

Running head: GENDER DIFFERENCES

Large-Sample, Single Experiment Estimates of the Size of Gender
Differences on Visual Illusions, Maze Learning, and Mirror Drawing

Kenneth O. McGraw

University of Mississippi

Address correspondence to

Kenneth O. McGraw
Dept. of Psychology
University, MS 38677

662-915-5192 (phone)
pymcgraw@olemiss.edu (e-mail)

Abstract

Using data from PsychExperiments, a public data archive that has been accumulating data from a set of cognitive and perceptual experiments since 1998, this report provides large-sample evidence of gender differences on three visual illusions (Poggendorff, Mueller-Lyer, and Ponzo), a maze learning task, a dynamic line motion illusion, and a mirror drawing task. For comparison purposes, data from a three-dimensional mental rotation task are included. The large-sample, exhaustive datasets permit estimation of effect size in the absence of meta-analysis, which is subject to the file-drawer problem and which requires pooling data across methodologically different studies. The data show effects that are consistent with the well-established finding that females are slower and less accurate than males in performing a variety of “spatial” tasks. The novel contribution is that the tasks reported—except for mental rotation—are not among those previously known to produce gender effects.

Large-Sample, Single Experiment Estimates of the Size of Gender Differences on Visual Illusions, Maze Learning, and Mirror Drawing

Reviews of the vast literature on gender differences in cognitive abilities (Halpern, 2000; Linn & Petersen, 1985; Maccoby & Jacklin, 1974; Voyer, Voyer & Bryden, 1995) provide substantial evidence of reliable differences on a great variety of tasks. Some of the most cited are differences in the speed with which males and females complete mental rotation problems or extract forms from embedded figures and in the accuracy with which they perform spatial perception tasks (e.g., the Rod and Frame task and Piaget's water-level task). This report adds to the literature on gender differences by providing large-sample, single-task effect size estimates for a variety of tasks not widely known to produce gender differences. The tasks include a set of static, geometric visual illusions; a maze learning task; a dynamic, line motion illusion; and a mirror drawing task. For good measure, data are included from a large-sample replication of the task that has historically produced the largest gender difference in cognitive performance—a three-dimensional mental rotation task.

Data for the current report come from a publicly available data archive at PsychExperiments (<http://psychexps.olemiss.edu>), an on-line psychology laboratory used since 1998 as an instructional resource in college classrooms in the U.S. and abroad. The site features a variety of interactive experiments that can be conducted on-line by student participants, who then submit their data to the archive, where it is available for download and analysis. Typically students download their own data and that of their classmates, but the entire archive is available. Data for the current report are the full set of archived data through December, 2003. Sample sizes for the experiments are given in Table 1 for "full data sets" and "reduced data sets." The reduced sets were created by removing outliers, as described later.

This report is divided into multiple sections, with each section being a mini-report (introduction, method, and results) dealing with a single task or, in the case of the three classic visual illusions, a coherent set of tasks. A general discussion follows the task-specific sections.

Classic Visual Illusions

Not appearing on reviewers' lists of tasks known to produce reliable gender differences are visual illusions. The reason is simple. The most credible evidence of gender differences has appeared since the most recent review, which is Halpern's 2000 text.

In fact, prior to 2000, the most frequently cited study of gender differences in illusion perception--Porac, Coren, Girgus, and Verde (1979)--reported that there was no difference, at least on any the 13 common illusions investigated in that study. Holland, Wilson, and Goddard (1990) concurred with this view after investigating a single illusion--the Baldwin illusion. Dewar (1967) found only a slight, statistically non-significant advantage for men on the Mueller-Lyer illusion. Then Miller (1998; 1999) came across a gender difference on the Ponzo illusion, which he followed up in a report (Miller, 2001) specifically designed to investigate the gender difference. He reported effect sizes of $d=.50$ and $.46$ on two administrations of the Ponzo illusion, effect sizes that were slightly larger than the $.42$ obtained for an embedded figures task included in the experiment protocol. In an effort to explain why a gender difference of this magnitude would have gone undetected in earlier investigations, Miller (2001) distinguished between simple and complex presentations of the Ponzo, with simple presentations employing just two converging lines and complex ones employing more. Miller's version presented the two parallel line segments against a background of 11 converging lines. Thus Miller attributed his discovery of the gender difference to his use of this more compelling version of the illusion.

The present report follows up on Miller's evidence of a gender difference on the Ponzo illusion with a large-sample replication of this effect, along with evidence of gender effects on the Mueller-Lyer and Poggendorff illusions. There have been intimations in the literature that gender differences might be found on these other illusions. Previously mentioned was Dewar's (1967) suggestive, but non-significant, finding that males experience the Mueller-Lyer illusion to a lesser degree than females. Interestingly, Porac et al. (1979) also found some evidence of a gender difference on at least one version of the Mueller-Lyer illusion, though they discounted this finding as a probable Type I error. The same study also found a difference on the Poggendorff illusion, though this too was discounted, as it was but one of two significant effects found in a total of 45 gender comparisons on variants of the 13 illusions of interest. But in another study conducted since Halpern's most recent review (Halpern, 2000), DeClerck and DeBrabander (2002) reported a strong gender difference on the Poggendorff, which in retrospect adds credibility to the result from Porac et al.'s large-scale study. The present study serves to confirm that there are in fact substantial gender differences in susceptibility to the Ponzo, Poggendorff, and Mueller-Lyer illusions.

Method

Tasks

Poggendorff Illusion. Figure 1a is a screen shot showing the version of the Poggendorff illusion used at PsychExperiments. The experiment employs 48 presentations of this illusion in which the length of the diagonals and the separation of the verticals are varied. On each trial participants drag the upper right diagonal to the position they perceive to be collinear with the fixed lower left diagonal. The initial position of the upper right diagonal is randomly set to be both above and below the true collinear position of the lower left diagonal. The dependent

variable is the deviation in pixels between the height of a true collinear diagonal line segment and the height of the adjusted diagonal line segment (the right diagonal). A positive value indicates that the right diagonal was placed too low, which is the placement that reflects the Poggendorff illusion. For the purposes of this report, adjustment errors for the 48 trials were averaged to provide a single measure of susceptibility to the Poggendorff illusion.

Ponzo Illusion. Figure 1b shows the Ponzo illusion used at PsychExperiments. On each of five trials, the participant adjusts the bottom line segment to the perceived length of the upper line segment. The initial size of the lower line is randomly set to be either shorter or longer than the fixed length, 100-pixel upper line, as required by the method of adjustment. (Screen size is 640 x 480 for all experiments at PsychExperiments.) Participants who experience the illusion adjust the lower line to be longer than the upper line, so a positive difference (lower minus upper) is evidence of the illusion. Randomly interspersed with the five illusion trials are five control trials in which the converging lines are removed from the display, but these data were ignored for the purposes of this report. An average of the adjustment errors on the five illusion trials was used as the measure of susceptibility to the Ponzo illusion.

Mueller-Lyer Illusion. The Mueller-Lyer task at PsychExperiments consists of just half of the famous illusion, as shown in Figure 1c. In the experiment, there is a single illusory display—the line with fins attached. Research participants adjust the red line segment on the left to be equal to the red, vertical line segment in the illusory figure on the right. Because the experiment was designed to give students the opportunity to determine the effect of fin angle on perceived line length, the angle of the fins in the illusory figure is one of 11 possible angles ranging in 15° steps from 15° (an extreme “fins out” display) to 165° (in an extreme “fins in” arrangement), making for 11 different experimental conditions. Each condition is repeated four

times by using four blocks of 11 trials. The angle on any one trial within a block is sampled randomly without replacement from the set of 11.

To simplify the data analysis for this report, data from just the four 45° (fins out) trials and the four 135° (fins in) were used. On these and all other trials, the red line of the illusory figure has a random length of between 100 and 150 pixels in the 640 x 480 pixel screen. The adjustable line is randomly set to either 90 or 160 pixels at the start of each trial. A positive difference (adjustable line length minus target line length) indicates the standard illusory effect for “fins out” stimuli, which is that the red line segment appears longer than it in fact is. A negative difference indicates the standard illusory effect for fins-in stimuli. Data for this report are the sum of the absolute errors on the two stimulus types--45° (fins out) trials and 135° (fins in) trials. They were summed in order to estimate from the single-stimulus, PsychExperiments data the illusion that would be present in the traditional dual-stimulus presentations of the Mueller-Lyer illusion. (Traditionally, a fins-out figure of fixed angle is presented alongside a fins-in figure using the same angle, thus creating a situation where the two distorting effects on perception work simultaneously.)

On-Line Demonstrations. The three tasks described above, along with those to be described later, are available in demonstration form at <http://psychexps.olemiss.edu/exps/demos>. To experience the complete experiments, readers should choose the experiment by name from the list at <http://psychexps.olemiss.edu/Exps/labexperiments.htm>. Because the experiments were created in Authorware, the Authorware web player needs to be installed in one's browser to view the experiments and demonstrations. It is available free from Macromedia.com using the link at <http://psychexps.olemiss.edu/need.htm>. Methodological details in addition to those given above

are available from experiment links on the Instructor's Only page, <http://psychexps.olemiss.edu/Instructors/instructors.htm>.

Data Screening. In developing a public website to receive data from remote participants, the developers must decide whether to write all the data to the database or to screen it for outliers before recording the data. PsychExperiments was built on the model that data screening should be done by the data users rather than the data collectors; therefore, all data that participants elect to send is accepted without screening into the database. For the present analysis, data for the Ponzo, Mueller-Lyer, and Poggendorff experiments were screened for values that met Tukey's (1977) statistical criterion for an outlier; that is, a value more than 1.5 times the interquartile range greater than the 75th percentile or more that 1.5 times the interquartile range less than the 25th percentile. Applying this screening is prudent for the data from the illusion experiments because it is possible to generate data for these experiments by clicking through the trials rapidly without actually making line adjustments. Since 2000, PsychExperiments has added a time-in-experiment data value that is useful for screening participants who complete trials improbably quickly or slowly, but this value was not available for all of the data used in the present analysis. For this reason, a statistical definition of an outlier was adopted.

Results

The analysis was conducted both on the full and reduced datasets for each illusion, as summarized in Table 1, which gives sample sizes and effect size measures for the two datasets. In addition to the effect size measure d (the mean gender difference in pooled standard deviation units), effects are expressed using McGraw and Wong's (1992) common language statistic CL , which for the present data is the probability that a randomly sampled female will be more strongly affected by the illusion than a randomly sampled male. For example, in the case of the

reduced dataset for the Poggendorff, the value of .658 indicates that in head-to-head comparisons of males to females, females would be more affected with probability of .658 (i.e., in nearly 66% of the pairings, the female would be more affected than the male). CL can be converted into an odds ratio, which in this case is $.658/.342$ or 1.93:1 that a female will be more affected than a male.

As is evident in Table 1, the largest gender difference was for the Poggendorff illusion, followed by the Ponzo 1.48:1, and the Mueller-Lyer, 1.39:1. The values of d from Table 1 for the reduced datasets are reiterated in Figure 2, which is a set of box plots that contrast key features of the distributions of illusion effects observed for females and males in the reduced sets. Plots for the full datasets would look nearly the same at the mean and median. Plots for the full datasets give good evidence that illusion effects are distributed normally around their means, because medians and means overlap and kurtosis values are close to normal kurtosis.

Maze Learning

The data on gender differences in maze learning in a variety of mammalian species is extensive (see references in Moffat, Hampton, and Hatzipantelis, 1998, p. 74), but in humans, evidence of a gender difference derives from two relatively recent, small-scale studies. Astur, Ortiz, and Sutherland (1998) used a computerized version of the Morris water task to show a difference between samples of 20 males and 20 females. Moffat et al., using samples about twice as large, showed that males were both faster and less error prone in learning routes through three-dimensional computer-presented mazes. Maccoby and Jacklin (1974, p. 93) cited one study of maze performance that produced evidence of a gender difference, but the

“perceptual” maze used in this research was more of a mathematics problem than a route learning task.

Earlier studies using finger mazes, which provide a good human analog to the maze learning tasks used with small mammals, have failed to indicate a male advantage. Alvis, Ward, and Dodson (1989) report a specific finding of no difference, and Biersner (1980) reports no difference when mazes are completed using the left hand. When completing mazes with the right hand, females were in fact faster than males in Biersner’s study, but both of these studies using finger mazes had small sample sizes (male and female $Ns < 30$).

The null results of early studies combined with the paucity and relative recency of studies showing gender differences on credible analogs to maze learning by animals serve to explain why the gender difference in maze learning has not yet appeared on the lists of tasks that produce reliable gender differences. However, a maze learning task at PsychExperiments, similar to the one created by Moffat et al. (1998), shows that maze learning is indeed a cognitive task that produces gender differences.

Method

The maze learning task at PsychExperiments employs a 16-chamber maze along with a start box, and a finish box. An overhead view of the maze is given in Figure 3a. The task for participants is to navigate their way through the maze. To do so, they are shown a straight ahead view from their current position in the maze. The view shows the open paths that can be chosen. Directional arrows are provided for choosing a path. For example, Figure 3b shows the view from the start box. Note that there are just two directional options active from the start box. One can go straight ahead or to the right. On subsequent choices, only viable options are

active. That is, if there is no open path to the right, the right choice is deactivated. When participants reach a dead end, the only active choice will be a backward movement.

While the maze task at PsychExperiments collects data on errors and time, just the time data were used for this report. Time measurement begins when participants make a choice that results in their leaving the start box and ends when they successfully enter the finish box. If the participants return to the start box before entering the finish box, the timing is not restarted. The task requires that participants complete the maze 15 times. Data used for this report were averages for the 15 runs through the maze.

Results

Figure 4 is a box plot that shows males were consistently faster on average than females to complete the maze. The difference was 1.73 sec, or 16% faster. The faster average was due to consistently faster performance on each of the 15 trials. As in the Moffat et al. study, therefore, the male-female difference was consistent across trials (i.e., no trials by gender interaction). Figure 4 presents data from which outliers were removed. Given the positive skew in the distributions of both male and female performance times, there were no fast times identified as outliers, only slow times. The effect of eliminating these was to reduce the overall variability in performance and, thereby, increase the effect size estimate, which is shown in Table 1 to be .438. This effect implies that males will be faster than females in 62 random pairings out of 100, which translates to an odds ratio of 1.645:1.

Line Motion Illusion

In addition to gender differences in judgments about line lengths and alignments in static illusory displays, research has shown differences in dynamic spatial perception. Halpern (2000,

p. 101) refers to the ability tapped by these tasks as spatiotemporal ability. Examples are tasks that require estimating the time of arrival of objects that are viewed in motion before being occluded and tasks requiring the judgment of coincidence. The task used in the present research is a somewhat novel measure of spatiotemporal ability. It involves the perception—or, rather, the misperception—of line motion, a phenomenon first discovered by Hikosaka, Miyauchi, and Shimojo (1992). Illusory perception of line motion is created by preceding a line display with an attentional cue that is at one end of the line. The line in fact appears as a single object, instantaneously on the screen, but when the line is preceded by a brief attention grabbing cue at one of its ends, the line is perceived to grow from that point, thus creating the illusion of left to right or right to left motion.

Method

A dynamic version of the line motion illusion is available in demonstration form at <http://psychexps.olemiss.edu/Exps/demo.htm>. This demonstration mixes trials in which the line appears simultaneously on the screen with times when the line is moved—at varying speeds—onto the screen. Research participants are asked to identify the motion as “Right to left,” “Left to right,” or “Simultaneous,” with the last choice indicating that there was no perceived motion. The purpose of the manipulation is to determine a time parameter on the illusion; namely, how strong is the illusion in the face of actual contrary movement? When lines move right to left, for example, how slow can they move and still be misperceived as right to left movement when preceded by an attentional cue on the right side?

For the purposes of this report, only data from trials when lines were presented simultaneously on the screen were used. This is a small portion of the total, but the non-moving, simultaneous lines are the ones that create the strongest demonstration of the illusion. The

question, therefore, was whether males and females would differ in susceptibility to the perception of motion in its absence. The data were the proportion of correct simultaneous judgments that were made on trials when the line appeared without motion.

Results

The data showed that females experienced the illusion of motion in its absence more often than males, with females perceiving motion on 88.7% of the non-motion trials and males, on 82.7% of the non-motion trials. Because the proportion data were naturally constrained to the range 0.00 to 1.00, there was no need to consider outliers. As shown in Table 1, the effect size was a relatively small .388. The corresponding CL value of .595 translates to an estimated 595:405 (1.47:1) male advantage in one-to-one pairings. The box plot in Figure 5 gives a more complete picture of the differences found between the distributions of proportion data for males and females.

Mirror Drawing

Having subjects trace outlines or perform connect-the-dot exercises while viewing their drawing hand in a mirror was a popular laboratory means for studying trial-and-error learning and learning transfer in the heyday of learning studies, beginning in the early 1900's (Carmichael, 1927; Dearborn, 1910). Gender differences on mirror drawing tasks were investigated in the early part of the last century, but the lack of consensus in the findings is striking. There are indications that females outperform males (Yoakum & Calfee, 1913), that males outperform females (Balinsky & Stone, 1940), that there is a shift in adolescence from male to female superiority (Clinton, 1930), and that there is no gender difference at all (Reynolds & Stacey, 1955). There has been no effort in the literature to reconcile the different findings,

and the lack of consensus in the few studies to address the issue may be the reason that reviewers beginning with Maccoby and Jacklin (1974) have ignored this literature. The mirror drawing experiment at PsychExperiments provides an opportunity to re-address the question of gender differences on a perceptual motor task where visual and proprioceptive cues are reversed.

Method

The mirror-drawing task at PsychExperiments simulates actual mirror drawing by reversing mouse movements. Upward movements of the mouse cause the cursor to move down, left movements cause it to move right, and so on. Using the mouse, research participants trace the outline of a five-pointed star, first with one hand and then with the other. Progress is marked by a yellow highlighter that indicates the part of the star that has been successfully traced. Participants continue until the highlighter reaches the start point. When the cursor moves off the star's outline, participants must return it to the departure point before continuing. This means that the star must be traced continuously. The performance measure is the time taken to completely trace the star. There are separate completion time measurements for the two hands, but for the purpose of this analysis, average times for the two hands were used.

Results

The mean tracing time for females was about half again as long as the mean tracing time for males in both the full and reduced datasets. The absolute difference was better than 20 seconds. The box plot in Figure 6 provides a graphical depiction of the rather dramatic differences in the distribution of completion times. The effect size was .65 or about 2/3 of a standard deviation, which translates into a 2:1 odds ratio that a randomly chosen male will complete the mirror drawing task faster than a randomly chosen female.

Mental Rotation

For comparison purposes, this report concludes with data from the mental rotation task at PsychExperiments. Mental rotation tasks, which measure people's speed and accuracy in visualizing how objects will appear when rotated in two or three dimensions, provide a measure of what is generally referred to as spatial reasoning ability (Casey, Nutall, Pezaris, & Benbow, 1995). This is a core cognitive ability, one on which males and females have been known to differ for some time. The most frequently used measure of mental rotation is the Vandenberg and Kuse (1978) test, which uses three-dimensional objects and a multiple-choice format. This test produces gender effect estimates that are two-thirds of a standard deviation ($d=.66$) according to Voyer et al.'s (1995) meta-analysis. Another category of test examined in that analysis, referred to as "Generic mental rotation" consisted of computer or slide presentations of three-dimensional stimulus pairs like those first used by Shepard and Metzler (1971). These tests, which correspond to the mental rotation test at PsychExperiments, produced a gender effect of .37. A weighted average of the effects in the methodologically diverse literature on mental rotation produced a gender effect estimate of .56.

Method

The mental rotation task at PsychExperiments requires research participants to make judgments about three-dimensional objects patterned after ones first used by Shepard and Metzler (1971). The objects, presented side-by-side, are either the same or mirror image reversals of each other. In each pair, the object on the left is presented in an upright position and the object on the right is rotated. The research participant's job is to determine as rapidly as possible whether the figures are the same (i.e., a copy that differs only in rotation angle) or different (i.e., mirror image objects). Responses are judged for correctness and timed.

On each trial, a different one of 16 possible object pairs is presented in random order. Eight are identical object pairs and eight are mirror image pairs. The eight in each set are distinguished by their angular discrepancy, which goes from 0° to 315° in 45° steps. In the original 1998 implementation of this experiment, participants could choose to conduct 16, 32, 48, or 64 trials. In September 2002, the experiment was changed to require 48 trials. The change from an optional number of trials to a fixed number produced no change in mean response times, which were identical at 3.35 seconds pre and 3.35 seconds post (full data set). The data for this analysis were the mean decision times across all trials.

Results

Figure 7 is a box plot that shows the differences in the distributions of response times for males and females. On average, females were slower by half a second, which is an increase of 18% over the mean male response time of 2.8 sec. The effect size estimate from the full datasets and reduced datasets were about the same--.474 versus .451 (see Table 1), which makes the odds just under 2:1 that males will be faster than females in one-on-one competitions.

General Discussion

Because such large samples were used for the present analyses of gender differences, there can be no question about the existence of gender differences in the population on the tasks reviewed here, with the population being defined primarily as undergraduate psychology students in the United States and Canada. This characterization derives from the fact that 71% of the data contributed to the site comes with a psychology classroom designation (e.g., U. of Mississippi, Brown's Psy 214) rather than the default entry "Interested person" (which is an entry that may also be used by many psychology students). The association of the data with psychology students is also evident in the predominance (2.04:1) of female data in the database.

Given that there can be no doubt about the existence of gender effects on the tasks in question, the issue of interest is the size of the effects, which is why estimates were created using both Cohen's distance measure and McGraw and Wong's (1992) probability measure. Ordered by magnitude, the task producing the largest effect was mirror drawing, followed by the Poggendorff illusion, maze learning, the line motion illusion, the Ponzo illusion, and the Mueller-Lyer illusion. The set of effect size estimates varied from just over $1/3$ to about $2/3$ of a standard deviation, which implies effects that are moderate to moderately large using the effect size anchor points suggested by Cohen (1977). Another anchor point is provided by the gender effect on PsychExperiment's mental rotation task, which is just under $1/2$ a standard deviation. Mental rotation is typically taken to be the task that produces the largest gender difference in cognitive performance, so relative to this standard, the tasks reported here produce effects that range from slightly smaller to slightly larger than mental rotation.

Although maze learning is not a task that has often been shown to produce reliable gender differences, it is no surprise to learn that it does. Maze learning is quite clearly a "spatial" task, requiring the use of directional memory and, perhaps, a cognitive map. Probably the only reason such tasks are not already on the list of tasks known to produce reliable differences is that creating mazes for humans was challenging prior to the widespread use of computers. The most frequently used simulations were finger mazes. Moffat et al. found that the best predictor of maze learning ability was mental rotation speed. In that mental rotation is the primary example of a task that produces male superiority, it is again no surprise that maze learning joins the list.

Given the effect size estimates reported here for classic visual illusions (.35-.65), one can wonder why these gender differences in illusion perception have gone largely unreported. One

reason may be that samples have not been large enough to provide adequate power. To detect the gender difference in the Mueller-Lyer illusion with power of .80 when alpha is set at .05 requires a sample size of 90, the Ponzo—at least the version used at PsychExperiments, which has five converging lines, 66. Only the Poggendorff effect would be reliably detected in the traditionally small sample research used in psychology experiments, because an N of 26 would give power of .80 to detect this effect.

Line motion provides an intriguing new entry on the list of tasks that produce gender differences. There are some recent studies showing differences in what Halpern (2000) calls spatiotemporal ability, with Schiff and Oldak's (1990) time-of-arrival task being a prime example. It seems likely that line motion tasks tap spatiotemporal ability, though factor analytic studies would be needed to know for sure.

With regard to mirror drawing, the current data serve to resolve the “Is there or Isn't there a gender difference” question that has been left open by the literature; however, as with line motion, it is not exactly clear how the cognitive demands of this task align with those from other tasks known to produce gender differences. Successful mirror drawing requires inhibition of dominant response tendencies, which does not seem to be an ability tapped by other tasks that produce gender differences; but, at the same time, mirror drawing can be conceptually cast as a “visuo-spatial” task. As with line motion, where this task fits in the sets of tasks that assess a general cognitive ability on which the genders differ will have to be determined by factor analytic studies.

In evaluating the current effect size estimates, it is important to keep in mind that they should not be expected to overlap exactly with effect size measures that are obtained from meta-analyses. On the one hand, meta-analyses are forced to combine studies using different methods.

If different methods produce different effects, the average effect will be larger or smaller depending upon the proportion of studies using the more “effective” method. In conducting meta-analyses, researchers employ procedures that screen for heterogeneity in effect size estimates, but the differences must be large for an effect to be eliminated. Small variations in effect created by methodological variations are tolerated.

Another factor to consider when comparing effect size estimates obtained from meta-analysis with those obtained from single experiments is the file drawer factor. Meta-analysis effect size estimates are invariably biased by the file drawer problem; single experiments are not. All else being equal, this would serve to make large-sample, single experiment estimates less than meta-analysis estimates in any case where the file drawer problem is substantial.

References

- Alvis, G. R., Ward, J. P., & Dodson, D. L. (1989). Equivalence of male and female performance on a tactual spatial maze. *Bulletin of the Psychonomic Society, 27*, 29-30.
- Astur, R. S., Ortiz, M. L.; Sutherland, R.J. (1998). A characterization of performance by men and women in a virtual Morris water task: A large and reliable sex difference. *Behavioural Brain Research, 93*, 185-190.
- Balinsky, B. & Stone, I. R. (1940). High school norms for the mirror-drawing test of the six-pointed star. *Journal of Genetic Psychology, 56*, 207-210
- Biersner, R. J. (1980). Sex differences in right- and left-hand tactuomotor acquisition practice. *Perceptual & Motor Skills, 50*, 986.
- Carmichael, L. (1927). The history of mirror drawing as a laboratory method. *Pedagogical Seminary, 34*, 90-91. Casey, Nutall, Pezaris, & Benbow, 1995.
- Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender difference in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697-705.
- Clinton, R.J. (1927). Nature of mirror-drawing ability: norms on mirror-drawing for white children by age and sex. *Journal of Educational Psychology, 21*, 221-228.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J. : L. Erlbaum Associates.
- Dearborn, W.F. (1910). Experiments in learning. *Journal of Educational Psychology, 1*, 373-388.
- Declerck, C. & De Brabander, B. (2002). Sex differences in susceptibility to the Poggendorff illusion. *Perceptual and Motor Skills, 94*, 3-8.

- Dewar, R.E. (1967). Sex differences in the magnitude and practice decrement of the Muller-Lyer illusion. *Psychonomic Science*, 9, 345-346.
- Halpern, D. F. (2000). *Sex differences in cognitive abilities*, 3rd ed. Mahwah, N.J. : L. Erlbaum Associates.
- Hikosaka, O., Miyauchi, S. & Shimojo, S. (1992). Focal visual attention produces illusory temporal order and motion sensation. *Vision Research*, 33, 1219-1240.
- Holland, M., Wilson, A.E., & Goddard, M. (1990). Lack of sex difference with the Baldwin illusion. *Perceptual and Motor Skills*, 71, 305-306.
- Linn, M.C. & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability. *Child Development*, 56, 1479-1498.
- Maccoby, E.E. & Jacklin, C.N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- McGraw, K.O. & Wong, S.P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.
- Miller, R. J. (1998). Depth cue orientation and perceived depth in pictures. *Visual Arts Research*, 24, 80-90.
- Miller, R. J. (1999). The cumulative influence of depth and flatness information on the perception of size in pictorial representations. *Empirical Studies of the Arts*, 17, 37-57.
- Miller, R.J. (2001). Gender differences in illusion response: The influence of spatial strategy and sex ratio. *Sex Roles*, 44, 209-225.
- Moffat, S.D., Hampson, E., & Hatzipantelis, M. (1998). Navigation in a "virtual" maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior*, 19, 73-87.

- Porac, C., Coren, S., Girgus, J.S., & Verde, M. (1979). Visual-geometric illusions: Unisex phenomena. *Perception*, 8, 401-412.
- Reynolds, W.F. & Stacey, C.L. (1955). A comparison of normals and subnormals in mirror drawing. *Journal of Genetic Psychology*, 87, 301-308
- Schiff, W. & Oldak, R. (1990). Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception & Performance*, 16, 303-316.
- Shepard, R.N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Vandenberg, S.G. & Kuse, A.R. (1978). Mental rotation, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47, 599-604.
- Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117, 250-270.
- Yoakum, C. S. & Calfee, M. (1913). An analysis of the mirror drawing experiment. *Journal of Educational Psychology*, 4, 283-292.

Table 1

Sample Size and Gender Effect Size in Original and Reduced Datasets for Experiments

Experiment		Full datasets			Reduced datasets		
		N	d	CL ^a	N	d	CL ^a
Mueller-Lyer	Female	1475			1425		
	Male	719	.206	.558	691	.293	.582
Poggendorff	Female	991			980		
	Male	520	.496	.634	510	.575	.658
Ponzo	Female	1561			1546		
	Male	922	.324	.590	913	.343	.596
Maze	Female	699			661		
	Male	412	.235	.566	386	.438	.622
Mirror Drawing	Female	2124			2041		
	Male	874	.54	.649	843	.653	.678
Line Motion	Female	296			296 ^b		
	Male	155	.388	.595	155 ^b	.388	.595
Mental Rotation	Female	4099			3924		
	Male	1828	.47	.630	1734	.451	.625

^aCL estimates the probability that a randomly sampled male will be more accurate than a randomly sampled female on the illusions tasks and faster on the Mirror Drawing, Maze, and Mental Rotation tasks (see McGraw & Wong, 1992, for details).

^bNo outliers were removed because the data for this experiment—proportion correct—were constrained by the natural limits of 0.00 and 1.00.

Figure Captions

Figure 1. Screen shots showing the portrayals of the Poggendorff, Ponzo, and Mueller Lyer illusions at PsychExperiments.

Figure 2. Box plots showing gender effects on the magnitude of the Poggendorff, Ponzo, and Mueller-Lyer illusions. Dotted lines represent the means and solid lines represent the quartiles of the distributions.

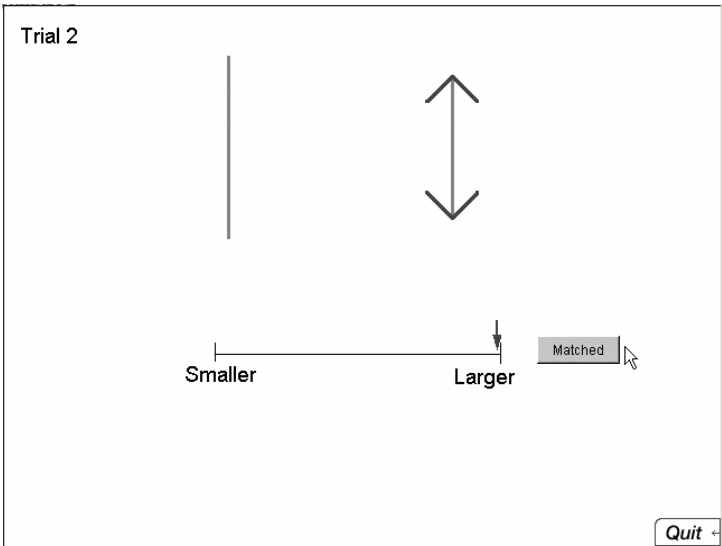
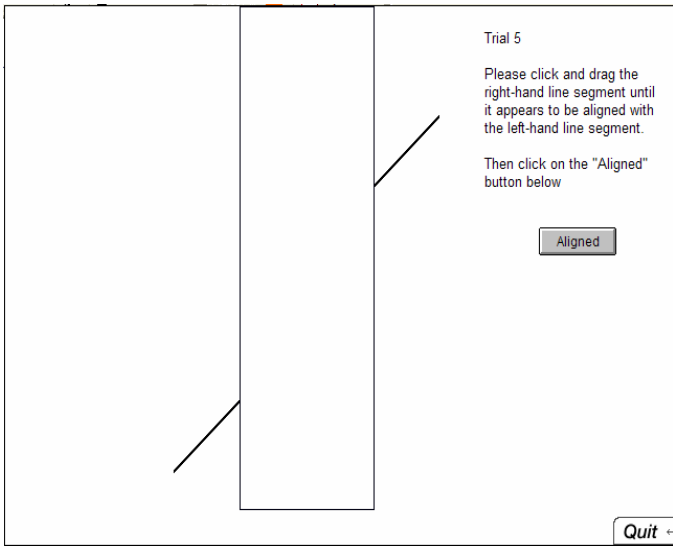
Figure 3. Map of the 16 chamber maze used at PsychExperiments and a screen shot showing the user's view from the start box of the maze. Active arrows show directional choices, which are selected by clicking on an arrow. Clicking an arrow moves the user into a new chamber where new directional choices appear.

Figure 4. Box plot showing gender effect on maze completion times.

Figure 5. Box plot showing gender effect on proportion of correct judgments on simultaneous trials in the line motion experiment.

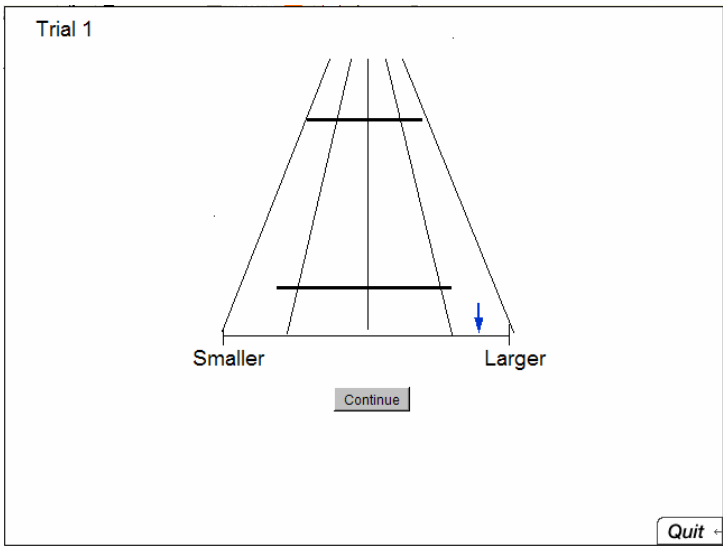
Figure 6. Box plot of mean time to completely trace a five-pointed star in the mirror drawing experiment.

Figure 7. Box plot showing the gender effect on response time in the mental rotation experiment.



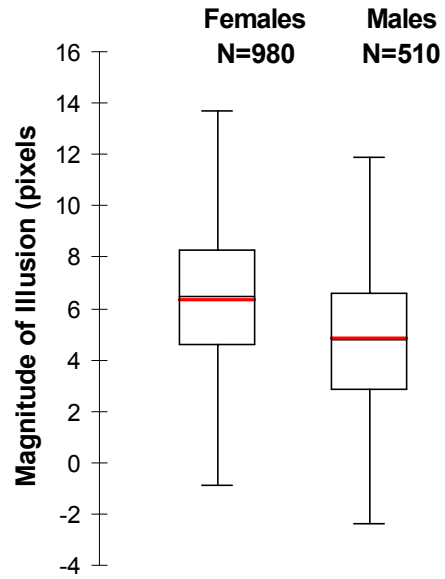
(a) A trial from the Poggendorff task

(b) A trial from the Mueller-Lyer task

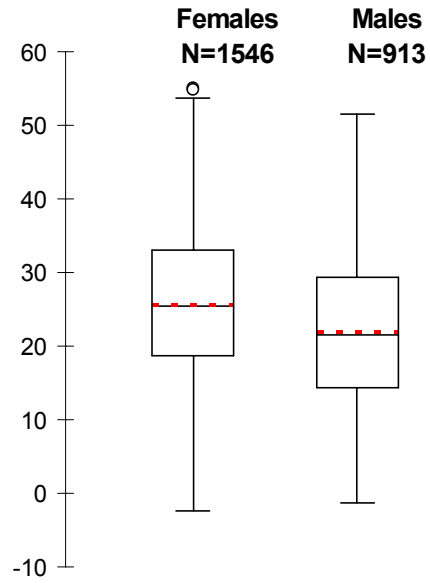


(c) A trial from the Ponzo task

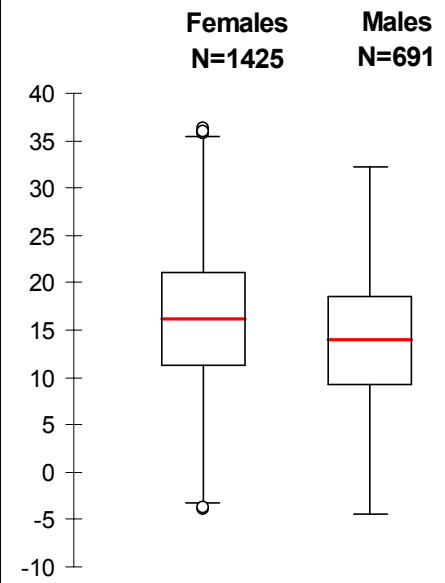
Poggendorff Illusion (d=.575)

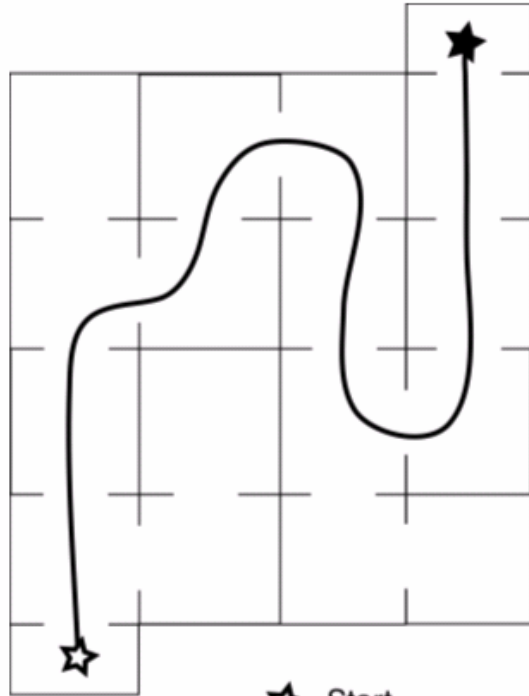


Ponzo Illusion (d=.343)



Mueller-Lyer Illusion (d=.293)

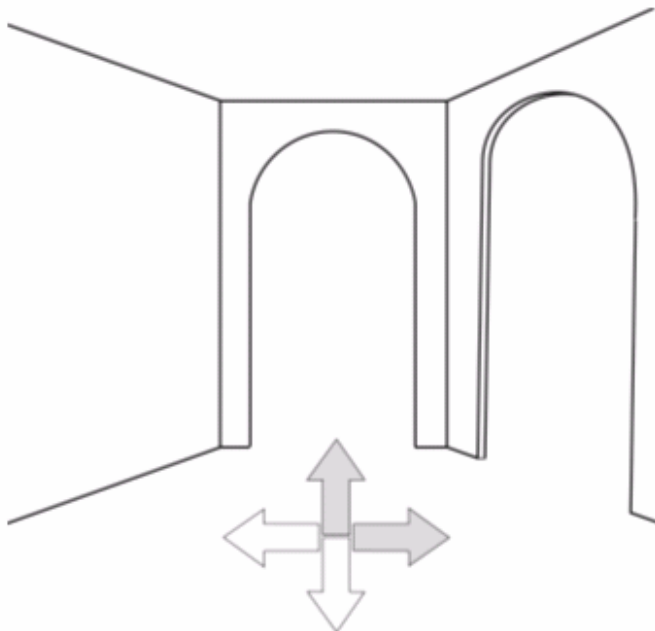




☆ Start

★ Finish

Maze Layout and Solution



View from Start Box

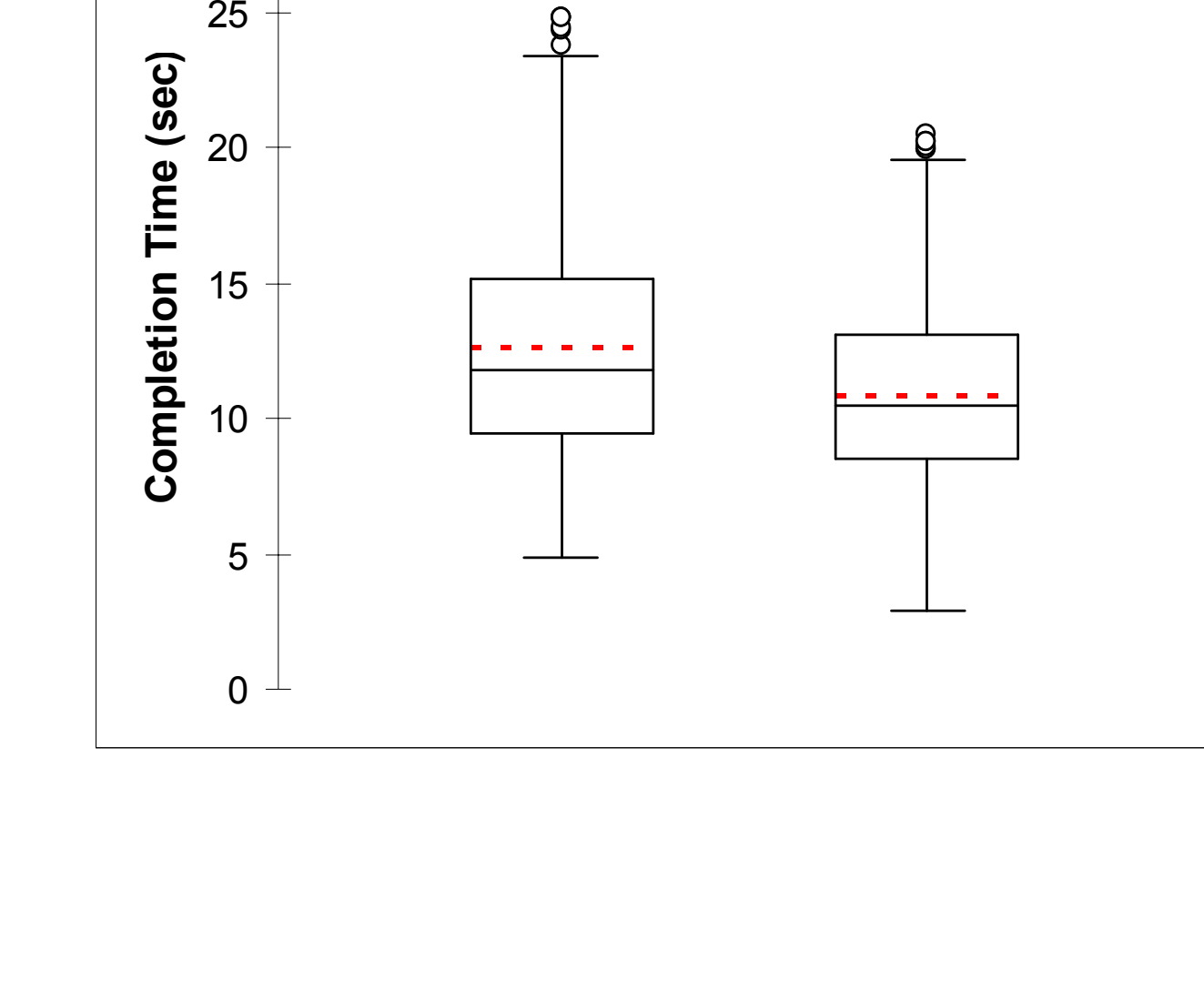
Maze Learning (d=.438)

Females
N=661

Males
N=386

Completion Time (sec)

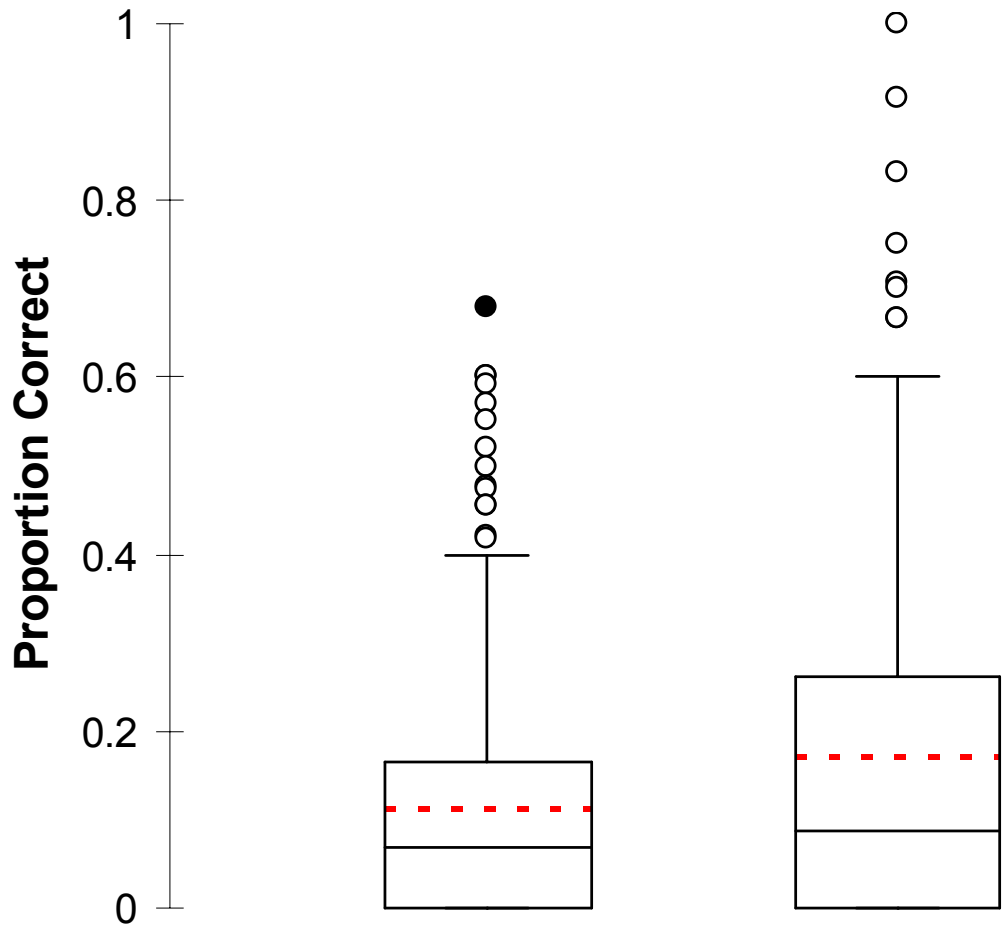
30
25
20
15
10
5
0



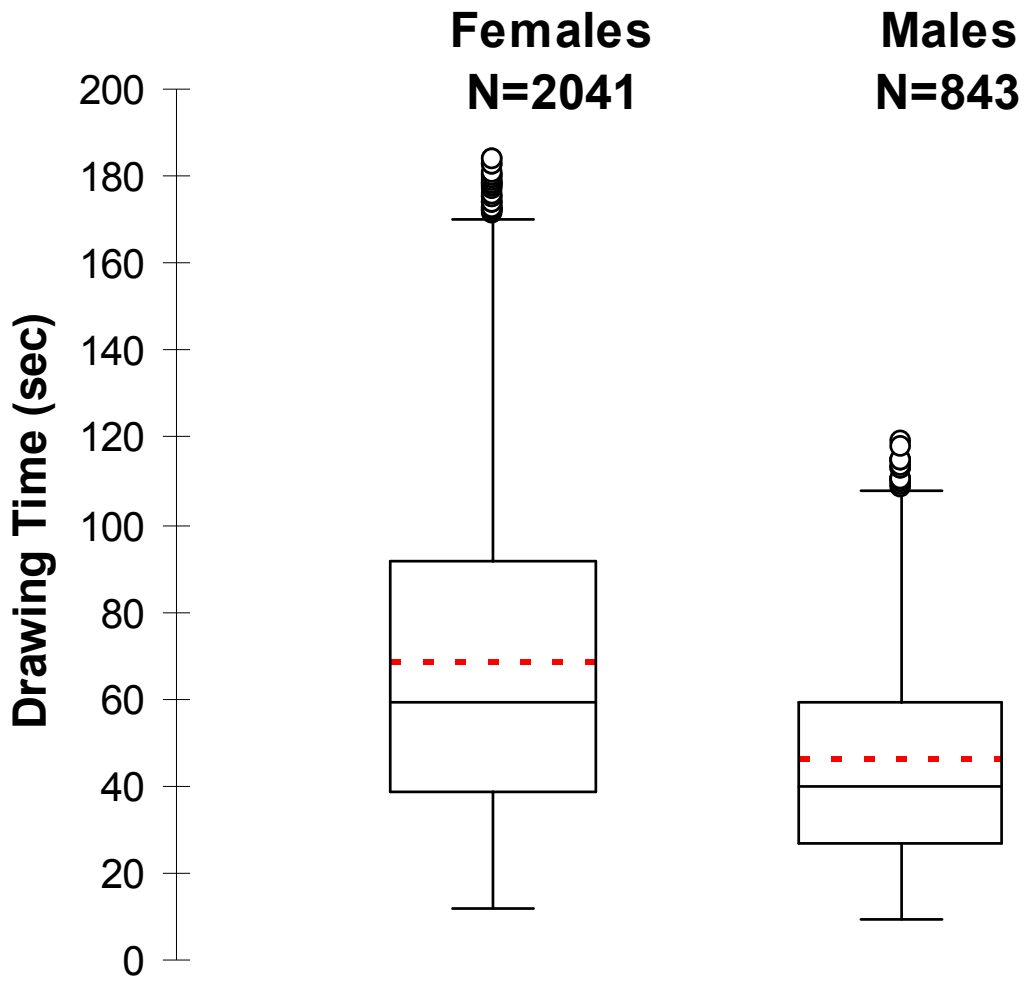
Line Motion (d=.338)

Females N=296

Males N=155



Mirror Drawing (d=.653)



Mental Rotation (d=.451)

Females
N=3924

Males
N=1734

