Not-So-Strong Evidence for Gender Differences in Risk Taking

Julie A. Nelson

WORKING PAPER 2013-06

DEPARTMENT OF ECONOMICS

UNIVERSITY OF MASSACHUSETTS BOSTON

Not-So-Strong Evidence for Gender Differences in Risk Taking

Julie A. Nelson
Professor and Chair
Department of Economics
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125 USA
Julie.nelson@umb.edu
+011 (617)287-6925 phone
+011 (617)287-6976 fax

December 22, 2013 (corrected)

Abstract

In their article "Strong Evidence for Gender Differences in Risk Taking," Gary Charness and Uri Gneezy (2012) review a number of experimental studies regarding investments in risky assets, and claim that these yield strong evidence that females are more risk averse than males. This study replicates and extends their article, demonstrating that its claims and methods are problematic. While it Charness and Gneezy assert that differences exist at the *individual* level, the data indicate only that modest differences (at most) exist at *aggregate* levels, such as group means. The evidence in favor of gender difference is thus considerably less strong than claimed. This analysis has important implications both for the design of behavioral and economic research and for policies related to discrimination and stereotyping.

Keywords: risk aversion; gender; risk; stereotyping; effect size

JEL codes: D03, D81, J16, B41, C12

1 Introduction

In their article "Strong Evidence for Gender Differences in Risk Taking," published in this journal, Gary Charness and Uri Gneezy (2012) review over a dozen studies in which subjects decide how much ($*x*) of a an endowment they will invest in a risky asset, rather than keep. They conclude that "The striking and consistent result is that despite the large environmental differences among the sets of experiments, a consistent gender difference is reported: Men choose a higher *x* than women do" (2012). The evidence is thus compelling, they say, in favor of the claim that females are more risk-averse than males.

This study replicates and extends Charness and Gneezy's review, and demonstrates that the empirical methods used were problematic. Re-examining the evidence they present using widely accepted statistical techniques demonstrates that the case for gender differences in risk taking is considerably less substantial than they portray. This result has implications both for behavioral and economic research and for policy.

2 What Are "Gender Differences"?

Before examining the appropriateness of the claims made and statistical techniques applied by Charness and Gneezy (2012), it will be helpful to distinguish between two possible meanings of "gender difference." These will be referred to as GDI, for individual-level differences, and GDA, for aggregate level differences.

2.1 Gender Differences at the Individual-Level (GDI)

GDI categorically distinguish between men and women. Suppose, for example, men "have" or "do" a particular value of a variable *x,* while women "have" or "do" the value $x - d$.

The trait "number of Y chromosomes per cell," for example, would be such a categorical classifier, with *$x=1$* and *$d=1$*. A statement such as "men have a higher *x* than women" is perfectly and universally valid in this case (assuming a chromosomal definition of sex). Furthermore, any pairwise comparison of randomly selected individual men and women will confirm the statement.

The GDI case can be made slightly more complicated by allowing for some amount of variability among men, and among women, in the value taken on by *x*, as long as the within-group variability is small relative to the value of *d*. Even with such variability, we would still expect that that the great preponderance of pairwise comparisons would confirm the statement. One might confirm such a GDI case by, for example, taking independent samples of men and women, computing $d_i = x_{Mi} - x_{Wi}$ for each randomly matched pair (defining $d_i > 0$ as a binomial distribution "success"), and apply the binomial test to H_o*:* $p(d_i > 0) \leq .95$ *or* $.99$ or some other desired level of confidence, versus H_A*:* $p(d_i > 0) > .95$ *or* $.99$. Finding that the men's *x* values are generally higher $(d_i > 0)$ while rejecting the null would confirm that men are consistently (to within one's desired confidence level) observed to have a different, and greater, value of *x*. Applying this result to the real world, it would mean that observing an individual man and an individual woman (for example, two job candidates), one could predict with a high level of confidence that the man has more *x* (for example, if risk differences were found to be GDI, more tolerance for risk) than the woman. Reversing this and taking a Bayesian view, if a difference is truly GDI and categorical, one would also be confident judging someone's sex by observing their level of *x*.

2.2 Gender Differences at the Aggregate-Level (GDA)

GDA, on the other hand, are found at the aggregate or distributional level, such as differences between men's and women's *average x* values.[1] While GDI implies GDA, the converse is not true: The smaller the substantive difference between the men's and women's averages, and the larger the intra-sex variability, the more the men's and women's distributions overlap, and less the GDA model resembles the categorical case of GDI.

In the case of GDA, conventional tests for statistical significance of the difference in averages would be appropriately applied. When the case is quite far from GDI, one may only be able to detect the existence of a GDA in large samples. Since the difference is detected at the aggregate—not individual—level, no strong expectations are created about what pairwise random comparisons would reveal. Examples could include phenomena such as academic test scores or wages. Cases of $x_{Mi} - x_{Wi} < 0$ (that is, "this individual woman has a higher *x* than this individual man") would generally exist among random pairings at the individual level, even though $\bar{x}_M - \bar{x}_W > 0$ (men's average is higher), and statistically significantly so. The *substantive* size of the difference should also be of interest. To the extent that the overlap between men's and women's distributions is considerable, one would not, then, be able to take the sex of individual males and females as a reliable indicator of their relative levels of *x*. To do so would be to stereotype: that is, to invalidly generalize from a group characteristic to an individual.

[1] While claims about differences in means are the most common in the literature, GDAs could also be found by comparing medians, by comparing variances or other moments of the distributions, via permutation tests, or by other methods that compare distributions. Claims about differences can also be examined in a multivariate context, by distinguishing between groups of variables that define "taxa" or categories, and groups that of variables that reflect only "dimensional" or distributional differences (see Carothers and Reis, 2013).

2.3 A Note on GDI, GDA, and the Causes of Difference

It may be tempting to think of a GDI as indicating a fundamental biological sex difference, and certainly a substantial literature now exists claiming that behavioral differences by sex are large and caused by differences in fundamental neurological structures (see reviews in Fine, 2010; Jordan-Young, 2010). However, this is not an accurate understanding of the distinction being made here, which is statistical.

A GDI that is sex-based should hold for *all* males and females from the entire human population. However, statistical GDI at times may also be distinguishable in subpopulations due to cultural norms or socialization. For example, in a highly gender-differentiated society, one may categorize an individual by sex, with a high degree of reliability, merely by observing hair length or the quantity of jewelry worn—or conversely, predict his or her apparel from his or her sex. Such cultural factors may also cause other behavioral differences.

On the other hand, it is possible that a GDA, while more weakly related to sex due to variations within each sex and overlap between them, may yet have biologically-influenced causes or reflect some interaction of nature and nurture.[2] While these issues of causes may be debated elsewhere, the statistical point being made here is that when a "difference" is GDA, one cannot validly claim that it holds at an "individual" level, nor imply that it should hold in pairwise comparisons.

[2] Over a long enough time span, of course, distinctions between nature and nurture become less clear cut: Over evolutionary time for a species or developmental time for an organism, many biological features come to be shaped by experience, and thus by culture.

3 GDI vs. GDA in Charness and Gneezy (2012)

Charness and Gneezy (2012) (henceforth, CG) appear to frequently confuse (possible) GDA evidence with (strong) GDI evidence, and depart from standard practice in their use of statistical tools. Their approach, the present study demonstrates, is overwhelming more consistent with a GDI (individual-level) mental model of difference, than with a GDA (aggregate-level) model of difference. There are also suggestions that CG lean towards a categorical and biological-level explanation for their findings. This evidence is presented immediately below.

Applying more standard and appropriate statistical tools, Section 4 of this study will show that the data CG examine actually, at best, supports an aggregative (GDA) model of gender differences in risk taking. In Section 5, it will be noted that social norms may contribute in important ways to risk GDA, and that distinguishing between GDA and GDI in risk preferences is important for both for scientific investigation and for correct economic policies.

3.1 Evidence for a GDI Model

There are several aspects of CG that are consistent with a GDI model and less suggestive of, or even inconsistent with, a GDA model. These include:

- References to the differences being studied as "individual" differences (p. 50).
- The offering of evolutionary explanations (footnote, p. 50) and hormonal (prenatal testosterone, p. 51) explanations for observed differences, suggestive of biologically-linked sex-based dichotomous traits.

- The repeated statement of differences in generic, biological terms, as occurring between males and females (e.g., "females are substantially more financially-risk-averse than males," p. 51, and similar throughout the work), generally with no caveat about the difference referred to as applying only in the aggregate (that is, no clarification that these apply only "on average").
- The use of a categorical, binary distinction to characterize differences in risk preferences: Male investments in the risky asset are categorized as either greater than, or not greater than, female investments, even though the underlying data are distributional and continuous.
- For ten (out of the fourteen unique) studies, no testing of the statistical significance of the difference in means is performed. Differences between *point estimates* of the male mean and the female mean are interpreted as evidence for a difference between male and female behavior, regardless of statistical significance or the lack thereof.
- In those same ten cases, no acknowledgement of intra-sex variability is made--that is, there is no presentation of standard deviations or standard errors of the means.
- Having (for the most part) foregone conventional tests of the statistical significance of differences in means, the authors adopt a very unconventional approach: Binomial tests of H_o: $p(d_i > 0) = .5$ are applied to paired results of male and female *means* across various treatments within two studies (found in CG Tables 1 and 3), and across 10 studies (CG Table 4), regardless of whether the differences in means are statistically significant or not. This appears to be an odd hybrid of GDI reasoning (perhaps suggesting binomial tests *on individuals*) and GDA reasoning (setting the

null hypothesis at 50/50 odds). It is very unclear what sort of statistical theory would justify this approach.

- Figures presented in the publication are designed to contrast values for males and females, without labeling the vertical axis as representing an average or including any information about with-sex distributions.

3.2 Evidence for a GDI Model

The evidence that GDA model might be implied is more limited:

- There is one reference to "difference in risk taking between groups" (p. 50).
- The text and tables refer to the values given as "averages."
- Tests of statistical significance for the difference between male and female means or medians, or for a regression coefficient on gender, are presented for four (of the fourteen) studies reviewed.

The next section explores which of these models—GDI or GDA—is most supported by the data.

4 Evidence from the Data

CG (2012) review over a dozen studies in which subjects decide how much ($*x*) of a an endowment they will invest in a risky asset, rather than keep. In their work, they compare the average (usually the mean) value of *x* chosen by men to the average value chosen by women. The current study replicates the CG investigation, augmenting their analysis with additional statistical measures and tests. In a few cases, the CG article provides statistical information sufficient to make these extensions; otherwise, raw data was requested from CG and/or a study's authors.

For comparison across the several studies, it will be helpful to express the differences in means found in a standardized way that takes into account both intra- and inter-group variability. This is easily done using *Cohen's d*, a standardized measure of effect size in wide use in the psychology and education literatures (e.g., Byrnes et al., 1999; Cross et al., 2011; Hyde, 2005; Wilkinson and Task Force on Statistical Inference, 1999).[3] For the case of a male versus female comparison, it is conventionally calculated as

$$d = \frac{\bar{X}_m - \bar{X}_f}{s_p}$$

where $\bar{X}_m$ is the male mean, $\bar{X}_f$ is the female mean, and s_p is the pooled standard deviation, a measure of the average within-group variation.[4] As conventionally set up in the psychological literature on gender differences, a positive value for *d* represents a case where the male score exceeds the female score. The difference is now expressed in standardized (standard deviation) units, and the measure quite sensibly gives a reduced measure of "difference" as the within-group variability (reflected in a rising s_p) increases.

Some insight into its interpretation, relative to the tests appropriate for detecting GDI, can be derived from assuming (for the moment) that the men's and women's

[3] The expression of findings in terms of substantive difference is considered best practice in the psychology literature (Wilkinson and Task Force on Statistical Inference, 1999), and--unlike in economics--the implications of various methodologies for the study of gender differences has also been the topic of intense and high-profile professional discussion (Archer, 1996; Eagly, 1995; Hyde and Plant, 1995; Martell et al., 1996 and other articles in the March 1995 and February 1996 issues of the *American Psychologist*.).

[4] This is most often estimated as:

$$s_p = \sqrt{\frac{(n_m - 1)s_m^2 + (n_f - 1)s_f^2}{n_m + n_f}}$$

where s_m, s_f, n_m and n_f are the standard deviations and sample sizes for the male and female samples. In Stata 12, this can be calculated using the user-written "cohend" command, using the result "corrected for uneven groups." In Stata 14, it and related statistics are directly available (Huber, 2013).

distributions are each normal. If *d*=3 and the distributions are normal, the probability that one could correctly guess whether a person was a male or female depending on their score would be 93%, and a random matching of a male and female would yield a higher score for the male 98% of the time (Coe, 2002). In this case, inter-group difference would be sizeable, relative to intra-group variation, and so there would be relatively little overlap between the distributions. When *d*=1*,* these probabilities drop to 69% (compared to 50% by chance alone) and 76%. The distributions would have a considerable range of overlap. When *d*=0.3*,* one would correctly guess the sex from the score only 56% of the time, and would find that the male scores higher in only 58% of random matches (Coe, 2002). On a rule-of-thumb basis, then, one might think of classifying a difference as GDI if it exhibits *Cohen's d* values of at least 3, if not higher.

While labeling a *Cohen's* d value as "large" or "small" can be a matter of interpretation and context, it would seem that a strong level of predictive ability on the part of any proxy for risk preference (such as sex) should be expected. The literature on risk implicitly assumes that one can accurately determine a subject's level of risk aversion using only a few survey questions, or a fairly brief laboratory experience. That being the case, it would clearly be very inefficient to, instead, judge the relative level of risk aversion of an individual based on a proxy that has only perhaps 76% or 58% accuracy.

The other additional statistics added in this replication and extension are conventional *t*-tests for the differences between means, and (when the necessary data are available) Wilcoxon Rank Sum tests for differences between medians. To ease exposition, they will both be presented in relation to *Cohen's d* values, although strictly speaking only the *t*-tests refer to means. All of the studies reviewed below involve experiments in which subjects decide how much ($*x*) of an endowment they will invest in a risky asset.

4.1 Dreber et al. (2010) and Dreber and Hoffman (2007)

CG begin by reviewing Dreber et al. (2010) and Dreber and Hoffman (2007). These represent two of the four cases for which CG present standard errors, standard deviations, and evidence about statistical significance. CG state that the gender differences in means and medians are statistically significant at the 1% level. *Cohen's d* can be computed as 0.94 for the former study, and 0.68 for the latter. While the differences in averages are statistically significant, the *Cohen's d* values indicate a GDA—not a GDI—structure.

Relevant for later discussion, is it worth noting that the Dreber and Hoffman study specifically focused on evidence for a biological gender difference (via measurement of digit ratios), possibly making gender a salient issue for the research subjects.

4.2 Charness and Gneezy (2010)

The next study CG review is Charness and Gneezy (2010). The averages and sample sizes reported in the first two columns of Table 1 replicate Table 1 in CG (2012, p. 52). Based on these means and their positive differences (in favor of men), CG apply a binomial test to the eight distinct treatments. They conclude that the "likelihood that either gender would invest more than the other in all eight treatments [if there were no underlying gender difference] is p=.0008, indicating a significant difference" (2012, pp. 52-3).

To investigate whether the data is more consistent with GDI or GDA, the raw data was requested and received from the authors, and a replication and extension were undertaken.[5] Within-sex standard deviations were computed and, using those, *Cohen's d* and test statistics for differences in means (t-test) and differences in medians (Wilcoxon Rank Sum Test). These are listed in bold typeface in Table 1. Intra-sex variability, it turns out, is far from trivial.

Table 1
Extended analysis of Charness and Gneezy (2010)

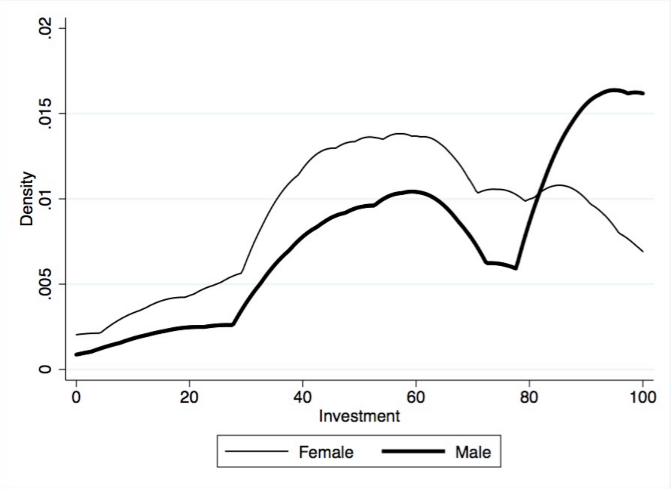
Treatment	Avg. male investment **(Std. Dev.)**	N	Avg. female investment **(Std. Dev.)**	N	**Effect Size (*Cohen's d*)**
Illusion – free choice	76.11 **(28.67)**	18	57.22 **(24.48)**	9	**.72**
Illusion – investor rolls	79.69 **(26.61)**	16	49.29 **(20.30)**	7	**1.27**$^{**,++}$
Illusion – experimenter rolls	71.20 **(30.45)**	20	69.83 **(21.28)**	6	**.05**
Illusion – costly choice	83.21 **(27.98)**	14	58.33 **(28.06)**	9	**.93**$^{++}$
All illusion choices	76.97 **(28.32)**	68	58.19 **(23.95)**	31	**.70**$^{***,+++}$
Ambiguity –free choice	75.26 **(22.45)**	19	61.43 **(33.38)**	7	**.56**
Ambiguity – known only	64.69 **(27.17)**	16	62.75 **(33.26)**	8	**.07**
Ambiguity – unknown only	70.81 **(67.50)**	16	67.50 **(22.52)**	10	**.14**
Ambiguity – costly choice	82.22 **(27.02)**	18	55.63 **(33.64)**	8	**.95**$^{**,++}$
All ambiguity choices	73.59 **(25.88)**	69	62.18 **(29.34)**	33	**.42**$^{**,++}$
All choices	75.27 **(27.07)**	137	60.25 **(26.73)**	64	**.56**$^{***,+++}$

Bold type indicates additions to the CG analysis.
*, **, and *** represent statistical significance of the difference between means (by *t*-test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test
$^{+}$, $^{++}$, and $^{+++}$ represent statistical significance of the difference between medians (by Wilcoxon Rank Sum Test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test

[5] Many very small differences between the many replications presented in the current study and those published in CG occurred, due to a factors such as slightly different number of observations in the data received or small errors in CG tables. Where these are immaterial to the argument, they are not discussed.

In only two out of the eight individual treatments is the difference in means statistically significant at the 5% or better; only three individual treatments out of eight yield statistically significantly different medians. While there is an overall statistically significant difference between means (and medians) across all treatments, there are not—as implied by CG's use of the binomial test--eight independent, statistically meaningful confirmations of larger male investment. The estimated size of the overall difference in means is a little more than half a standard deviation. None of the estimates of *Cohen's d* suggest a GDI structure.

The fact that there is considerable overlap between the distributions for men and women is confirmed by the kernel density plot of individual investments in Figure 1, drawn using the "all choices" data. While a larger proportion of men than women invest the highest possible amount ($100) in the risky asset, there are also women who invest the maximum. Similarly, while a higher concentration of women is found at very low levels of investment, some males invest at low levels as well. Both men's and women's distributions show some bunching up at the midway point of $50.



Note: The curves are obtained using a Epanechnikov kernel density smoother with a bandwidth of 10.

Fig. 1. Kernel densities of individual investments, Charness and Gneezy (2010) "all choices"

In summary, these data show that, for this experimental set-up and population of subjects, a statistically significant difference between male and female mean investments in the risky asset, of a little more than half a standard deviation in size, is found in a large sample. The difference, however, is not always statistically significant in small samples, and is not categorical.

4.3 Charness and Gneezy (2004)

Charness and Gneezy (2004) present subjects with the investment task in a "Natural" treatment and a "Frame" treatment, where the latter involves a small change in wording. CG (2012) report that the coefficient on gender in Tobit regressions is statistically significant with "p=0.006, one-tailed test" in the Natural condition and "p=0.036, one-tailed test" in the Frame condition.

These results were not precisely replicable, however. While the authors generously provided the data, they (Charness, personal correspondence) noted that "the paper [CG] contains an error. It had inadvertently combined data that were not relevant." Using the smaller sample of relevant data only, and including the same extended statistics as for the studies above, yields Table 2.

Table 2
Extended Analysis of Charness and Gneezy (2004)

Treatment	Avg. male investment **(Std. Dev.)**	N	Avg. female investment **(Std. Dev.)**	N	**Effect Size (*Cohen's d*)**
Natural	48.3 **(30.63)**	15	33.46 **(13.23)**	13	**.64**
Frame	89.51 **(18.70)**	14	73.57 **(28.11)**	14	**.69***
All choices	68.20 **(32.72)**	29	54.26 **(29.88)**	27	**.45**

Bold type indicates additions to the CG analysis.

*, **, and *** represent statistical significance of the difference between means (by *t*-test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test

Standard two-tailed tests of statistical significance are more appropriate than the one-tailed tests chosen by CG, since the one-tailed tests make a pre-judgment that women will not, on average, invest more in the risky asset than men. (As will be seen in later studies, this is not a safe assumption.) Using one-tailed tests inflates the inferred statistical significance by halving the p-value. Table 2, using the corrected data and two-tailed tests, shows that the difference in means is only statistically significant at the 10% level for the Frame treatment. Overall, and in the Natural treatment, the difference in means is not statistically significant. The difference in medians is not statistically significant for any case, using the Wilcoxon Rank Sum test.

While the values of *Cohen's d* for the differences by gender are all well under one standard deviation, it is interesting to note that a comparison of the frame to the natural condition (pooling data from both sexes) yields a considerably larger *Cohen's d* of 1.64. The difference between the natural and frame distributions is also judged statistically significant at the 1% level by both the *t* and Wilcoxon tests. The issue of the importance of framing will be returned to later.

In addition to summarizing this study in a table, CG (2012) also plot graphs for the Charness and Gneezy (2004) data which show the male value lying above the female value in each of the 10 periods occurring in each of the two treatments. They noted that "males consistently invest at a higher rate in each period" (53). Those lines represent, of course, *the means* for each gender. Figure 2(a) below recreates these as solid lines, for the natural treatment, using the corrected data. Adding 95% confidence bands for the means, by sex, in

Figure 2(b) however, makes it clear that the gender difference is not categorical: Due to considerable intra-sex variability, the confidence bands for the means show substantial overlap. The overlap of *individual* investments is, of course, even wider. The range of individual investments (over all periods) made by men in these data, in fact, runs from $0 to $100, which is also the range of individual investments made by women.

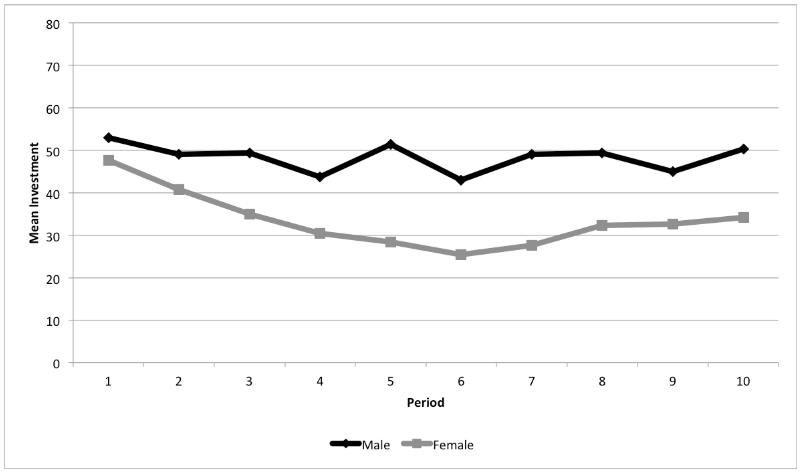


Fig. 2(a). Mean investments by sex, Charness and Gneezy (2004) "natural" treatment

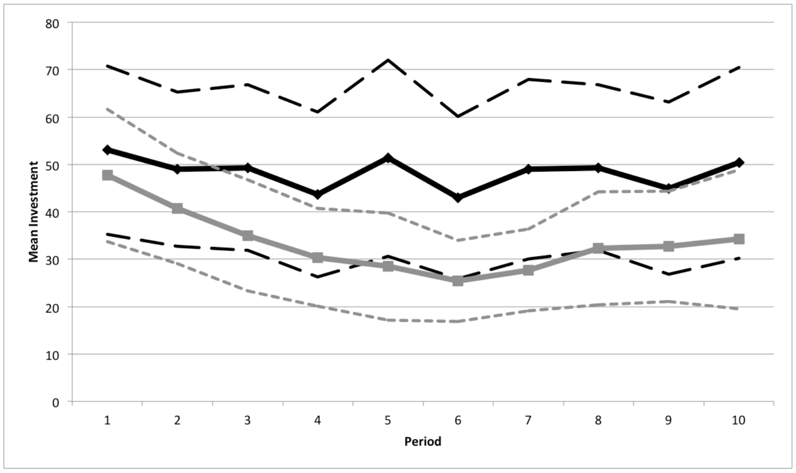


Fig. 2(b). Mean investments by sex with 95% confidence bands for the mean, Charness and Gneezy (2004) "natural" treatment

4.4 Yu (2006)

CG report that the six treatments (daily, intermediate, and weekly, each in both a laboratory and an internet setting) studied in Yu (2006) demonstrate a "substantial and consistent gender difference in investment choices" (55). They claim that "the binomial test gives p=0.031, two tailed test" (55) and report that Yu found a highly statistically significant gender dummy in an OLS regression. CG also include graphs corresponding to each individual treatment (56). Table 3 extends there analysis to include effect sizes and additional statistical significance tests.

Table 3
Extended Analysis of Yu (2006)

Treatment	Avg. male investment **(Std. Dev.)**	N	Avg. female investment **(Std. Dev.)**	N	**Effect Size (*Cohen's d*)**
Laboratory					
Daily	**52.69** **(26.56)**	19	37.26 **(19.25)**	9	**.65**
Intermediate	49.03 **(24.86)**	13	35.53 **(19.35)**	15	**.63**
Weekly	67.06 **(32.71)**	17	47.81 **(24.82)**	14	**.64***
All laboratory choices	56.70 **(28.91)**		40.46 **(23.31)**		**.62**$^{***,++}$
Internet					
Daily	61.52 **(29.16)**	22	38.07 **(24.26)**	16	**.89**$^{**,++}$
Intermediate	58.39 **(24.26)**	22	**34.70** **(27.28)**	11	**.67***
Weekly	72.23 **(31.00)**	19	45.81 **(29.30)**	16	**.90**$^{**,++}$
All Internet choices	63.66 **(33.83)**		**40.09** **(26.75)**		**.76**$^{***,+++}$

All choices	60.62 **(31.82)**	40.26 **(25.04)**	**.70**$^{***,+++}$

Bold type indicates additions to the Yu analysis.
*, **, and *** represent statistical significance of the difference between means (by *t*-test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test
$^{+}$, $^{++}$, and $^{+++}$ represent statistical significance of the difference between medians (by Wilcoxon Rank Sum Test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test

Two of the six treatments yield differences that are not statistically significant at any level, for either means or medians. Two more differences are statistically significant at only the 10% level, and then only for means, not medians. Only at the large-sample level do highly statistically significant differences (at the 1% level) appear. Substantively, all *Cohen's d* values indicate a GDA structure.

4.5 "Other" Studies

In Table 4 of their article, CG list results from ten “other” experimental studies. They note that “In 9 of the 10 experiments in Table 4 men invested more than females” and “apply the simple binomial test, where the null hypothesis is that there is no difference between the number of studies finding the males are more financially risk averse than females and the number of studies that show the opposite” (56). They conclude that, given the result of the binomial test, “these data support the consistent pattern of greater risk-taking by males in investment choices” (56).

Table 4 replicates and extends their results. All average investment levels are expressed as proportions of endowments. One of the “other” studies CG listed was Dreber and Hoffman (2007), already discussed earlier in their paper, and so only nine studies are listed here.

Table 4
Replication and Extension of "Other" Studies

Authors	Treatment	Avg. male investment (N) **(Std. dev.)**	N	Avg. female investment (N) **(Std. dev.)**	N	**Effect size (*Cohen's d*)**
(Langer and Weber, 2004)	All treatments	64.62 **(27.49)**	93	58.70 **(23.02)**	14	**.22**
(Haigh and List, 2005)		58.30 NA	50	55.59 NA	8	NA
(Fellner and Sutter, 2004)	All treatments	57.44 **(32.57)**	39	49.04 **(27.08)**	79	**.29**
(Bellemare et al., 2005)[a]		65.03 (95) **(25.06)**	95	61.36 **(26.72)**	40	**.14**
(Charness and Genicot, 2009)		59.22 **(24.72)**	41	52.23 **(22.48)**	53	**.30****
(Gneezy et al., 2009)[b]		68.70 **(32.32)**	**96**	69.29 **(30.62)**	91	**-.02**
(Ertac and Gurdal, 2012)	Positive expected return only	72.32 **(25.87)**	79	54.29 **(21.50)**	49	**.75**^{***,c}
(Gong and Yang, 2012)	Matrilineal society	53.9 **(35.4)**	31	32.5 **(29.1)**	36	**.68**^{***,c}
(Gong and Yang, 2012)	Patrilineal society	37.3 **(33.1)**	37	4.3 **(15.5)**	28	**1.24**^{***,c}

Bold type indicates additions to the CG analysis.
NA indicates "not available": The original authors were willing to share their data, but had not themselves compiled the data on gender. CG, who received original experimental sheets from Haigh and List declined to supply these data.
*, **, and *** represent statistical significance of the difference between means (by *t*-test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test
$^{+}$, $^{++}$, and $^{+++}$ represent statistical significance of the difference between medians (by Wilcoxon Rank Sum Test) at the 10%, 5% and 1% levels, respectively, in a two-tailed test
[a] Converted to express investments as a percentage of endowment.
[b] Neither the means nor sample sizes listed in CG could be replicated using the data supplied. The calculations shown here (based on the data supplied) are, however, consistent with tables published in Gneezy et al.
[c] Wilcoxon Rank Sum Test was not performed, because calculations were made from published summary statistics rather than raw data.

In only four out of the eight studies for which statistics are available, are the differences in means by gender statistically significant. In no cases where raw data made the test possible, were the differences in medians statistically significant as judged by the Wilcoxon Rank Sum test.

Additional light can be shed on the variability of gender differences by examining some of the different treatments used within some of these studies, although these were not explored in CG. In a note to their own table in regard to Ertac and Gurdal (2012), CG write that "We include only the individual risk decisions where there is a positive expected return from investing in the risky asset. A similar gender difference applies in the other cases." Examining at the negative expected return case in Ertac and Gurdal, however, reveals a considerably smaller difference (*d*=.17) than that shown in Table 4 for the positive return case, and one which is statistically insignificant. While the Fellner and Sutter results are not statistically significant when pooled, two of their four treatments result in statistically significant (at the 5% and 10% levels) higher average investments by males. Their other two treatments, however, result in higher average *female* investments, though the differences are not statistically significant. Two of the three treatments in Bellemare et al. give rise to higher male average investment (one of them significant at the 10% level), but the third results in a higher average female investment (not statistically significant).[6] Were binomial tests (invalidly) applied to point estimates for these cases with no regard to statistical significance, as they were (invalidly) applied to other treatments by CG, a somewhat different picture would have appeared.

5 Discussion

While the current study focuses on the CG analysis, that study can be put in the context of a larger research literature, and the implications of beliefs about gender and risk for real-world economic behavior and policy can be briefly discussed.

[6] These results are available from the author.

5.1 Confusions in the Literature

While the CG article particularly merits replication and extension due to the breadth of its claims, the high-profile nature of its publication, and the unique nature of its statistical methods, some degree of misunderstanding appears to be, in fact, rather widespread in the gender and risk literature.

Most other studies in the economics literature on gender and risk do, in contrast to CG, at least examine the statistical significance of any differences found. Attention to the substantive size of differences found, or to the degree of overlap between distributions, however, is still rare (Nelson, 2012). As a result, it is often not made clear whether the claims made refer to GDI or GDA, or whether they refer to a subpopulation or to all humans.

Statements such as "men have a higher *x* than women" are literally true only for GDI, and only true in the most general sense for those GDI that distinguish *all* human males and females. One problem arises when differences found within a particular subpopulation are inappropriately generalized (Henrich et al., 2010). Another, less recognized, problem is that statements of the form "men have a higher *x* than women" are, unfortunately, likely to be understood as confirming a GDI (categorical) difference (e.g., Khemlani et al., 2012), even if the phrase is intended only as a convenient shorthand for GDA ("our study found a statistically significant difference between the mean levels of *x* for men and women, with men having the higher average level").[7]

[7] Similar problems of interpretation may occur, of course, for aggregate differences found by other groupings, such as race, ethnicity, income level or region of residence, and these may also have significant real-world implications. Yet since the human propensity to stereotype by sex seems to be particularly strong (Prentice and Miller, 2006), investigating the case of gender is also of interest on its own.

To the extent that researchers, ourselves, do not carefully distinguish between GDI and GDA, we may not only convey a false impression to the public, but be prone to the same sorts of cognitive errors—such as stereotyping and confirmation bias—that we study in our research subjects. A prior belief that gender differences in risk aversion are GDI, for example, might help explain the methodological difficulties evident in CG. Nelson (2013) demonstrates how a number of other works in the economics literature exhibit evidence of confirmation bias, in the form of selectively publishing or highlighting only evidence that agrees with the expectation of a greater male preference for risk.

Research on gender and behavioral differences could be improved by an increased awareness of possible confirmation bias; by examining the substantive size of differences and the degree of overlap between distributions; and by care in making generalizations. In addition, the larger psychological literature suggests that, for experimental research in particular, issues of framing deserve careful treatment. Some psychological studies have manipulated the salience of gender in the framing of their experiments, and found that differences in risk aversion (measured by *Cohen's d*) grow larger in sex-stereotypical directions when gender stereotypes are evoked, and weaken considerably in size and statistical significance when they are not (Carr and Steele, 2010; Ronay and Kim, 2006; Weaver et al., 2012). For example, Carr and Steele (2010) evoke stereotype threat among one group of female subjects by presenting a lottery-type exercise as a "math problem" and asking about gender before it was administered; calling it a "puzzle" and asking about gender afterwards dramatically reduces the measured "gender differences." It may be that GDA in risk could be in substantial measure be *caused* by the social belief that risk-taking is a masculine activity. Such evidence should further discourage confusing GDA with GDI, or asserting biological difference, and encourage more careful attention to (perhaps unintentional) gender framing effects that may be present in the protocols for experimental studies.

5.2 Implications for Policy

Recent research in social psychology suggests that human brains are very prone to making sweeping and invalid generalizations based on gender, and to applying these generalizations to individuals—that is, to stereotype. Stereotyped judgments have been found even among scientists (Moss-Racusin et al., 2012), and to be expandable in their scope by experiences involving sample sizes as small as two (Prentice and Miller, 2006). Neurological research has found correlates of stereotyped beliefs in brain activity (Knutson et al., 2007).

To the extent that a high-profile claim such as that made in CG unfortunately contributes to and reinforces a now widespread stereotype of risk-taking as a masculine activity, it therefore is highly likely to distort our understanding of economic reality. It has been claimed, for example, that the "fact" that "women are more risk-averse than men" may contribute importantly to explaining labor market pay differentials or differences in entrepreneurship (Booth and Nolen, 2012; Hartog et al., 2002; Lindquist and Säve-Söderbergh, 2011), and prescribe categorically sex-segregating financial advising or employment in finance (Beckmann and Menkhoff, 2008; Olsen and Cox, 2001). Without that "fact," the validity of these explanations and prescriptions would be more deeply examined, and alternatives entertained more readily.

To the extent that research encourages a belief that risk-taking is GDI among the general public, it directly encourages invidious discrimination by creating false beliefs about the characteristics of individuals, based on their group membership. Women may, for example, be judged according to stereotype, and therefore systematically excluded from realms of employment and social interaction where risk-taking is valued. This runs counter to goals of efficient and equitable use of human resources in employment, and to national anti-discrimination policies.[8] To the extent that men may attempt to confirm their masculine identity (Akerlof and Kranton, 2000) by taking excessive risks, stereotyped expectations may additionally have very negative consequences in male-dominated areas such as large-scale investment banking.

6 Conclusion

Both in methodology and in presentation, Charness and Gneezy (2012) tend to treat gender differences in risk aversion as categorical, applying to males and females at the individual level. A re-examination of the data, however, reveals that the gender differences found (when they occur) are more correctly understood as applying to aggregates, such as means, and are modest in size. These on-average differences are not substantively large enough, relative to the amount of intra-sex variability, to justify making inferences about individuals. In a number of studies, the differences in average risk-aversion by gender are not even statistically significant, or are of the "wrong" sign. Evidence from psychology also suggests that aggregate differences, when found, may be attributable least in part to social stereotyping and framing effects present in the experiments themselves.

[8] While arguments may be made for cases of efficient "statistical discrimination" when underlying differences are large and are expensive to observe, this is not the case here, as noted earlier.

Greater attention to intra-sex variability and inter-sex overlap of distributions would improve the objectivity of economic research, and prevent economic research from contributing to social and economic problems of invidious stereotyping, discrimination, and excessive risk-taking.

Acknowledgements

I thank Gary Charness, Sabine Kröeger, Kenneth Leonard, and Frank Yu for graciously providing original datasets. Mayara Fontes, Sophia Makemson , and Matthew P.H. Taylor and provided expert research assistance.

References

Akerlof, George A. and Kranton, Rachel E., 2000. Economics And Identity. Quarterly Journal of Economics, 115(3), 715-753.

Archer, John, 1996. Comparing Women and Men: What is Being Compared and Why? American Psychologist, 51(2), 153-154.

Beckmann, Daniela and Menkhoff, Lukas, 2008. Will Women Be Women? Analyzing the Gender Difference among Financial Experts. Kyklos, 61(3), 364-384.

Bellemare, Charles, Krause, Michaela, Kröger, Sabine and Zhang, Chendi 2005. Myopic loss aversion: information feedback vs. investment flexibility. Economics Letters, 87, 319-324.

Booth, Alison L. and Nolen, Patrick, 2012. Gender Differences in Risk Behaviour: Does Nurture Matter? The Economic Journal, 122(558).

Byrnes, James P., Miller, David C. and Schafer, William D., 1999. Gender Differences in Risk Taking: A Meta-Analysis. Psychological Bulletin, 125(3), 367-383.

Carothers, Bobbi J. and Reis, Harry T., 2013. Men and Women are from Earth: Examining the Latent Structure of Gender. Journal of Personality and Social Psychology, 104(2), 385-407.

Carr, Priyanka B. and Steele, Claude M., 2010. Stereotype Threat Affects Financial Decision Making. Psychological Science, 21, 1411-1416.

Charness, Gary and Genicot, Garance, 2009. An experimental test of inequality and risk-sharing arrangements. Economic Journal, 119(April), 796-825.

Charness, Gary and Gneezy, Uri, 2010. Portfolio choice and risk attitudes: an experiment. Economic Inquiry, 48(1), 133-146.

Charness, Gary and Gneezy, Uri, 2012. Strong Evidence for Gender Differences in Risk Taking. Journal of Economic Behavior & Organization, 83(1), 50-58.

Charness, Gary and Gneezy, Uri 2004, 2004, Gender, Framing, and Investment, mimeo (citation from CG; work not located).

Coe, Robert, 2002, It's the Effect Size, Stupid: What effect size is and why it is important, Annual Conference of the British Educational Research Association, University of Exeter, England.

Cross, Catherine P., Copping, Lee T. and Campbell, Anne, 2011. Sex Differences in Impulsivity: A Meta-Analysis. Psychological Bulletin, 137(1), 97-130.

Dreber, Anna and Hoffman, Moshe, 2007, 2D:4D and Risk Aversion: Evidence that the Gender Gap in Preferences is Partly Biological, mimeo (citation taken from CG; work not located).

Dreber, Anna, Rand, David G., Garcia, Justin R., Wernerfelt, Nils, Lum, J. Koji et al., 2010, Dopamine and Risk Preferences in Different Domains, Working Paper Series. Harvard University, John F. Kennedy School of Government

Eagly, Alice H., 1995. The Science and Politics of Comparing Women and Men. American Psychologist, 50(3).

Ertac, Seda and Gurdal, Mehmet Y., 2012. Deciding to decide: gender, leadership and risk-taking in groups. Journal of Economic Behavior & Organization, 83(1), 24-30.

Fellner, Gerlinde and Sutter, Mattias, 2004, How to Overcome the Negative Effects of Myopic Loss Aversion – An Experimental Study, Mimeo, MPI Jena. (Citation from CG; work not obtained).

Fine, Cordelia 2010, Delusions of Gender: How Our Minds, Society, and Neurosexism Create Difference. W.W. Norton, NY.

Gneezy, Uri, Leonard, Kenneth L. and List, John A., 2009. Gender Differences in Competition: Evidence from a matrilineal and patriarchal society. Econometrica, 77(5), 1637-1664.

Gong, Binglin and Yang, Chun-Lei, 2012. Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi. Journal of Economic Behavior & Organization, 83(1), 59-65.

Haigh, Michael S. and List, John A., 2005. Do professional traders exhibit myopic loss aversion? an experimental analysis. Journal of Finance, 60(1), 523–534.

Hartog, Joop, Ferrer-i-Carbonell, Ada and Jonker, Nicole, 2002. Linking Measured Risk Aversion to Individual Characteristics. Kyklos, 55(1), 3-26.

Henrich, Joseph, Heine, Steven J. and Norenzayan, Ara, 2010. The weirdest people in the world? Behavioral and Brain Sciences, 33(2/3), 1-23.

Huber, Chuck, 2013, Measures of effect size in Stata 13, The Stata Blog. Stata Corporation.

Hyde, Janet Shibley, 2005. The Gender Similarities Hypothesis. American Psychologist 60(6), 581-592.

Hyde, Janet Shibley and Plant, Elizabeth Ashby, 1995. Magnitude of Psychological Gender Differences: Another Side to the Story. American Psychologist, 50(3), 159-161.

Jordan-Young, Rebecca M., 2010, Brain Storm: The Flaws in the Science of Sex Differences. Harvard University Press, Cambridge.

Khemlani, Sangeet, Leslie, Sarah-Jane and Glucksberg, Sam, 2012. Inferences about members of kinds: The generics hypothesis. Language and Cognitive Processes, 27(6), 887-900.

Knutson, Kristine M., Mah, Linda, Manly, Charlotte F. and Grafman, Jordan, 2007. Neural Correlates of Automatic Beliefs About Gender and Race. Human Brain Mapping, 28, 915-930.

Langer, Thomas and Weber, Martin, 2004, Does Binding or Feedback Influence Myopic Loss Aversion? An Experimental Analysis, CEPR Discussion Paper Series. Center for Economic Policy Research, London.

Lindquist, Gabriella Sjögren and Säve-Söderbergh, Jenny, 2011. “Girls will be Girls”, especially among Boys: Risk-taking in the “Daily Double” on Jeopardy. Economics Letters, 112(2), 158-160.

Martell, Richard F., Lane, David M. and Emrich, Cynthia, 1996. Male-Female Differences: A Computer Simulation. American Psychologist, 51(2), 157-158.

Moss-Racusin, Corinne A., Dovidiob, John F., Brescollc, Victoria L., Grahama, Mark J. and Handelsman, Jo, 2012. Science faculty’s subtle gender biases favor male students. PNAS (Proceedings of the National Academy of Sciences) Early Edition(September 17), 1-6.

Nelson, Julie A., 2012, Are Women Really More Risk-Averse than Men? , INET Resarch Note.

Nelson, Julie A., 2013, The Impact of Gender Stereotyping and Confirmation Bias on Economic Research: Theory and a Case Study, Working paper, based on "Are Women Really More Risk-Averse than Men?" INET Research Note #12, 2012.

Olsen, Robert A. and Cox, Constance 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(1), 29-36.

Prentice, Deborah A. and Miller, Dale T., 2006. Essentializing Differences Between Women and Men. Psychological Science, 17(2), 129-135.

Ronay, Richard and Kim, Do-Yeong, 2006. Gender differences in explicit and implicit risk attitudes: A socially facilitated phenomenon. British JOurnal of Social Psychology, 45(2), 397-419.

Weaver, Jonathan R., Vandello, Joseph A. and Bosson, Jennifer K., 2012. Intrepid, Imprudent, or Impetuous?: The Effects of Gender Threats on Men's Financial Decisions. Psychology of Men & Masculinity, Advance online publication., 1-12.

Wilkinson, Leland and Task Force on Statistical Inference, 1999. Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist, 54(8), 594-604 (591-524 in downloadable document).

Yu, F., 2006, Information Availability and Investment Behavior. University of Chicago. .