study looks in particular at the question of whether economists seem to be prone to stereotyping (the tendency to draw on overly simple beliefs about groups to make judgments about individuals) and confirmation bias (the tendency to perceive and seek out information that confirms one's pre-existing beliefs, and avoid information that conflicts).

The claim that 'women are more risk averse than men,' for example, has become widespread in the economics literature (for example, Arano, Parker, & Terry, 2010;

*Email: julie.nelson@umb.edu

Bernasek & Shwiff, 2001, Booth & Nolen, 2012, Borghans, Golsteyn, Heckman, & Meijers, 2009 – and nearly every other article on risk reviewed below). Results from empirical studies showing statistically significant differences between men's and women's behavior, on average, in experimental lotteries or retirement investments, or in responses to survey questions, are given as evidence. Reviewing a swath of this literature in the high-profile *Journal of Economic Literature*, Croson and Gneezy (2009) conclude that there is a 'fundamental' (p. 467) difference between men and women in risk aversion (p. 448).

Such a 'finding' confirms popularly held stereotypes of men as the braver and more adventurous sex. But could such a finding also be, at least in part, a *result* of such stereotypes affecting economics research? Could researchers be tending to 'find' results that confirm socially held prior beliefs? This article examines the relationship between the empirical evidence and the claims made in a sample of economic studies of risk aversion, using an expanded set of quantitative and qualitative tools.

The argument is as follows: suppose that the average value of some variable derived from data about women is found to be substantially below that for men, and, in addition, within-sex variability is found to be minimal. The men's and women's distributions would then have little-to-no overlap; drawing conclusions about individuals based on the group averages would be justifiable; and – were this to hold for all human males and females – one might even say that the difference in means reflects a truly fundamental or 'essential' sex difference.[1] Drawing on my (Nelson 2014) empirical review of the literature on gender and risk aversion, however, this study first shows that substantive differences actually found in studies published in economics journals on gender and risk aversion are small, and the degree of overlap between women's and men's distributions is considerable. This article then demonstrates that, in spite of such evidence, many works still make claims about 'essential' differences. In addition (1) earlier literatures are inaccurately cited in a stereotype-confirming way, (2) results that confirm the stereotype are emphasized, while results that do not are downplayed, (3) stereotype-confirming results are more likely to be published, (4) the effect of confounding variables is neglected, and (5) the areas of risk studied are selectively chosen. With the claims made in the literature about differences in risk aversion having been shown to be exaggerated and over-generalized far beyond what the data actually support, this article then draws on a larger scholarly literature to explore the sources and persistence of stereotyping and confirmation bias. Methodological innovations that could help to reduce these biases are also discussed.

The case of gender and risk aversion is of more than simply methodological importance, however, since the perception that there are substantial sex differences in risk preference has become part of public and academic discussions about financial market stability (e.g., Kristof, 2009); labor market, business, and investment success (e.g., Booth & Nolen, 2012, p. F56; Eckel & Grossman, 2008, p. 14); and environmental policy (e.g., Kahan, Braman, Gastil, Slovic, & Mertz, 2007). To the extent that illegitimate gender stereotyping at a cultural level is reinforced – rather than weakened – by social science research, negative repercussions can occur in many realms of real-world experience.

The relation of claims to evidence

Before turning to a re-examination of economics studies on gender and risk, a discussion of the relation of claims to evidence is necessary. Additional data-analysis tools must also be introduced.

Empirical versus essentialist statements

Consider the following two statements:

A. In our sample, we found a statistically significant difference in mean risk aversion between men and women.

B. Women are more risk averse than men.

While the two statements are often taken as meaning the same thing, they in fact convey very different meanings. Statement A is a narrow statement, referring to an aggregate (that is, group-level) finding, which can be factually correct within the confines of a particular study. It is empirical. Statement B, on the other hand, will be widely interpreted as implying something about stable characteristics of individual people according to their presumed male or female natures or essences. Statement B tends to create an expectation that, in a comparison between an individual man and individual woman, the woman will be found to be, by virtue of her womanly nature, more risk-shy.[2]

The essentialist meaning that will generally be drawn from Statement B may, of course, not have been intended by a researcher who states it. It is likely that some (perhaps many) researchers who make Statement B or report finding a 'sex difference' may primarily mean simply that they found an empirical difference on average (Statement A), and are merely writing in shorthand (Statement B). Certainly, such phrasing can also be found in academic studies regarding groups classified by race, nationality, income, and so on – and may create similar tendencies toward mis-interpretation and stereotyping.

But suppose one *begins* with a belief in essential, fundamental sex differences. Then it will seem obvious and natural to assume that Statement B reflects an underlying reality, of which Statement A is a consequence. Furthermore, it will seem obvious that one can take Statement A as evidence that confirms Statement B. Furthermore, since one is already assuming that men and women are in distinct categories, investigating the actual substantive *size* of a difference may seem irrelevant. Lastly, if one happens to come upon evidence that conflicts in some way with Statement B, one may tend to overlook it, or find fault with some aspect of the evidence rather than with one's belief.

These are exactly the sort of fallacies ('hypothesis-determined information seeking and interpretation,' Nickerson, 1998, p. 177) which constitute confirmation bias. And, unfortunately, evidence of all of these fallacies can be found in the economics literature on gender and risk aversion.

A tool for measuring substantive difference

Clearly it is more legitimate to think in terms of essential differences the more that, *within* a group, group members are identical or highly similar to each other, and, *across* groups, people are substantively very different from each other. Conversely, the more within-class heterogeneity there is, and the less substantive difference and more overlap there is between the group distributions, the less support there is for a clear-cut categorization.

Economists have, however, neglected these issues, tending to describe in categorical language cases in which, simply, the point estimate of the mean (or median) of a measure of risk taking for men exceeds that for women. Statistical significance is also generally checked – though not always, as will be seen below.[3] Discussions of the substantive size of the gender differences found are rare.[4] Discussions of the degree of overlap of men's and women's distributions are virtually non-existent.

Fortunately, measures of the ‘substantive significance’ of an observed difference (a very different thing from ‘statisical significance,’ as pointed out by Ziliak and McCloskey (2004) and Miller and van der Meulen Rodgers (2008)) have been much discussed in psychology, education, and other fields. There, the common use of ‘effect size’ measures of the substantive size of a difference helps address the issues of within-group variation and across-group overlap.

Cohen’s *d*, a very commonly used measure of effect size, expresses the magnitude of a difference between means in standard deviation units (e.g., Byrnes, Miller, & Schafer, 1999; Cross, Copping, & Campbell, 2011; Hyde, 2005; Wilkinson & Task Force on Statistical Inference, 1999). For the case of a male versus female comparison, it is conventionally calculated as

$$d = \frac{\bar{X}_{\mathrm{m}} - \bar{X}_{\mathrm{f}}}{s_{\mathrm{p}}}, \tag{1}$$

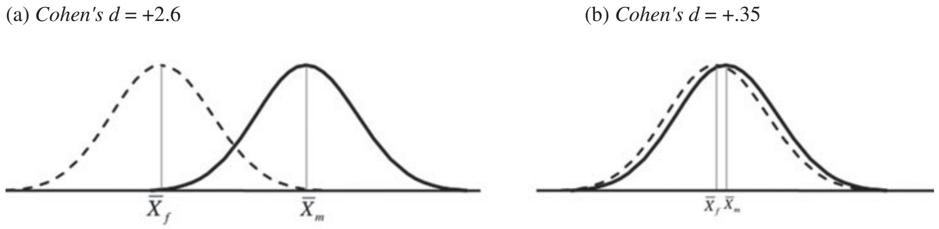
where $\bar{X}_{\mathrm{m}}$ is the male mean, $\bar{X}_{\mathrm{f}}$ is the female mean, and s_{p} is the pooled standard deviation, a measure of the average within-group variation.[5]

To visualize what this represents, Figure 1 illustrates, using normal distributions, Cohen’s *d* values of + 2.6 and + 0.35. A Cohen’s *d* value of + 2.6 corresponds to the commonly observed difference in male and female heights (Eliot, 2009). While the male and female distributions are quite distinct, they still overlap throughout a substantial range. Figure 1(b), in contrast, illustrates $d = 0.35$, corresponding to a level used by psychologist Hyde (2005) in a meta-analysis of 124 sex-related effect sizes.[6] Clearly, in such a case, ‘difference’ is considerably less substantial. A truly categorical difference – for which one could very reliably conclude that ‘men are more *X* than women’ even on the basis of random pairwise comparisons, then, would require a Cohen’s *d* that is even larger than that for height.

In the psychological literature on behavior, the expression of findings in terms of substantive difference is widespread, standardized, considered best practice (Wilkinson & Task Force on Statistical Inference, 1999). Unlike in economics, the implications of various methodologies for the study of sex differences have also, in this literature, been the topic of intense professional discussion (Archer, 1996; Eagly, 1995; Hyde & Plant, 1995; Martell, Lane, & Emrich, 1996, and other articles in the March 1995 and February 1996 issues of the American Psychologist).

A tool for measuring similarity

The index of similarity (IS) introduced by Nelson (2014) is an easily computable and understandable measure of the degree of *overlap* between two distributions. It is



(a) *Cohen's d* = +2.6

(b) *Cohen's d* = +.35

Figure 1. Illustration of Cohen’s *d*.

calculated as

$$\mathrm{IS} = 1 - \frac{1}{2}\left(\sum_i \left|\frac{f_i}{F} - \frac{m_i}{M}\right|\right), \tag{2}$$

where f_i/F is the proportion of females within category i and m_i/M is the proportion of males in that same category. IS has an intuitive interpretation as (in equal-sized groups) the proportion of the females and males who are similar, in the sense that their characteristics or behaviors (on this particular front) *exactly match up* with someone in the opposite sex group. IS is derived from the 'index of dissimilarity' (also called 'Duncan's *D*') that has been long used to study racial housing segregation (Duncan & Duncan, 1955), which is also the formula that underlies the 'index of occupational segregation' used to study gender segregation of occupations (Blau, Ferber, & Winkler, 2010, p. 135; Reskin, 1993). Using this alongside Cohen's *d* creates a sort of symmetry: one technique measures difference, while the other measures similarity.

Re-examining the economics literature on gender and risk

Using summary statistics from the articles and/or re-analysis of data provided by authors, Nelson (2014) calculates Cohen's *d* and IS values for 35 studies of gender and risk preferences or risk perceptions from the literatures in economics, finance, and psychology. The types of studies done vary from experimental studies in which subjects are offered lotteries of various types, to analysis of survey questions asking people about their attitudes toward various risks (including financial and/or employment risks), to studies of financial asset allocations among risky or less-risky assets. Most of the sample sizes are large, with the number of respondents usually in the 100s or 1000s. The subjects in nearly all the studies are from Western industrialized countries.

This study begins with a subset of the articles reviewed in Nelson (2014), focusing only on articles published in journals that mention economics in their self-descriptions, which were reviewed prior to 2013, and which specifically study risk aversion (rather than risk perception). First, Cohen's *d* and IS results are reproduced here, in order to illustrate the extent to which the underlying data structure might be consistent with a categorical and individual-level model of 'difference.' Following that, the texts of the articles are analyzed to determine the extent to which the claims made are actually consistent with the data.

Actual magnitudes of sex difference and similarity

Table 1 summarizes the findings for men versus women for the 18 of the publications surveyed in Nelson (2014) that are economics related and for which data were available, looking only at the questions in each study which specifically focus on risk aversion.

Cohen's *d*, expressed such that a positive number signifies greater mean male risk taking is reported for only for differences between means that were statistically significant (at a 10% level or better).[7] Note that in two articles differences that are statistically significant in the direction of greater female risk taking ($d < 0$) are among the findings, and many more articles include some variables or samples for which no statistically significant difference is found, in spite of generally large sample sizes. A finding of a *d* value exceeding $+0.50$ – that is, half standard deviation, in favor of greater male risk taking – occurs in only 4 of 18 articles, and the finding of a difference of more than 1

Table 1. Magnitudes of male vs. female differences and similarities related to risk.

Author(s)	Cohen's *d*	Index of similarity
Harris, Jenkins, and Glaser (2006)	− 0.34 to NSS to 0.74	–
Fehr-Duda, De Gennaro, and Schubert (2006)	− 0.25 to NSS to 0.49	–
Arano et al. (2010)	NSS	–
Gneezy and Leonard (2009)	NSS	–
Bernasek and Shwiff (2001)	NSS	0.87
Lindquist and Säve-Söderbergh (2011)	NSS	–
Holt and Laury (2002)	NSS to 0.37	0.83 to 0.86
Booth and Nolen (2012)	NSS to 0.38	0.84
Beckmann and Menkhoff (2008)	NSS to 0.46	0.67 to 0.91
Dohmen et al. (2011)	NSS to 0.48	0.80 to 0.88
Meier-Pesti and Penz (2008)	NSS to 0.85	–
Powell and Ansic (1997)	0.06 to 0.17	0.90 to 0.93
Sunden and Surette (1998)	0.08 to 0.16	0.95 to 0.96
Barber and Odean (2001)	0.09 to 0.26	–
Eriksson and Simpson (2010)	0.19 to 0.22	0.89 to 0.91
Hartog, Ferrer-i-Carbonell, and Jonker (2002)	0.22 to 0.29	0.85 to 0.96
Borghans et al. (2009)	0.32 to 0.55	–
Eckel and Grossman (2008)	0.55 to 1.13	0.60 to 0.80

Note: Adjusted so that positive *d* values indicate relatively greater risk taking on the part of males, compared to females, on average. NSS indicates no statistically significant difference (see text for further explanation).
Source: Nelson (2014).

standard deviation of difference occurs in only 1 article. In most cases – and even within the same articles reporting the largest *d* values – smaller *d* values are (also) found.

Table 1 also reports Indexes of Similarity for comparisons reported as statistically significant in the source articles. Since these figures measure *similarity* but are only reported here for *statistically significant differences*, the numbers in Table 1 represent the low end of possible IS values that could be found in these data. IS values range from 0.60 to 0.96, with most studies yielding no values below 0.80.[8]

In light of the results in Table 1, the existence of a categorical male/female difference in risk-taking by sex can clearly be ruled out. Statistically significant differences in group averages are sometimes not found, or go in the 'wrong' direction. More importantly, it is easily seen that even when a statistically significant *difference* in *means* is found, the degree of *overlap* among *individuals* in each group is considerable. Instead of difference, similarity seems to be the more prominent pattern, with well over half of men and women 'matching up' on risk-related behaviors in every study.

Essentialist assertions

Are the results of these studies portrayed as representing simple empirical differences, on average, as would be merited by the data? Or are they presented as confirming a presumably essential sex difference?

Among economics articles on gender and risk, one can find titles such as 'Will Women Be Women?' (Beckmann & Menkhoff, 2008) in *Kyklos* and 'Girls will be Girls' in *Economics Letters* (Lindquist & Säve-Söderbergh, 2011). The apparent presumption in such titles is that were a group of women or girls found to *not* be relatively more risk averse, they would somehow be abnormal relative to their own female natures. Other articles, while not so clear in their titles, treat risk-aversion explicitly as a sex-linked 'trait' (e.g., Borghans et al., 2009; Powell & Ansic, 1997), presumably stable across time and cultural contexts.

In addition, many studies hypothesize evolutionary explanations for female risk aversion or male risk-seeking (e.g., Cross et al., 2011; Hartog et al., 2002; Olsen & Cox, 2001), or hypothesize links to sex-related hormones or other genetic factors commonly thought to define an essence of femaleness or maleness (see examples cited in Meier-Pesti and Penz (2008) and Croson and Gneezy (2009)). Croson and Gneezy (2009) claim that the literature has 'documented fundamental differences between men and women' (p. 467).

Other researchers take observed differences in means, and from these make sweeping recommendations consistent with an essentialist view. Two articles reviewed for this study, for example, recommend that women investors should be paired with women investment advisers (Beckmann & Menkhoff, 2008, p. 381; Olsen & Cox, 2001). Because the within-sex patterns actually observed in these studies, however, are quite heterogeneous and the male and female distributions have considerable overlap, pairing by sex would in fact add little to the likelihood of congruence in risk preferences. Consider one of the relatively sizeable differences found by Beckmann and Menkhoff (2008) in risk aversion among fund managers, for which Cohen's $d \approx 0.4$, IS ≈ 0.7, with the difference in means statistically significant. Assuming that the same distribution of risk preferences holds among female clients as among female fund managers, one can calculate that the chance of a randomly selected female client being matched on risk aversion with a randomly selected female manager is only 37.5%. If the randomly selected manager is, instead, male, the chance of a match is not much lower, at 25%. Of course, it is implicitly assumed in all of these studies that risk aversion can be measured accurately by a few survey questions or by a short laboratory experiment. Therefore, it seems quite odd that one would use sex to match advisors and clients, when simply asking clients about their risk preferences would allow them to be matched with an advisor far more accurately.

While evidence of an essentialist view is not found in all articles in this sample – as noted earlier, some researchers may report empirical (Statement A) results in essentialist (Statement B) language simply out of habit or convenience, without fully realizing the implications – enough examples of explicit essentialism are present to suggest that *a priori* beliefs are playing a role.

Digging for difference?

Of course, if one *begins with* the belief that women are by nature more risk averse from men in a substantial and important way – and so one expects this sex difference to be manifested across all types of studies and populations – it may be difficult to see what all the fuss is about. Each particular study that shows a statistically significant difference is then seen as confirming one's belief, and the logical fallacies described previously (and, in fact, this article – which will be merely skimmed, at best) may seem to be beside the point.

Because the purpose of this article is to point out profession-wide tendencies to diverge from the goal of objectivity due in large part to commonly shared cognitive biases, the following critique should not be taken as criticism of particular, individual authors or their peer reviewers. A whole literature can, apparently, drift in a particular direction due to widespread (though possibly erroneous) cultural beliefs combined with generally accepted (but in actuality, non-rigorous) methodological practices. At least five kinds of biases can be identified.

(1) Inaccurate citations of earlier literature

In reviewing the literature, one economics article states that 'Previous surveys of economics... and psychology (Byrnes et al., 1999) report the same conclusions: women are more risk averse than men in the *vast majority* of environments and tasks' (Croson & Gneezy, 2009, p. 449, emphasis added). Another article cites Byrnes et al. (1999) as demonstrating that 'females' lower risk preferences and less risky behavior *is robust across a variety of contexts*' (Eriksson & Simpson, 2010, p. 159, emphasis added). In fact, what Byrnes et al. (1999) actually conclude, after surveying studies of 322 different effects, is that 'the majority (i.e., 60%) of the effects support the idea of greater risk taking on the part of males' and 'a sizable minority (i.e., 40%) were either negative or close to zero' (Byrnes et al., 1999, p. 372).[9]

(2) Overemphasis on difference within a study's own results

In another study (Arano et al., 2010), a difference between married men and married women in the expected direction is highlighted in the text (Arano et al., 2010, p. 153), even though it is statistically insignificant. Meanwhile, differences between single men and single women go in the opposite direction. (That is, the point estimate of Cohen's *d,* should one calculate it, is negative, though also not statistically significant). Regression analysis is then pursued on the married subsample and, with the addition of various covariates, a statistically significant regression coefficient on gender in the expected direction is found. No further investigation of the single subsample is reported. Similarly, Bernasek and Shwiff (2001) undertake a long and painstaking analysis of gender differences in the percentage of defined contribution pension funds invested in stock. This is in spite of the fact that, in their data, there is no significant difference between the means for men and women. Unless one holds a prior belief that a difference *should* exist, it is hard to explain why such additional analysis would be pursued.

In another study (Beckmann & Menkhoff, 2008), the difference between male and female subjects on the most direct measure of risk aversion is statistically significant in only one of the four countries studied, and then only at a 10% significance level. Considering two other less direct questions as well, differences in only 5 out of 12 measures (four countries by three questions) are statistically significant: four at the 10% level and one at the 5% level. In a later section of the article, it is found that females are statistically significantly *more* likely than males, on average, to *increase* investment risk taken on, under certain circumstances (p. 378). Rather than taking this as evidence of possible higher male risk-aversion on average, a convoluted explanation is presented to justify this result as due to a greater presumed preference for conformity on the part of women ('strong ambition to stick close to the benchmark's performance,' Beckmann & Menkhoff, 2008, p. 378). While the data used in the study would seem to suggest that the evidence for greater female risk aversion is, at best, mixed, the article concludes that the data reveal 'a victory for gender difference' and (p. 367) 'robust gender differences' (p. 379) in the direction of women being 'significantly more risk averse' (p. 379).

Another study notes a 'striking gender difference' in probability weights calculated from a combination of data and a particular theoretical framework, while skipping quickly over the fact that the distribution of men and women in major decision-making types is found to be 'quite similar' (Bruhin, Fehr-Duda, & Epper, 2010, p. 1402). While the discussion of the 'striking gender difference' goes on for several pages, the practice of showing confidence bands – established in earlier sections of the article – is suddenly dropped. Earlier work on some of the same data (Fehr-Duda et al., 2006) had noted that male and female confidence bands for probability weights overlapped in 3.5 of the 4 treatments studied.

Another study reports, based on survey data, that 'Women are significantly less willing to take risks than men in all domains.' (Dohmen et al., 2011, p. 535) ($d \approx +0.40$, IS > 0.80). The same study, however, also includes a field experiment with 100s of subjects. An analysis of the field experiment data reveals only a marginally statistically significant difference by sex on one risk question ($d = 0.17$, IS $= 0.88$, $p = 0.07$) and no statistically significant sex difference ($d = 0.05$, IS $= 0.90$, $p = 0.60$) on the other. The published article does not report these weak to non-existent results.

In yet another study, findings of sex differences in the aggregate on one variable of interest are highlighted throughout, while the lack of any statistically significant difference in a risk experiment is addressed far more briefly, late in the article (Gneezy, Leonard, & List, 2009, p. 1652).

It appears that many researchers downplay evidence that fails to confirm a belief in women being more risk averse than men. No cases have been found, in this review of the literature, of bias in the other direction (that is, of playing *up* the results that fail to confirm the stereotype).

(3) Publication and confirmation bias

Biases in publication can occur if authors, reviewers, and editors tend to favor statistically significant over statistically insignificant results, or tend to favor results that are consistent with a prior belief to results which contradict a prior belief. Stanley and Doucouliagos (2010) have suggested a clever way for economists to detect such bias in the literature using only published works.

Based on the work of Light and Pillemer (1984), Stanley and Doucouliagos note that, in the absence of publication and confirmation bias, a plot of the precision of a parameter's estimates (calculated as the inverse of the standard error, SE) against the magnitudes of the estimates, taken from many studies, 'should be symmetric and shaped approximately as an inverted funnel' (p. 174). That is, the most precise estimates (generally from the largest samples) should come quite close to a 'true' parameter value, while one expects a wider – but symmetric – distribution of estimates as sample size falls and the effect of sampling error increases. Failure to report statistically insignificant estimates, they note, would result in a funnel graph that is 'hollow' (that is, contains few points in the region where the magnitude is smallest relative to the SE). A tendency to prefer estimates with a particular sign will result in a funnel graph that is asymmetric, with points on one side or the other of zero being scarce or missing. They give an example of estimates of an own-price elasticity of demand, for which the expectation that this should be negative seems to have created biased reporting.

Plots of the precision (1/SE) of the estimates of Cohen's d from Table 1 against their magnitudes are shown in Figures 2 and 3.[10] Figure 2 contains the main findings. Three studies have been omitted from Figure 2 and shown separately in Figure 3 simply because their scale or density would make Figure 2 difficult to read.

The highest precision cluster of points in Figure 2 comes from Sunden and Surette (1998), who examined data on 1000s of retirement portfolios.[11] The right-most point of $d = 1.13$ comes from Eckel and Grossman (2008) and is statistically significant. The left-most point of -0.79 comes from Holt and Laury (2002) and is not statistically significant. Points that are (as discussed above) over-emphasized in relation to their statistical significance and/or to other – conflicting – results within the same study are marked with solid circles. Those very near the horizontal axis, for example, are the non- or marginally-statistically significant results that Beckmann and Menkhoff (2008) claim show 'robust gender differences.' 'Downplayed' points, marked by hollow circles, are the statistically insignificant or 'wrong' sign results whose neglect has just been discussed above (from Arano et al., 2010; Bernasek & Shwiff, 2001; Dohmen et al., 2011; Gneezy et al., 2009).

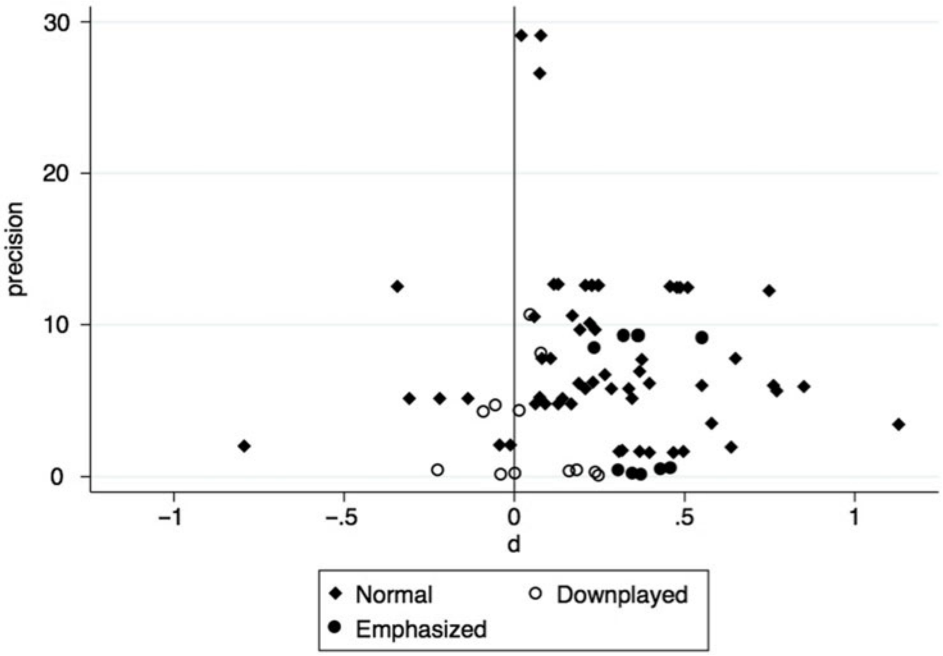


Figure 2. Funnel graph plotting precision versus magnitude of Cohen's *d*, for 15 studies.

Figure 3 shows results for three additional studies. Barber and Odean (2001) and Hartog et al. (2002) use data-sets with 10s of 1000s of observations (from actual stock investment accounts and a newspaper survey, respectively), giving their results (the top cluster in Figure 3) an unusually high level of precision. Both these studies show

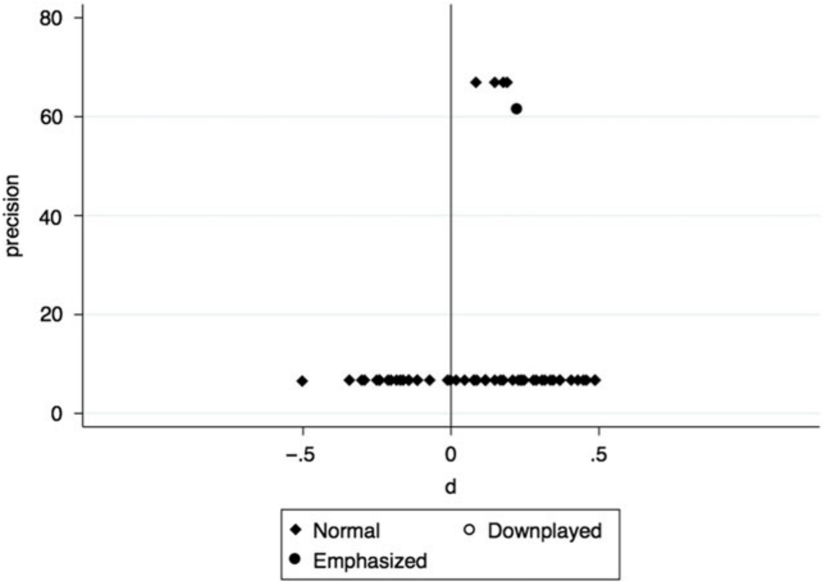


Figure 3. Funnel graph for the remaining three studies.

differences that, while highly statistically significant, fall in the range of $0.08 < d < 0.23$. Fehr-Duda et al. (2006) in a very unusual study present 181 student subjects with differently framed lotteries (as well as various probabilities) and explicitly address the issue of stereotyping. While the sheer volume of their not-easily-summarized results would make Figure 2 less readable, Figure 3 shows a range of results that span approximately $-0.50 < d < +0.50$. Those differences beyond approximately 0.30 (0.40) standard deviations from zero, in either direction, are statistically significant at the 5% (1%) level.

As can be seen, there is something of a 'hollow' aspect to the funnel graph in Figure 2, with statistically insignificant points tending to be downplayed. This indicates publication bias. More dramatic, however, is the fact that Figure 2 is markedly asymmetric. A simple average of the eight most precise estimates (from both figures) yields $d = 0.13$. Yet one sees in Figure 2 far fewer Cohen's d values to the left of this value than to the right. Confirmation bias is strongly indicated.

Stanley and Doucouliagos (2010) also point out that this analysis can – by suggesting what an average value over *all* studies, without bias, would be like – be used to 'provide an estimate of the overall magnitude of the empirical effect in question' (p. 172). While $d \approx 0.13$ may seem like a good candidate for such an estimate of gender differences in risk aversion, some caution is in order. It should first be asked whether there actually is some single risk aversion difference parameter 'out there,' independent of cultural and experimental contexts, which is being estimated.

(4) Failure to consider confounding variables

The essentialist explanation for observed differences is that greater risk aversion is a trait, characteristic, or essence shared by women by virtue of their being women. Differences in risk that appear to be strongly related to sex (when they occur) may, however, in fact be due (in part or completely) to a third, confounding variable, such as societal pressures to conform to gender expectations or to locations in a social hierarchy of power. Or they may no longer be seen when the sampling universe is broadened.

A literature in psychology has, in fact, grown up around the question of whether manipulations of gender salience and stereotype threat could explain observed sex differences (Carr & Steele, 2010; Ronay & Kim, 2006; Weaver, Vandello, & Bosson, 2013). Carr and Steele (2010), for example, create a 'stereotype threat' situation for women in a laboratory experiment, by asking subjects to record their gender before they were asked to do a lottery exercise, and describing these exercises as testing their mathematical abilities. An extensive psychological literature suggests that such a set-up may tend to erode women's performance by triggering anxiety about confirming a cultural belief (in this case that women have less mathematical ability). The stereotype-threat situation results in substantial and statistically significant sex differences in mean risk aversion between men and women ($d > 1$). In another group, however, subjects are not asked their gender until later, and the (exact same) exercises were described as being about puzzle solving. In this case, no statistically significant difference between the average behavior of men and women was found.

Differences in socialization patterns could also contribute to observed differences. The effects of same-sex versus co-ed schooling (Booth & Nolen, 2012), variations across cultural groups outside of Western industrialized societies (Gneezy et al., 2009), and studies of race and cultural worldviews (Flynn, Slovic, & Mertz, 1994; Kahan et al., 2007) suggest further variables that may have confounding effects. The variation – or disappearance – of sex differences across cultural contexts found in these studies makes the biological explanation appear less plausible, since an 'essential' sex characteristic should presumably not vary with social context.

In addition, considering that some of the studies deal with investment behavior, it is notable that few discuss the role of investment advice. A long tradition of treating 'widows and orphans' (de Goede, 2004) differently from male investors, as well as stereotypes about the presumed risk aversion of female investors (Eckel & Grossman, 2008, p. 15; Schubert, Brown, Gysler, & Brachinger, 1999, p. 385) may contribute to differences in average investment patterns being observed, independent of investors' own inclinations.

Stanley and Doucouliagos (2010) suggest that their funnel graph analysis could be 'used to identify moderator variables that explain the wide variation in the reported findings routinely found in economics research' (p. 172). Such an empirical exploration is left for future research.

(5) Examination of a narrow range of risks

The variety of types of risk studied is also quite limited, with lottery, gambling, and investment scenarios dominating the economic analysis. To what extent do these measure attitudes toward 'risk'?

Many authors seem to assume the existence of a general sex-identified risk-aversion utility parameter applicable to all contexts – in one case, for example, hypothesizing that risk tendencies observed in lottery choices could be extrapolated to preferences concerning marriage and the afterlife (Hartog et al., 2002, p. 16). Other studies examine somewhat broader phenomena such as driving behavior. The studies that claim that 'women are more risk averse than men,' however, do not in general include in consideration areas of life in which women on average take on elevated risks relative to men, for example in pregnancy and childbirth or in relation to domestic violence.[12]

The primary focus on lottery-type scenarios also tends to draw attention toward situations of Knightian 'risk,' in which both payoffs and probabilities are known. 'Uncertainty' is often narrowly interpreted in the literature as describing a case in which probabilities are not known, though the payoffs still are. Situations concerning the true sort of uncertainty generated simply by the fact that human beings live in a complex world that generates an unknown future receive less attention. Yet unforeseen events – e.g., new inventions, bursting asset bubbles, or negative environmental consequences – regularly surprise us, and can be of very large economic consequence (Randall, 2009; Taleb, 2010). It may be argued that lottery experiments have the advantage of being more amenable to study, but if a focus on tractability drives economists to only 'look under the lamppost' in studying risks, any generalization to larger scale real-world concerns should be considered epistemologically suspect.

How biases arise and persist

Many possible reasons could be given for the presence of gender-stereotyped biases in economics, including the positing of an explicit effort on the part of (still disproportionately male) economists to maintain male dominance. However, in the spirit of behavioral economics research, this work will explore cognitive phenomena that seem to effect even well-meaning researchers who do not, at a conscious level, endorse sexist views.

Essentializing and 'natural kinds'

Psychologists, along with other researchers including neuroscientists, linguists, and philosophers, have in recent decades called into question the notion that humans perceive simply what is 'out there,' and that human minds operate following rules of logic. Rather than thinking of mental processes as simply accepting all inputs and then processing them according to rules of rational calculation, cognitive scientists have increasingly come to see the mind as developed through bodily experience, with tendencies to selectively perceive information and process it in ways that serve the goals (such as survival) of the organism or species (Burton, 2008; Damasio, 1994; Lakoff & Johnson, 1999; Williams, Huang, & Bargh, 2009). Habitual modes of perception that very quickly sort the potentially dangerous from the benign, and cognitive 'short cuts' that save on mental effort, for example, may have a great deal of practical value for the maintenance of life, even though they fail on criteria of logic or accuracy (Gelman, 2005; Gigerenzer, 2007; Lakoff, 1987; Leslie, 2008; Most, Verbeck Sorber, & Cunningham, 2007). Stereotyping is one such mental 'short cut.'

Empirical studies suggest that from a very early age, humans create simple mental categories, taxonomies, or 'cognitive schema.' Items mentally placed within a given category are thought to be of the same 'natural kind,' and to all share in some deep-lying 'essence' (Gelman, 2005; Leslie, 2008). These essences are assumed to be not only generalizable to all members of a kind, but also to be immutable (Prentice & Miller, 2006): 'We essentialize a kind if we form the (tacit) belief that there is some hidden, non-obvious, and persistent property or underlying nature shared by members of that kind, which causally grounds their common properties and dispositions.' Items that belong in a different category are thought to lack that particular 'essence' (Leslie, in press).

Sciences take a taxonomic approach when identifying different species of animals or chemical elements. When applied to some kinds of physical phenomena, assertions about persistent, common, essential properties seem non-controversial, and we can forget that the creation of 'kinds' involves a sort of folk theorizing based on hidden properties.

However, the evidence suggests that our brains carry the habit of essentializing over into far less clear-cut realms. From a very young age, children also observe sexual dimorphism, meaning the constellations of different physical traits (height, voice pitch, etc.) that are typically associated with men and women. It is not surprising that such observations give rise to a folk belief in gender 'essences.' In fact, gender is one of the strongest examples of delineation of 'natural kinds' found in the psychological literature (Prentice & Miller, 2006, p. 130). The attributions of 'essential' sex-related characteristics commonly extend well beyond the constellations of observable preponderant physical traits to beliefs about natural, distinct, stable traits of personality and social behavior (Carothers & Reis, 2013; Prentice & Miller, 2006, p. 130).

Why isn't this solved by empirical study?

Social scientists are used to looking to empirical evidence for confirmation or disconfirmation of claims. Beliefs in essences and natural kinds, however, have a complex relationship to empirical evidence. Are they based on universal observations? Many observations? A single observation? No observation?

Research indicates that they do not seem to be based on universal observations. Linguistically, when we make a statement about the (presumed) essential properties of a natural kind, we use a 'generic' noun (Gelman, 2005, p. 3). Examples include 'tigers have stripes.' Such a statement is a generalization that 'refers to a category rather than a set of individuals' and 'express[es] essential qualities' implying that a category 'is coherent and permits categorywide inferences' (Gelman, 2005, p. 3). In a generic of the form 'Fs are G,' that is,

> one is saying *of a kind of thing*, specified in the statement, that its members are, or are disposed to be G (or to [do] G) *by virtue of being of that kind*. The speaker conveys that being G is somehow rooted in what it is to be an F: G-ing is what Fs do (or are disposed to do) by virtue of being F. (Haslanger, 2011, p. 13)

The belief that there is an 'innate, genetic, or biological basis' for a statement is also a characteristic of many generics (Gelman, 2005, p. 1).

While generic claims such as 'Tigers have stripes' appear to state a universal quality, the existence of albino tigers, for example, is not generally seen as nullifying the statement that 'Tigers have stripes.' Having stripes is considered to be part of the intrinsic, essential nature of tigerhood even if stripes are not manifested in a particular case (Khemlani, Leslie, & Glucksberg, 2012).

Nor do claims about essences seem to be based on a majority of evidence. While one might suppose that the generic statement 'Fs are G' might be functionally equivalent to quantitative statements such as 'Fs are *more* G, *on average*,' '*most* Fs are G,' or '*a majority of* Fs are G,' this turns out not to be the case, either. The more nuanced quantitative statement (e.g., 'Fs are *more* G, *on average*') has been called an '*aggregate*-type proposition' in the statistical literature (Bakan, 1955, 1966, p. 433), as contrasted to the generic proposition. The complicated relationship between generic statements and aggregate statements is a topic of discussion in the psychological and philosophical literatures. While some generic statements seem to be accepted as true based on statistical prevalence (e.g., 'cars have radios' discussed in Khemlani et al. (2012)), and some research has suggested Bayesian models for the formation of generics, statistical prevalence cannot, in fact, explain many cases (Pelletier, 2009). Prevalence in a majority of a kind seems to be neither necessary nor sufficient for a generic to be considered true. For example, the generic statement 'Ducks lay eggs' is generally accepted as true. In fact, only a minority of ducks (i.e., those that are female, mature, and non-sterile) lay eggs (Khemlani et al., 2012). We seem to reason that ducks are birds, and birds, as a 'kind' or category, have the 'characteristic' of reproducing by way of eggs. On the other hand, 'Canadians are right-handed' is rejected, even though a majority of Canadians are right-handed (Khemlani et al., 2012).

Apparently, it may take only a *single* observation of a 'difference' for human minds to embellish a belief in natural kinds with a new 'essential' characteristic. In a recent study, psychologists had subjects take an arbitrary test (counting dots on slides) that had no relation to gender, but which the subjects were told measured their 'perceptual style.' Those who took the test alongside an opposite-sex partner and were informed that they had the opposite style to their partner were markedly more inclined, on average, to judge that their 'style' would be typical for their sex, and rare in the opposite sex, than those tested under other conditions (Prentice & Miller, 2006).

What about reasoning in the opposite direction, from natural kinds to individual instances? Research indicates that the propensity to essentialize is so strong and basic that a statement phrased as a generic and accepted as true predisposes people to believe that *individual members* of a class will have the stated property (Khemlani, Leslie, & Glucksberg, 2009, p. 447) – that is, to interpret generic statements in ways that essentialize or even universalize the association. For example, given the statement 'Quacky is a duck,' people tend to agree with the statement 'Quacky lays eggs' (Khemlani et al., 2012, p. 10). In the social realm of beliefs about men and women, generalization from presumed traits of groups to expectations about individuals is, of course, the essence of prejudice and discrimination.

So – tempting as it may be to think so – the existence or non-existence of essences cannot be proven through statistical analysis. Rather, the tendency to class things in terms of kinds seems to be part of the structure of our *inside* worlds – that is, of evolved, developmental human cognition.

The persistence of biases

Once a belief in a 'natural kind' is established, it is extremely difficult to shake. Evidence from psychological studies indicates that, when required to complete a task that requires working *against* stereotypes, subjects need more time and even use different parts of their brains (as indicated by functional magnetic resonancing imagery scans) than when allowed to use their well-established mental categories (Knutson, Mah, Manly, & Grafman, 2007).

The phenomenon of confirmation bias has been known since Frances Bacon's time (see quote that opens this article) and is very well documented in the psychological literature (Nickerson, 1998). When we believe the world to be characterized by essences, the very 'naturalness' of these 'core conceptual beliefs' (Khemlani et al., 2012, p. 1) makes them feel uncontestable, and we are drawn to evidence that confirms our pre-existing belief. Evidence that *psychologically* confirms a pre-existing belief is often mistaken as *logically* confirming that belief (Nickerson, 1998, p. 179). Meanwhile, counterevidence regarding non-conforming individuals, or the fuzziness or overlapping of categories, or alternative (non-'essential') explanations of causality or association may have little or no effect.

Effects in scientific fields

Unfortunately, while highly intelligent scholars who seek to do scientific research may like to believe ourselves to be above such failings, psychological research suggests that we are not immune. Nickerson's (1998) review of the confirmation bias literature, for example, provides numerous examples of the phenomenon affecting scientific fields. Nor have cultural developments moved us past sex stereotyping. A 2012 study, for example, found that biology, chemistry, and physics faculty at research-intensive universities rated identical application materials differently (both statistically and substantively), on average, according to the presumed sex of the applicant (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). Another recent study found that the more one feels that one is an 'objective, rational actor,' the *more* likely one is to have confidence in one's stereotyped beliefs and act on them (Uhlmann & Cohen, 2007).

Sex stereotyping may be particularly prevalent in current neuroscientific research. The attribution of (on average) different psycho-social behaviors to (fundamental) sex differences in hormones and/or brain structure, further explained as caused by differences in evolutionary pressures on bodies with different reproductive roles, can currently be found in many studies. It is possible, however, that these stories may also reflect a good deal of folk theorizing combined with confirmation bias. While plausible and compelling, the biological and evolutionary explanations may not hold up when examined under more rigorous standards. Scholars including Barnett and Rivers (2004), Eliot (2009), Hyde (2005), Fine (2010), and Jordan-Young (2010) have pointed out numerous methodological flaws in a range of such studies. Many of these flaws, such as publication bias, also occur in economics, as demonstrated here.

Additional methodological notes

This article points out how paying attention to overlaps in distributions (using tools such as Cohen's *d* and IS) can help diagnose and prevent invalid stereotyping. However, more general methodological lessons can also be learned from this case.

Statistical significance versus inductive reasoning

The case of gender and risk aversion provides a good example of a statistical fallacy that has a long history. In the psychological methods literature, it was long ago suggested by Bakan (1966) that some confusion among researchers between Fisherian statistical inference and inductive reasoning may be behind considerable misinterpretation of statistical results. Fisherian inference means going from sample results concerning an aggregate, such as a sample difference in means, to inferences about the corresponding population aggregate. To reason inductively, on the other hand, means to go from specific observations to hypothesizing general propositions that invite conclusions about the *nature* of the subjects of study.

Fisherian significance testing about a difference in means only (at best) justifies the inference that a difference in means in a sample corresponds to a difference in means in a population: That is, Fisherian inference creates the basis for Statement A: 'In our sample, we found a statistically significant difference in mean *X* between men and women.' Fisherian inference does *not* justify generalizing an (sample) aggregate statement to a generic statement, such as Statement B, 'Women are more *X* than men,' in which 'women' and 'men' are considered as different natural kinds. It also does not justify generalizing from the particulars of a study's construction – e.g., exact wording of questions; age, and other demographic characteristics of the particular population studies; what the subjects were doing before the survey or experiment – to what one would find if these particulars were varied (Bakan, 1966). Finally, Fisherian inference does not justify generalizing from a study of a particular variable – e.g., a particular type of risk such as in lottery playing – to other variables, e.g., risk taking on the job.

The making of unfounded inductions from Fisherian inferential studies is not confined to the economics literature on sex and risk. Deaton's (2010, pp. 439–442) discussion of randomized control experiments in development economics makes a similar point: in the case where the only information a statistical study yields is a difference in means, this does not justify inferences about other aspects of the distributions; does not itself supply the backstory (or 'mechanism') explaining why the difference occurs; and does not (in the presence of heterogeneity) provide a clear-cut guide for making decisions concerning individual instances. Only when we read into the statistical results, the existence of a generic relationship – e.g., 'dams harm development' or 'women are more risk averse' – are we tempted to draw hard-and-fast conclusions about a specific dam or a specific woman.

Taxometric methods

While *d* values and IS are univariate, other researchers have introduced 'taxometric' measures, which look at differences from a multivariate perspective (Carothers & Reis, 2013; Meehl, 1992). If belonging to the natural kind (or 'taxon') called 'men' has categorical implications over a range of physical and behavioral characteristics, then it should be possible to find sets of variables that tend toward accurately distinguishing 'men' from their complement, 'women.' On the other hand, if men and women are actually quite similar on other variables, differing only in the *degree* to which observables are manifested, sorting will be less accurate and the difference can be regarded as 'dimensional' rather than 'taxonic.' Using such techniques, Carothers and Reis (2013) found that a set of variables measuring performances in four strength-related track-and-field events and another set of variables measuring anthropomorphic characteristics (e.g., weight, height, arm circumference), both by and large showed taxonic structures.[13] For these variables, effect sizes tended to be large and consistent, and the scores of individual men across the different variables tended to show a consistent pattern of being above (or below) the female means. On the other hand, for nearly all of the groups of psycho-social variables they studied (e.g., sexual attitudes, care orientation, fear of success), the results indicated a dimensional construct. While differences in averages could be detected, the frequent occurrence of individuals scoring in a sex-stereotyped direction on one variable, but in a counter-stereotyped direction on another, did not permit categorical sorting. Application of this technique to behavioral economics research is a topic for future study.

The need for a wider community of scholars

When a stereotype is believed by all members of a community, confirmation bias is likely to be rife. As has been demonstrated for the literature on gender and risk aversion, confirmation bias has resulted in misleading knowledge claims even among researchers who are using the methods (such as regression analysis) commonly thought to be 'objective' and 'rigorous.' How, then, can a better knowledge of methodology help to root it out?

While techniques that can be adopted by individual researchers, such as Cohen's *d* and IS, can be helpful, the mere availability of techniques does not assure that they will be used. Cohen's *d* has been used in other disciplines for years.

A more sensible notion of scientific objectivity and rigor defines reliable knowledge as that which passes the test of evaluation by larger, more diverse communities that bring to bear a variety of perspectives (e.g., Keller, 1985; Kitcher, 2011; Longino, 1990; Nelson, 1996; Sen, 1992). An economics profession that is widely diverse across the lines of gender, race, class, nationality, and other identities would therefore not merely be more socially representative, but also is likely to be less prone to confirmation bias arising from locally held social beliefs.

Conclusion

The economics literature on gender and risk aversion reveals considerable evidence of 'essentialist' prior beliefs, stereotyping, publication bias, and confirmation bias. The claims made about gender and risk have gone far beyond what can be justified by the actual quantitative magnitudes of detectable differences and similarities that appear in the data. While such tendencies toward stereotyping and confirmation bias have a cognitive and neurological basis, a discipline that aspires to be a social 'science' should take steps to correct and prevent such bias when possible.

This article demonstrates how several methods, including calculating Cohen's *d* values and the IS, and creating funnel graphs, can help to diagnose cases of confirmation bias. To prevent future cases, both a widespread adoption of these expanded methods and a more inclusive community of scholars are required.

Acknowledgement

This work was supported by the Institute for New Economic Thinking [grant number INO1200012]. Matthew P. H. Taylor, Sophia Makemson, and Mayara Fontes supplied expert research assistance.

Notes

1. While inclusion of intersex, transgendered, or transsexual subjects would clearly complicate – and enrich – this analysis, this study focuses on the economics literature, in which a sex binary is assumed.
2. The scholarly literature supporting these claims is discussed in the section on ‘How Biases Arise and Persist,’ below.
3. Nelson (2013) discusses an additional notable case in which such testing is neglected.
4. In the literature reviewed here, Eckel and Grossman (2008, p. 15) and Dohmen et al. (2011, pp. 530, 540) are notable for providing any extended discussion of substantive economic significance.
5. This is most often estimated as:

$$s_{\mathrm{p}} = \sqrt{\frac{(n_{\mathrm{m}} - 1)s_m^2 + (n_{\mathrm{f}} - 1)s_f^2}{n_{\mathrm{m}} + n_{\mathrm{f}}}},$$

where s_{m}, s_{f}, n_{m}, and n_{f} are the standard deviations and sample sizes for the male and female samples.

6. Hyde found that 78% of reported empirical sex differences were smaller than this value.
7. A 10% level was chosen, rather than 5% or 1%, to give the existence of ‘difference’ the maximum benefit of doubt. Numeric values for d (or IS) were not calculated when differences were not statistically significant, because of the rather wild values that occurred in some of the small samples.
8. Because IS is non-directional, it is worth noting that one instance of IS $= 0.67$ (Beckmann & Menkhoff, 2008) is for a case where *fewer* men than women chose a risky option.
9. Eriksson has since acknowledged the error in the statement made in his co-authored 2010 article (2012).
10. The formula for the standard error of Cohen’s d is from Cooper and Hedges (1994).
11. The risk-taking measure is created by examining the proportion of men and women whose portfolios were mostly stock. One measure is statistically significant at the 5% level, the other is so close to zero ($d = 0.02$) that it is not.
12. It may be objected that risks of pregnancy and childbirth are not relevant for comparisons between the sexes, since men do not participate in them. Anecdotally, at least, however, it seems that there is little reticence about making inferences about risk taking from men-only activities. Flying fighter planes, for example, or engaging in combat is commonly taken as indicating masculine bravery, even though institutional constraints have largely prevented women from engaging in these activities.
13. A study of preferred sex-stereotyped activities (e.g., watching boxing, scrapbooking), also showed a taxonic structure (p. 391), but this was in good part by construction (p. 393). All subjects in this study were self-identified heterosexual men and women.

References

Arano, K., Parker, C., & Terry, R. (2010). Gender-based risk aversion and retirement asset allocation. *Economic Inquiry*, *48*, 147–155.

Archer, J. (1996). Comparing women and men: What is being compared and why? *American Psychologist*, *51*, 153–154.

Bakan, D. (1955). The general and the aggregate: A methodological distinction. *Perceptual and Motor Skills*, *5*, 211–212.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*, 261–292.

Barnett, R., & Rivers, C. (2004). *Same difference: How gender myths are hurting our relationships, our children, and our jobs*. New York, NY: Basic Books.

Beckmann, D., & Menkhoff, L. (2008). Will women be women? Analyzing the gender difference among financial experts. *Kyklos*, *61*, 364–384.

Bernasek, A., & Shwiff, S. (2001). Gender, risk, and retirement. *Journal of Economic Issues*, *35*, 345–356.

Blau, F. D., Ferber, M. A., & Winkler, A. E. (2010). *The economics of women, men and work*. Upper Saddle River, NJ: Pearson Prentice Hall.

Booth, A. L., & Nolen, P. (2012). Gender differences in risk behaviour: Does nurture matter? *The Economic Journal*, *122*, F56–F78.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, *7*, 649–658.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, *78*, 1375–1412.

Burton, R. A. (2008). *On being certain: Believing you are right even when you're not*. New York, NY: St. Martin's Press.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383.

Camerer, C. F., & Loewenstein, G. (Eds.). (2003). *Advances in behavioral economics*. Princeton, NJ: Princeton University Press.

Carothers, B. J., & Reis, H. T. (2013). Men and women are from earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology*, *104*, 385–407.

Carr, P. B., & Steele, C. M. (2010). Stereotype threat affects financial decision making. *Psychological Science*, *21*, 1411–1416.

Cooper, H. M., & Hedges, L. V. (1994). *The handbook of research synthesis* (Vol. 236). New York, NY: Russell Sage Foundation.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*, 448–474.

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: A meta-analysis. *Psychological Bulletin*, *137*, 97–130.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: G.P. Putnam's Sons.

de Goede, M. (2004). Repoliticizing financial risk. *Economy and Society*, *33*, 197–217.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*, 424–455.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*, 522–550.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*, 210–217.

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*.

Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1–17.

Eliot, L. (2009). *Pink brain, blue brain: How small differences grow into troublesome gaps: And what we can do about it*. Boston, MA: Houghton Mifflin Harcourt.

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*, 159–163.

Eriksson, K. (2012). År kvinnor mindre riskvilliga ån mån? Blog. http://bloggar.tidningencurie.se/kimmoeriksson/ar-kvinnor-mindre-riskvilliga-an-man/

Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision*, *60*, 283–313.

Fine, C. (2010). *Delusions of gender: How our minds, society, and neurosexism create difference*. New York, NY: W.W. Norton.

Flynn, J., Slovic, P., & Mertz, C. K. (1994). Gender, race, and perception of environmental health risks. *Risk Analysis*, *14*, 1101–1108.

Gelman, S. A. (2005). Essentialism in everyday thought. *Psychological Science*, *Agenda*, (May), 1–6.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconcious*. New York, NY: Penguin Books.

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and patriarchal society. *Econometrica*, *77*, 1637–1664.

Harris, C. R., Jenkins, M., & Glaser, D. (2006). Gender differences in risk assessment: Why do women take fewer risks than men? *Judgment and Decision Making*, *1*, 48–63.

Hartog, J., Ferrer-i-Carbonell, A., & Jonker, N. (2002). Linking measured risk aversion to individual characteristics. *Kyklos*, *55*, 3–26.

Haslanger, S. (2011). Ideology, generics, and common ground. In C. Witt (Ed.), *Feminist metaphysics: Explorations in the ontology of sex, gender and the self* (pp. 179–208). Dordrecht: Springer.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*, 1644–1655.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592.

Hyde, J. S., & Plant, E. A. (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist*, *50*, 159–161.

Jordan-Young, R. M. (2010). *Brain storm: The flaws in the science of sex differences*. Cambridge, MA: Harvard University Press.

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, *4*, 465–505.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 1449–1475.

Keller, E. F. (1985). *Reflections on gender and science*. New Haven, CT: Yale University Press.

Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2009). *Generics, prevalence, and default inferences*. CogSci 2009: 31st annual meeting of the Cognitive Science Society, Amsterdam.

Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2012). Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes*, *27*, 887–900.

Kitcher, P. (2011). Science in a democratic society. Amherst, NY: Prometheus Books.

Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping*, *28*, 915–930.

Kristof, N. D. (2009, February 7). Mistresses of the universe. *New York Times*.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, NY: Basic Books.

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1–47.

Leslie, S.-J. (in press). The original sin of cognition: Fear, prejudice and generalization. *The Journal of Philosophy*.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lindquist, G. S., & Säve-Söderbergh, J. (2011). 'Girls will be girls', especially among boys: Risk-taking in the 'daily double' on Jeopardy. *Economics Letters*, *112*, 158–160.

Longino, H. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton, NJ: Princeton University Press.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male–female differences: A computer simulation. *American Psychologist*, *51*, 157–158.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, *60*, 117–174.

Meier-Pesti, K., & Penz, E. (2008). Sex or gender? Expanding the sex-based view by introducing masculinity and femininity as predictors of financial risk taking. *Journal of Economic Psychology*, *29*, 180–196.

Miller, J. E., & van der Meulen Rodgers, Y. (2008). Economic importance and statistical significance: Guidelines for communicating empirical research. *Feminist Economics*, *14*, 117–149.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*, 16464–16479.

Most, S. B., Verbeck Sorber, A., & Cunningham, J. G. (2007). Auditory Stroop reveals implicit gender associations in adults and children. *Journal of Experimental Social Psychology*, *43*, 287–294.

Nelson, J. A. (1996). *Feminism, objectivity and economics*. London: Routledge.

Nelson, J. A. (2013). Not-so-strong evidence for gender differences in risk taking. UMass Boston Economics Working Paper No. 2013-06.

Nelson, J. A. (2014). Are women really more risk-averse than men? A re-analysis of the literature using expanded methods. *Journal of Economic Surveys*. doi:10.1111/joes.12069/abstract

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Olsen, R. A., & Cox, C. M. (2001). The influence of gender on the perception and response to investment risk: The case of professional investors. *The Journal of Psychology and Financial Markets*, *2*, 29–36.

Pelletier, F. J. (2009). *Kinds, things, and stuff: Mass terms and generics*. New York, NY: Oxford University Press.

Powell, M., & Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. *Journal of Economic Psychology*, *18*, 605–628.

Prentice, D. A., & Miller, D. T. (2006). Essentializing differences between women and men. *Psychological Science*, *17*, 129–135.

Randall, A. (2009). We already have risk management: Do we really need the precautionary principle? *International Review of Environmental and Resource Economics*, *3*, 39–74.

Reskin, B. (1993). Sex segregation in the workplace. *Annual Review of Sociology*, *19*, 241–270.

Ronay, R., & Kim, D.-Y. (2006). Gender differences in explicit and implicit risk attitudes: A socially facilitated phenomenon. *British Journal of Social Psychology*, *45*, 397–419.

Schubert, R., Brown, M., Gysler, M., & Brachinger, H. W. (1999). Financial decision-making: Are women really more risk-averse? *American Economic Review*, *89*, 381–385.

Sen, A. (1992). The Lindley lecture: objectivity and position. Lawrence: University of Kansas.

Sunden, A. E., & Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. *The American Economic Review*, *88*, 207–211.

Stanley, T. D., & Doucouliagos, H. (2010). Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys*, *24*, 170–191.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable*. New York, NY: Random House.

Uhlmann, E. L., & Cohen, G. L. (2007). ‘I think it, therefore it’s true’: Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*, 207–223.

Weaver, J. R., Vandello, J. A., & Bosson, J. K. (2013). Intrepid, imprudent, or impetuous?: The effects of gender threats on men’s financial decisions. *Psychology of Men & Masculinity*, *14*, 184–191.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. (1–24 in downloadable document).

Williams, L. E., Huang, J. Y., & Bargh, J. A. (2009). The scaffolded mind: Higher mental processes are grounded in early experience of the physical world. *European Journal of Social Psychology*, *39*, 1257–1267.

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error of regressions in the American economic review. *The Journal of Socio-Economics*, *33*, 527–546.