The Sex Mosaic: Examining Multiple Approaches on a Visual Spatial Task

by

Erin Phares

Department of Psychology and Neuroscience

University of Colorado, Boulder

Defense Date: Friday, October 30, 2020

Presented to:

Dr. Robert Spencer, Department of Psychology and Neuroscience, Thesis Advisor

Dr. Heidi Day, Department of Psychology and Neuroscience, Honors Council Representative

Dr. Cathy Comstock, Department of Humanities, Outside Member

Abstract:

Identification of sex differences in the human brain are of high interest due to their support of a binary understanding of sex. Cleanly characterizing individuals into distinct binary categories of male and female helps justify a long history of differential treatment and expectations of females and males in education, in the work place and in the home. Piaget's 1946 Water-Level Test, originally designed to investigate a timeline of human development of physical principles like gravity, has been used numerous times since the 1960s to equate good performance on the test with males and poor performance with females. This study, while it acknowledges differences in performance of the water level test between sexes, paints a different picture of success and failure for this test. It provides evidence of an additional strategy that both females and males use. Besides drawing a horizontal line, as most studies have assessed, this investigation uncovered an approach of drawing a glass half-full. When looking at the test results in a broader context like this, it shows that no approach is exclusive to a particular sex. The biggest portion of females and males both approach the task as one of both attempting to draw a horizontal line and a glass half full, with males and some females then leaning on an approach of drawing a horizontal line and females and some males then leaning on an approach of drawing a glass half-full. Since none of the approaches are exclusive to one sex, this research seeks to influence future experiments to go beyond a binary approach when characterizing differences in human brains and behavior and recognize a spectrum of possibilities that represents the similarities of the human species.

Introduction:

This research was initially undertaken to explore underlying factors contributing to an observed and consistent pattern of sex differences in performance on an in class, paper and pencil water-level task derived from Piaget's Water-Level test. This variation of the water-level test was performed by 474 students across 12 semesters of teaching an upper division Psychology / Neuroscience class at University of Colorado - Boulder. Specifically, across all semesters, it was observed that males attempted to draw a horizontal line more than females. This thesis research project uncovered an alternative and reasonable approach to the task that was used by both males and females, which was previously unconsidered in other research related to variations of Piaget's Water Level Test. Namely, males and females often attempt to draw a glass half-full in addition to or instead of attempting to draw a horizontal line. Overall, males' and females' approach to the task overlapped no matter what approach they took. Additionally, there was not a significant effect of sex on whether they chose a reasonable approach (both a horizontal and glass half-full approach or just a horizontal approach or just a glass half-full approach). These findings challenge the traditional definition of success for variations of Piaget's test and challenge arguments that there are distinctive male or female brains. Instead, the findings of this research support research describing brains as "mosaics of features, some more common in females compared with males, some more common in males compared with females, and some common in both females and males" (Joel, et al., 2015).

Obvious adult secondary sexual characteristics such as genitals, hair growth and fat distribution have resulted in sex that has historically being described as a binary, with only two possible categories: male and female. Given this external dichotomy, much energy has been spent to find measurable differences in the human brain. "To insist that … evolution did not produce biologically based sex influences of all sizes and sorts in the human brain, or that these influences … produce little or no appreciable effect on the brain's function (behavior) is … denying that evolution applies to the human brain" (Cahill, 2014). This either/or mentality has given rise to debates regarding what traits are male versus what traits are female. Much of the science regarding neuronal and behavioral sex differences documents, and thus reinforces, dimorphic traits. For instance, prior research says sex has an effect favoring males in areas of quantitative problem solving (small to moderate effect), overall quantitative abilities (small to moderate effect), spatial perception (small effect), 2-D visual rotation (small effect) and 3-D visual rotation (large effect) (Nelson, 2011). These claims likely contribute to a male advantage in cognitively challenging (and more financially rewarding) occupations that rely on physics, architecture, and performing medical surgeries (Shea, Lubinski, & Benbow, 2001). They may also influence the growing divide between boys and girls in STEM (Science Technology Engineering and Math) classes during primary and secondary education. Cordelia Fine, in her book "Testosterone Rex" refers to the analysis of a highly cited researcher, Dr. Gijsbert Stoet, a professor at the University of Essex in the United Kingdom, whose view is that we "can't deny human biology and nature…If it's typically only in male nature to play with certain kinds of toys, to want to work in particular kinds of occupations…then that surely tells us something about what kind of society it's reasonable to hope for and aspire to" (Fine, 2017).

In 1956, Piaget introduced the original version of the water level test as a tool to study when children discover "real physical laws…such as the constancy of the surface of a liquid, no matter what the angle of the container" (Piaget & Inhelder, 1956). Piaget and Inhelder documented how before age four, children merely scribbled with no identifiable surface, but by age nine, they consistently drew horizontal lines. Since then, the experiment went viral and there have been many experiments with variations. The research to date viewed success as being drawing a horizontal line and failure as not drawing a horizontal line. The results primarily centered around sex differences on adult performance of "Piagetian" water level tests and contributed to the assessment that males excel over females in visual spatial tasks (Vasta & Liben, 1996).

Researchers analyzed previous experiments on variations of Piaget's Water Level Test and found virtually all results report that males were more accurate than females at performing this test. Yet, they also acknowledge that, like the results in this report, not all males are accurate and not all females are inaccurate. Thus they conclude that "any explanation that ties Water Level Test performance exclusively to the participant's biological sex is untenable" and "there are no differences between the males and females who do well on the task or between the males and females who do poorly (Vasta & Liben, 1996). Vasta and Liben discussed the various mechanisms, including biological (genes associated with the development of the horizontality principle being expressed more in males and exposure to sex-related hormones during prenatal periods) mechanisms and sociological (encouragement of boys in activities that promote spatial skills like playing with blocks or excelling in math and science) mechanisms. They acknowledge that it is more likely an interaction mode, a "specific combination of biological potential and prior experiences" that most influence the observed outcomes of sex differences when performing visual spatial tasks (Vasta & Liben, 1996).

Based on analysis of grey matter, white matter, and connections in 1,400 human brains, other researchers conclude that there are some traits that are more common to females and other traits that are more common to males. Of the extensive number of features in the brain, a single brain will rarely contain all the traits more common in one sex or the other. Rather, each brain is a mixture of attributes and contain some features that are more common in males alongside others that are more common in females. “Thus, most humans possess a mosaic of personality traits, attitudes, interests, and behaviors, some more common in males compared with females, others more common in females compared with males, and still others common in both females and males” (Joel, et al., 2015).

To examine whether reported sex differences in visual spatial task performance would be evident among college students, a simple paper and pencil Water-Level Test was administered to students in a classroom setting. The test was adapted from Piaget’s Water-Level test, that has previously been reported to show sex differences in performance (Voyer, Voyer, & Bryden, 1995). This test was administered over a number of years to students attending a college course at a large state University. The course was initially titled “Hormones and Behavior” and was taught primarily to Psychology Majors. The course eventually was revised and renamed “Behavioral Neuroendocrinology”, and taught to a combination of Psychology Majors and Neuroscience Majors.

This research study looked at performance on water-level tests on a variety of independent variables, including sex, handedness, and major to decipher not just whether males did better but also to understand the overlaps in males’ and females’ behaviors and identify potential underlying reasons for performing differently.

Part of this experiment was to take two dependent measures based on the line drawn. The first dependent measure was an angle representing the degrees off horizontal of the line drawn. Each angle measurement was sorted into two categories: **attempted** horizontal line versus **not attempted** horizontal line and horizontal versus not horizontal. The second dependent measure taken was a percentage representing the area in the glass underneath the line drawn. Each percentage, measuring fullness, was categorized as either **attempted** half-full or did not **attempt** half-full. The objective of this research study was to determine whether or not there were clear differences in performance between the independent variables of sex, major in school and handedness. Previous studies support a hypothesis that males would be expected to be better at drawing horizontal lines (Roberts & Bell, 2002).

This study provided some support for that result. Males **attempted** to draw a horizontal line and were also somewhat more **accurate** at drawing a horizontal line than females who attempted to draw a horizontal line. A surprising result of this in-class experiment, however, was that there was a very reliable sex difference in the number of individuals that attempted to draw a horizontal line, with a higher percentage of males attempting to draw a horizontal line than females. This result has not been previously reported and is likely a result of the fact that the task instructions did not explicitly direct the participant to draw a horizontal line.

While other water level test experiments focused solely on drawing a horizontal line (Liben L. , 1978) as the criteria for doing the task correctly, the design of this study allowed exploration of alternative approaches in addition to or instead of attempting to draw a horizontal line. A prominent alternative approach used by both males and females was attempting to draw a glass half-full in addition to or instead of focusing on drawing a horizontal line. This thesis, therefore, aims to uncover approaches to the water level task not previously considered. These approaches not only question the traditional criteria of success and failure but also demonstrate underreported commonality between sexes on the water level test.

Methods:

Participants:

The participants were college students enrolled at University of Colorado-Boulder. These students were enrolled in Dr. Robert Spencer's upper division/graduate level Behavioral Neuroendocrinology Class (NRSC 4092 - previously known as PSYC 4092/5092 - Hormones and Behavior). The class was listed as a Psychology class until 2013 when it became cross-listed as a Psychology and Neuroscience course. Typically, students taking the class were pursuing a psychology and/or neuroscience major. The water-level test was handed out about halfway into the semester for 12 semesters spanning 20 years and included 474 participants, 159 of those being male and 315 of those being female.

Methods and Materials:

An exercise was given to all students enrolled in 12 separate semesters of the University of Colorado - Boulder, Behavioral / Hormones and Behavior class (NRSC/PSYC 4092/5092) between 1998 and 2017. About the halfway point in the semester, the professor administered the water-level tests to the class (except for one semester, in which a graduate student administered the test). The water-level test data sheets were handed out with minimal instruction, in a classroom setting. Students took this test simultaneously and were in the same classroom together. After the water-level tests were completed, which only took a couple of minutes, they were collected and given back to the instructor. The test included a graphic of a tilted glass pouring water into a bowl that is on a table (Figure 1A). It came with instructions for the participants to "Draw one straight line that represents a 1⁄2 filled glass of water" in a tipped glass. Demographic questions about the participant were included on the test as well which included age, birthday, sex, major, class in school, handedness, and previous employment history.

Data Gathering:

The data was analyzed by first digitizing each water-level test data sheet. This was done by scanning each semester's water-level test data sheets into a PDF file, which was emailed to Dr. Spencer (professor) and to his undergraduate student research assistant, Erin Phares. Using the editing functions in Adobe Reader, each data sheet was numbered and a horizontal line (parallel with the top of the table and bowl and intersecting one edge of the line drawn) was drawn using the shapes tool in Adobe and was put on the glass to form an angle with the line that the participant drew (similar to the dotted line in Figure 1B). The marked-up PDF document was converted into .jpg format by using an online site, www.pdftojpg.net.

A computer program called ImageJ was used to measure each angle. To measure the angle in the ImageJ software, a line was traced over the superimposed horizontal line and the line the participant drew. The ImageJ software calculates the angle. The angle represents the degrees off horizontal the participant drew their line. ImageJ was also used to determine the area of the water in the glass for each participant. The area between the bottom of the glass and the line drawn was compared to the total area of the glass (a constant for all participants) to determine how full the glass was that each participant drew.

The dependent measures were whether the participant attempted to draw a horizontal line (parallel with the top of the table and bowl), whether those who attempted to draw a horizontal line were accurate or not, whether the participants attempted to draw a half-filled glass of water (lik, the angle off horizontal for those who attempted a horizontal line and the approach to the task the participant took. Objective criteria used for each of these dependent measures was determined empirically based on the distribution of values observed in the entire sample. Determination of these criteria are therefore presented in the Results section.

Statistical analysis:

Statistical tests were run using computer software IBM SPSS or Microsoft Excel. For categorical dependent variables, which include accuracy (horizontal/not horizontal) and whether or not they approached the task by attempting to draw a horizontal line or not (attempt horizontal/not attempt horizontal), and whether they approached the task by attempting to draw a glass half-full (attempt half-full/not attempt half-full) a Chi-Squared test was used. For the numerical dependent variable (average angle), either a t-test or an Analysis of Variance (ANOVA) was used. A t-test was used when the independent variable contained two groups (e.g., sex, major, or handedness). ANOVA was used when the independent variable contained more than two groups (e.g., age and class in school). The results of all inferential statistical tests are summarized in Table 2.

Results:

Set criteria for grouping results into binary categories concerning horizontal lines and accuracy:

One primary dependent measure for this experiment was the angle of the line the participant drew compared to a horizontal line. To determine this angle of the line drawn, a horizontal line, parallel with the top of the table and bowl and intersecting the line drawn on one side of the glass was overlayed on the participant's drawing (similar to the dotted line in Figure 1B). A horizontal line was zero degrees and the measurement for each individual was how many degrees off horizontal the line was. Measurements were taken using ImageJ software. Values ranged from zero degrees up to 87.00 degrees. Each result was rounded to two decimal places and grouped by frequency into bins of 2.5 degrees in order to help determine binary criteria for whether each participant was attempting to draw a horizontal line and whether they were successful at drawing a horizontal line. In Figure 2, angles are grouped in bins of two and a half degrees each (ranging from 0-2.49 up to 62.5-64.49). Each individual was sorted into a bin based on the angle of their line compared to horizontal. Only four individuals had angles 65 degrees or higher, so they were all grouped together in a 65+ category. The resulting graph was bimodal and skewed. It showed natural breaks between those who attempted to draw a horizontal line and did not attempt as well as those whose lines were horizontal versus not horizontal. Figure 2 illustrates a clear break point after the 27.5-27.99 bin. This point is recognized as differentiating between the categories of those who attempted horizontal (<30 degrees off of horizontal) and those who did not attempt horizontal (≥ 30 degrees).

Within the grouping of the category of those who "attempted horizontal", as seen in Figure 2, there was another clear break point at 10 degrees. Angles less than 10 degrees were categorized as "horizontal". Horizontal lines were scored accurate. Angles 10 degrees or higher, including those that were "not attempted horizontal" were categorized as "not horizontal".

Roberts and Bell also used a 10 degree cut-off value for accurate measurements in the water-level test that they performed (Roberts & Bell, 2002).

Males consistently attempted to draw a horizontal line more than females across all semesters:

Figure 3 shows a comparison by semester between males and females and whether they approached the exercise by attempting to draw a horizontal line. Remarkably, across all semesters, males consistently attempted to draw a horizontal line more frequently than females. This strong finding served as the basis for all further analysis and drove the investigation by looking at factors potentially involved in how an individual approached this exercise.

Males attempted a horizontal line more frequently and were more accurate at drawing a horizontal line:

Figure 4A shows that a greater percentage of males (77%) attempted to draw a horizontal line (a line < 30 degrees off horizontal) compared to females who attempted (only 56%). Sex had a pronounced and significant effect (Chi Square test, $p<0.0005$; Table 2) on whether they approached the task by attempting to draw a horizontal line. The question remained on how accurate the individuals were at drawing a horizontal line. As discussed earlier, anyone who drew a line that was within 10 degrees of horizontal, was considered to have been accurate at drawing a horizontal line. Sex also had a significant effect (Chi Square test, $p=0.007$; Table 2) on accuracy (whether they drew a line within 10 degrees of horizontal) for those who were trying to draw a horizontal line (Figure 4B). Of those that attempted to draw a horizontal line (i.e., those whose angles were within 30 degrees of horizontal), 83% of males accurately drew a line within 10 degrees of horizontal compared to only 69% of females (Chi Square test, $p=0.007$; Table 2). Based on the binary categories of either “attempted horizontal” or “not attempted horizontal” as well as “horizontal” or “not horizontal” described in Figure 2, it was clear that males more than females approached the task by attempting to draw a horizontal line. Furthermore, for all males and females who attempted to draw a horizontal line, males were more likely than females to draw a line within 10 degrees of horizontal line. However, when the angle drawn was treated as a continuous variable, there was not as clear of a difference for those who were trying to draw a horizontal line. The mean angle of the line drawn for males who were trying to draw a horizontal line was 7.52 degrees off horizontal, while females who attempted to draw a horizontal line on average drew a line with a slightly wider angle at 8.52 degrees off horizontal. Sex, therefore, did not have a significant effect, but instead, only a trend (t-test, p=0.1; Table 2) on the angle drawn by males and females who approached the task by attempting to draw a horizontal line (Figure 4C).

Females attempted drawing a half-filled glass more frequently than males:

Since females were not as concerned with their line being horizontal like males were, the question about what motivated their approach arose. As discussed in the prior section, males attempted to draw a horizontal line more frequently than females and males drew a line within 10 degrees of horizontal more frequently than females (Figures 4A and 4B). It was possible, however, that females, were most focused on drawing a line that showed the glass as being half-full, and with that focus they may have been as accurate or more accurate than males. To test this possibility, first criterion was established for determining whether the participant attempted to draw a line showing the glass half-full. Figure 5 graphs how full the glasses were based on the area of the total glass compared to the area in the glass underneath the line they drew, regardless of whether the lines were horizontal or not. The mean fullness for all people who drew the line in the glass and not in the bowl was 52% full with a standard deviation of 8%. It is reasonable to consider an attempt at half-full to be within one standard deviation of the mean. Anyone who drew a glass that was between 45% full and 58% full, therefore, was considered to have attempted to fulfill the instruction of drawing a glass half-full.

Figure 6 shows that a greater percentage of females attempted to draw a glass half-full than males, especially if the females were not trying to draw a horizontal line. Sex (male/female), though, did not have a significant effect (Chi Square test, p=0.37; Table 2) on whether they approached the task as one of drawing a glass half-full. 69% of all females attempted to draw a glass half-full and 65% of all males attempted to draw a glass half-full. For just those not attempting to draw a horizontal line, though, there was a significant effect (Chi Square test, p=0.02; Table 2) of sex on whether they attempted to draw a glass half-full. Of those who did not attempt a horizontal line, a greater percentage of females (76%) were instead attempting to draw a glass half-full than males (56%). It appeared that females more frequently approached the task by following the directions and attempting to draw a glass half-full when they were not trying to draw a horizontal line.

Left-handed people attempt to draw a horizontal line more frequently than right-handed people:

Handedness (left or right) was analyzed to look for other influences on whether participants were more likely to approach this exercise from a perspective of drawing a horizontal line. The data found in Figure 7 is interesting because it shows that more left-handed people attempted to draw a horizontal line than right-handed people, however, the difference did not reach statistical significance (Chi Square test, p = 0.11; Table 2). Out of the left-handed people, 74% attempted to draw a horizontal line, whereas 61% of right-handed people attempted to draw a horizontal line. In this study, 8% of females and 10% of males were left-handed, compared with global norms of left handers being 10% (Papadatou-Pastou, et al., 2020). The global average of the ratio of male and female left handers indicates that males are 1.23 more likely to be left-handed than females (Papadatou-Pastou, Martin, Munafò, & Jones, 2008). This study was in line with that. Even though males were 1.21 times more likely to be left-handed in this data set than females, no statistical tests show a correlation between handedness and performance on the water-level test (Chi Square test, p>0.05; Table 2). The data also showed no correlation between handedness and whether or not the participants attempted to draw a glass half-full by the criteria used in this investigation.

Natural science majors attempted a horizontal line more frequently and social science majors attempted a glass half-full more frequently:

Besides handedness, the data contained information on each participant's major and minor fields of study. The individuals were classified into social sciences or natural sciences. Table 1 documents the fields of study considered social sciences and natural sciences for the purpose of this investigation. If the participant had at least one of the Natural Sciences listed as either a major or minor, they were considered to be a Natural Science major, regardless of whether they also designated one of the social sciences as a field of study. Figure 8 shows that a greater percentage of natural science majors attempted to draw a horizontal line than social science majors. Major (natural science versus social science) had a significant effect (Chi Square test, p=0.001; Table 2) on whether they were more likely to draw a horizontal line. 69% of all natural science majors attempted to draw a horizontal line and 55% of all social science majors attempted to draw a horizontal line. Natural science versus social science majors did not have a significant effect (Chi Square test, p>0.05; Table 2) on accuracy (whether they drew a line within 10 degrees of horizontal) for those who attempted a horizontal line. Major classification also was not significant (t-test, p>0.05; Table 2) on angle for those who attempted to draw a horizontal line.

Major classification also influenced whether the participant attempted to draw a glass half-full. 71% of Social Science majors approached the task by drawing a glass half-full compared to only 62% of Natural Science Majors (Figure 8B). This result was significant (Chi Square test, p=0.02; Table 2). There was not a significant effect of major classification when drawing a glass half-full when looking at just those who attempted to draw a horizontal line or just those who did not attempt to draw a horizontal line.

We also looked at interactions between sex and major. Sex had a significant effect (Chi Square test, $p=0.012$; Table 2) on how they approached the task for natural science majors. 79% of the males attempted to draw a horizontal line and 63% of the females attempted to draw a horizontal line. Sex did not have a significant effect (Chi Square test, $p>0.05$; Table 2) on accuracy for natural science majors who attempted to draw a horizontal line. Sex was not significant (t-test, $p>0.05$; Table 2) on angle for natural science majors who attempted to draw a horizontal line. Sex had a significant effect (Chi Square test, $p<0.0005$; Table 2) on how they approached the task for social science majors. 73% of males attempted to draw a horizontal line and 47% of females attempted to draw a horizontal line. Sex did not have a significant effect (Chi Square test, $p>0.05$; Table 2) on accuracy for social science majors who attempted to draw a horizontal line. It also did not have a significant effect (t-test, $p>0.05$; Table 2) on angle for social science majors who attempted to draw a horizontal line.

Major (natural science versus social science) had a significant effect (Chi Square test, $p=0.003$; Table 2) on how females approached the task. 63% of the female natural science majors attempted to draw a horizontal line and 47% of female social science majors attempted to draw a horizontal line. Major (natural science versus social science) did not have a significant effect (Chi Square test, $p>0.05$; Table 2) on accuracy for females who attempted to draw a horizontal line. It also was not significant (t-test, $p>0.05$; Table 2) on angle for females who attempted to draw a horizontal line.

There was no significant effect (Chi Square test, $p>0.05$; Table 2) for major (natural science versus social science) on how males approached the task, on accuracy for males who attempted to draw a horizontal line, and on angle for males who attempted to draw a horizontal line.

Separated approach into Single Focus, Dual Focus or Neither Focus:

Based on the above results, participants were categorized into one of four categories: 1) Single focus on drawing a horizontal line only 2) Single focus on drawing a glass half-full only 3) Dual focus on drawing a horizontal line and glass half-full and 4) Neither focusing on drawing a horizontal line nor glass half-full. The most prominent approach for all males and females was 3) Dual focus on drawing a horizontal line and glass half-full (Figure 9A, in green).

For ease in contrasting the four categorical approaches to more traditional analysis of the results, Figure 9B documents how a traditional study might classify success and failure. That is, those who drew a horizontal line within the criteria defined (in the case of this study < 10 degrees off horizontal), would be considered successes and the remainder would be failures, regardless of their approach. Males in this case succeeded at a rate of 64% compared to females at 38% (Chi Square test, $p = \leq 0.0001$; Table 2). Figure 9C restates the same data as in Figure 4A for ease of contrasting with the four categorical approach. It shows males approached the task by attempting to draw a horizontal line 77% whereas females approached the task by attempting to draw a horizontal line 58% of the time (Chi Square test, $p = \leq 0.0005$; Table 2).

Discussion:

Recall that the mission of this study was to determine correlations for a consistent pattern of sex differences observed across separate groups in performance on a paper and pencil water-level task in a classroom setting. This investigation was initiated to help explain the phenomena and implications of the observation. Exactly 12 semesters of water level tests, which happened across many years, were examined in this study. A trend that males and females approached this variation of Piaget's Water Level Test differently from each other held up: throughout all semesters, males repeatedly and consistently attempted to draw a horizontal line more frequently than females (Figure 3).

For each of the 12 semesters, Dr. Robert Spencer, professor in the Neuroscience department at the University of Colorado-Boulder, administered a two-minute exercise in his Behavioral Neuroendocrinology class as part of the topic of hormones and structural differences in the brain that potentially could influence sexual dimorphism. The pencil and paper exercise instructed students to "Draw one straight line that represents a 1/2 filled glass of water" above a graphic of an empty glass in the pouring motion hovering over a tabletop that had an empty bowl sitting on top of it (Figure 1A). This study departed from similar water level test studies in that the instructions were not explicit to draw a horizontal line and there were no guide points on the glass itself prompting participants to draw a horizontal line. This ambiguity opened up different strategies on how to approach the task. Still, given that ambiguity, it makes it more interesting that so many people (both males and females) still inferred that they should attempt to draw a horizontal line, in addition or instead of the strategy to draw a glass half-full (for instance with a transverse line that bisects the glass, so that the line is parallel to the bottom of the glass).

Traditional water-level test variations focused on accuracy of drawing a horizontal line (Pascual-Leone & Morra, 1991) (Liben L. , 1991). Instead of looking just at accuracy, this investigation uncovered an additional way that both males and females approached the task of drawing a line besides drawing a horizontal line. It is true that males attempted to draw a horizontal line more often than females (Figure 4A). Also, like prior research, males, particularly those who attempted to draw a horizontal line, also more accurately drew a line within 10 degrees of horizontal (Figure 4B). Traditional studies stopped there and discounted any other non-horizontal performance, concluding that failures (mainly female) demonstrated deficits in visual spatial ability (Liben L. , 1991) (Robert & Morin, 1993). This study also supported those results (Figure 9B) and could have stopped there, contributing to the body of work that set the stage for explaining and justifying differences in male and female behavior. Instead, this research explored what others classified as failures. We found that for all participants, females attempted to draw a half-filled glass more often than males. But that result (Figure 6A) was not significant (Chi Square test, $p=0.37$; Table 2) meaning both males and females employed this approach without effect of sex. More interestingly, of the participants who did not attempt to draw a horizontal line (Figure 6B), sex had a significant effect (Chi Square test, $p = 0.02$; Table 2) on whether they instead attempted to draw a glass half-full, with females (76%) choosing this strategy more than males (57%). Since the instructions stated that drawing a half full glass was an objective of the exercise, this was a reasonable approach. Other studies would have classified these students (females and males) who did not even attempt a horizontal line as failures without exploring alternative motivations.

In the ongoing theoretical debate around sex differences, both biological and sociological mechanisms have been discussed, but they have done so more with the lens pointing at deficits in females’ performance (Vasta & Liben, 1996). This investigation introduced a reasonable alternative approach to just focusing on a horizontal line. It outlined how the approach to the water level task should not be a binary outcome of success or failure based on how horizontal the line drawn was, but instead an analysis of what factors, including sex, education, life experience, and perhaps even handedness that influenced how the line was drawn. It is worth considering based on the data presented here, that differences in approach between the sexes is not rigid or strictly male or female since so much overlap exists between the sexes (Figure 9A). Others have come to a similar conclusion, that even within what has historically been classified as gender-stereotyped strengths (such as visual spatial ability), a lot of variability and overlap occur (Hyde, 2005) (Carothers & Reis, 2013). This perspective allows for gradients of behavior with some males and females exhibiting a preference for drawing a horizontal line and some males and females exhibiting a preference for drawing a glass half-full. This exploration and results of this study broadened the dialogue to explore the possibility that approaching the exercise as depicting a glass half-full is just as sensible and realistic in the water level task as drawing a horizontal line, even though it may be rooted more on practicality than theoretical considerations**.** It illuminates how traditional studies on sex differences with two rigid contrasting and mutually exclusive categories, male or female, overshadow the range of behaviors for either sex. This study instead offers solid reasoning for abandoning rigid and dualistic categorization when it comes to debating differences between the sexes.

Handedness (left or right) may also have been a factor in attempting to draw a horizontal line. Prior research suggests that left-handedness is associated with advantages performing visuospatial tasks (Swirsky-Sacchetti & Genetta-Wadley, 1990). In this study, 9% overall (8% of females and 10% of males) were left-handed and males were 1.21 times more likely to be left-handed in this data set than females. These demographics were similar to those expected in the general population. The ratio of left-handed males to females was expected to be 1.23 times (Papadatou-Pastou, Martin, Munafò, & Jones, 2008) and the overall estimate worldwide of left-handers is 10% (Papadatou-Pastou, et al., 2020). Left-handed people attempted to draw a horizontal line more than right-handed people and ambidextrous people were split down the middle (Figure 7) but there was not a significant effect (Chi Square test, $p = 0.11$; Table 2) of handedness on whether the participants attempted to draw a horizontal line.

There was a significant effect of educational background (Figure 8A) with natural science majors attempting a horizontal line (69%) more than social science majors (55%) (Chi Square test, $p = 0.001$; Table 2). Social science majors also attempted to draw a glass half-full (71%) more than natural science majors (62%) (Figure 8B; $p=0.02$; Table 2). When attempting to draw a horizontal line (Figure 8A) there was still a significant effect of sex (Chi Square test, $p = 0.012$; Table 2) for natural science majors with males still attempting a horizontal line (79%) more often than females (63%) (Figure 8A). When just looking at females, there was a significant effect of major (Chi Square test, $p= 0.003$; Table 2) on whether or not females attempted to draw a horizontal line, with 63% of female natural science majors attempting a horizontal line compared with only 47% of female social science majors attempting a horizontal line. More male natural science majors attempted a horizontal line (79%) than male social science majors (73%) but there was not a significant effect of major for males (Chi Square test, $p=0.392$; Table 2). The implication of this correlation might be related to natural science majors and males having more experience or exposure to scientific concepts such as the effect of gravity on liquids. It also might imply that males overall and females who chose natural science majors already had an aptitude for these scientific concepts or think in a more critical way than social science majors.

Halpern, et al., acknowledge the interaction is complex and no simple answer exists: "… causal statements about brain differences and success in math and science are circular. A wide range of sociocultural forces contribute to sex differences in mathematics and science achievement and ability… early experience, biological factors, educational policy, and cultural context affect the number of women and men … in science and math and that these effects add and interact in complex ways" (Halpern, et al., 2007).

When looking at attempting to draw a glass half-full, 62% of all natural science majors and 71% of all social science majors (Figure 8B) approached the task by attempting to draw a glass half-full (Chi Square test, p=0.02; Table 2), based on the criteria used. Other groupings showed that social science majors and females overall were focused on the task of drawing a glass half-full more than natural science majors, or males. But, major did not have a significant effect on these other groupings. In particular, there was not a significant effect of major when looking at female social science majors (72%) compared to female natural science majors (65%) (Chi Square test, p=0.14; Table 2) or looking at male social science majors (69%) compared to male natural science majors (57%) (Chi Square test, p=0.12; Table 2). Likewise, though females within either major overall attempted to draw a glass half-full more frequently than males, there was not a significant effect of sex within major on whether someone attempted to draw a glass half-full: male social science majors (69%) attempted to draw a glass half-full less frequently than female social science majors (72%) (Chi Square test, p=0.597; Table 2); male natural science majors (57%) attempted to draw a glass half-full less frequently than female natural science majors (65%) (Chi Square test, p=0.238; Table 2). Since the instructions explicitly state to draw a line that "represents a 1⁄2 filled glass of water," it is not surprising that the percentage of participants overall that attempted a glass half-full are fairly high. What is interesting is that generally social science majors and females were more concerned with the explicit task of drawing a glass half-full, while natural science majors and males were more concerned with the theoretical task of drawing a horizontal line.

This research replicates what has been published regarding sex differences in the performance of Piaget's Water Level Test. Namely, males generally are more successful than females at the task of drawing a horizontal line within a tilted object (Kalichman S. C., 1988). Instructions in prior studies, though, may have been more explicit in their focus on drawing horizontal lines. Some instructions included cues such as a small dot on the glass as the starting point for where the water intersects one edge of the glass or an instruction to draw a line that "represented the surface of the water" (Hecht & Proffitt, 1995). Other experiments explicitly instructed participants to "to draw a horizontal line about half way up a tilted bottle" or "so that it looks horizontal (or straight across), regardless of whether or not you think it really is horizontal" (Sholl & Liben, 1995). Dr. Spencer's test, on the other hand, simply instructed the participants to "Draw one straight line that represents a 1⁄2 filled glass of water" and did not include a cue dot. Given the ambiguity of the instructions in this regard for Dr. Spencer's in-class exercise, it was remarkable that 77% of males and 56% of females (Figure 4A) still attempted to draw a horizontal line. But the puzzle about what underlies the participants' approach remained.

Other studies with variations of the water level test defined success as only a function of how horizontal the line drawn was. If drawing a horizontal line was the only definition of success, the results clearly favor males (Figure 9B). This study explored alternative mindsets, though, that might have influenced how the participant drew the line. This view paints a very different, more interesting and realistic picture. The instructional ambiguity set up a potential for expanding how an individual approached the task and uncovered previously unobserved sex differences in water level test performance. Participants were divided into one of four categories (Figure 9A): 1) those with a single focus on drawing a horizontal line representing gravity's effect on water; 2) those with a single focus on the task of accurately dividing the glass in half; 3) those with a dual focus on dividing the glass in half and horizontal; and 4) those not too concerned with either how full or how level the water is. Looking at the data in this manner was the most compelling reason for concluding that strict pass/fail criteria for drawing a horizontal line does not allow for expanded and reasonable approaches to this task. Both males and females attempted a dual approach (attempting both a horizontal line and a glass half-full) more than any other approach. Secondarily, males approached the task only with a focus on a horizontal line and for females with a focus on half-full. This nuanced approach is reasonable and has been overlooked in other studies. And that focusing just on a pass/fail approach short-changes other valid approaches and hides the fact that there is a lot of nuanced overlap between males and females.

These results demonstrate a smaller sex difference for those who attempt to draw a horizontal line when looking at natural science majors (Figure 8A) than the overall population of participants (Figure 4A). A 1986 study also looked at major in school and noticed it had a moderating effect on the difference in performance between males and females on an exercise that was also based on Piaget's water level test. In that study, in all cases, males did better with respect to drawing a horizontal line than females, and like this study, for those classified as natural science majors, the sex difference was not as pronounced compared to when all students were analyzed or when just social science students were analyzed (Kalichman S. C., 1986). One possible explanation is that the outcome of being enrolled in natural science courses is that educational history leading up to choosing a major in natural sciences might have exposed these students to more physical principles, such as the effect of gravity on water. On the other hand, it might mean that those with more of an aptitude for spatial activities such as this task were more drawn to natural sciences to begin with. The U.S. Department of Education reported in 2015 that “compared to males, lower percentages of female high school graduates reported that they liked mathematics or science and that mathematics or science was one of their favorite participants” (Cunningham, Hoyer, & Sparks, 2015). Still, this does not point to the degree that socialization and expectations of females and males early in education may influence outcomes like those described in this study or whether females and males are wired differently.

Future directions:

In future semesters, it would be interesting to see the results if Dr. Spencer modified the original exercise slightly by incorporating a dot on the glass as a starting point. He could randomly hand out this to half of each class and the original to half of the class then compare the two groups. He also might consider rewording the directions on the water level test to be less ambiguous. Instead of saying “Draw one straight line that represents a 1/2 filled glass of water” the instructions could say “Draw one straight line that represents a half-filled glass of water, based on gravity.” This would help determine whether the instructions had a significant effect on performance.

Participants could also be recruited to do the exercise from other colleges, such as the business college or the engineering college to get additional data on the effect of experience on the task.

This data should be further analyzed to see whether sex or major classification had a significant effect on the continuous variable of percent full. This analysis could yield more objective criteria for categorizing attempted half-full versus not attempted half-full as well as establishing success criteria for drawing a glass half-full.

Also, enough males and females approached the task in the very same way (Figure 9A), which calls into question the binary treatment of the sexes in all studies to date. Thus, it seems that a purely biological explanation is unlikely. One of the most fascinating aspects in this investigation was uncovering a novel perspective on the data, specifically looking at what the other studies would have lumped together as failures of the task, to find a previously unconsidered approach to this task. While the other studies explicit goal was to find sex differences with performance of the task, they focused only on the performance based on a narrow criteria of horizontality. What if they considered whether their instructions included some previously unnoted ambiguity as well? It would be interesting to revisit the data from the 20 studies discussed in Kalichman's 1988 review (Kalichman S. C., 1988) or the additional 17 studies in Voyers' and Bryden's 1995 review (Voyer, Voyer, & Bryden, 1995) to see how explicit the instructions really were and how the "failures" approached the task. The Water Level Test is understood to be a visual spatial task, but maybe it involves more competencies and skills than that if the data is looked at through a different lens.

Overall Conclusion:

Generally, males and natural science majors approached this task in a way that demonstrated their knowledge of physical principle of gravity's effect on water. Females and social science majors generally approached this task in a way that demonstrated their ability to divide an area in half. Still, so much overlap occurs between the approaches, with females and males both using as their primary strategy a dual approach, both attempting a glass half-full and a horizontal line, that it calls into question whether studies should conclude sex differences based on results of this task. Like the article "Sex beyond the genitalia: The human brain mosaic" advocates, these results support "replacing the currently dominant practice of looking for and listing sex/gender differences with analysis methods that take into account the huge variability in the human brain" (Joel, et al., 2015). Results of our work could lead to further investigation of alternative approaches to the task. Maybe there are biological mechanisms which influence how human brains develop that show that brains are not purely female or purely male but a mosaic. For instance, some researchers believe that prenatal exposure to androgens or estrogens cause organizational effects on brain development, maybe even the areas responsible for spatial processing (Collaer & Hines, 1995). Maybe there are activational effects of hormones that influence spatial performance. It is true there may be differences in sexes but they are exaggerated when we continue to view performance based on a binary approach to sex.

Appendix:

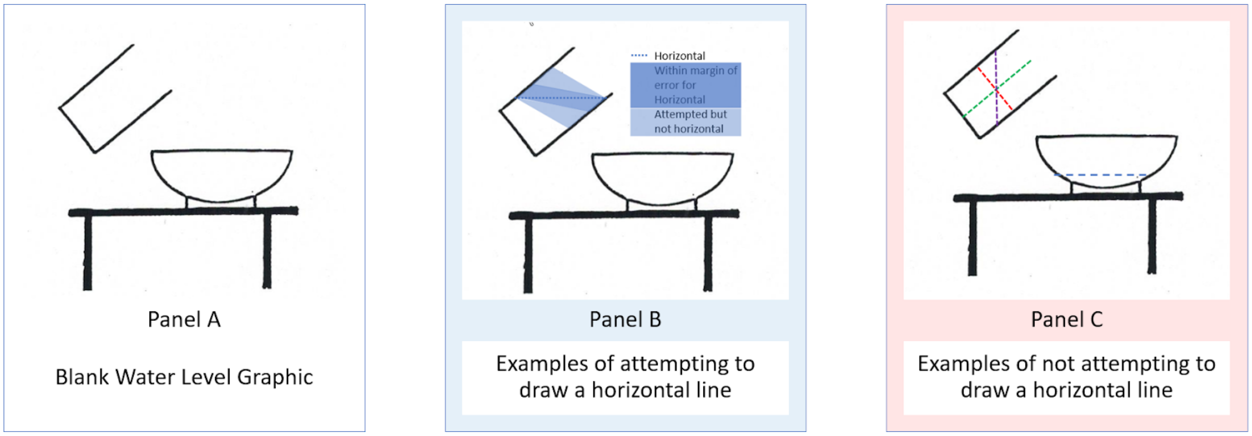


Figure 1, Panel A: Blank Water Level Graphic. Participants were instructed to "*Draw one straight line that represents a 1/2 filled glass of water*". **Panel B**: Examples of attempting to draw a horizontal line. The dashed line is a sample horizontal line. The criteria for being horizontal is to draw a line that is parallel with (or within 10 degrees of parallel) the top of the bowl and the top of the table. A horizontal line can be anywhere within the glass. Those categorized as accurately drawing a horizontal line were less than 10° off their attempted horizontal line (the darker of the two shadings). Those categorized as having attempted to draw a horizontal line but not drawing one accurately, drew a line whose angle was greater than or equal to 10° but less than 30° from horizontal (the lighter of the two shadings). **Panel C:** Examples of not attempting to draw a horizontal line. Those who did not attempt to draw a horizontal line typically would draw a line that was parallel with the bottom of the empty glass (red). Some would instead draw a line perpendicular to the table (purple), parallel with the side of the glass (green) or horizontal, but within the bowl and not the glass (blue).

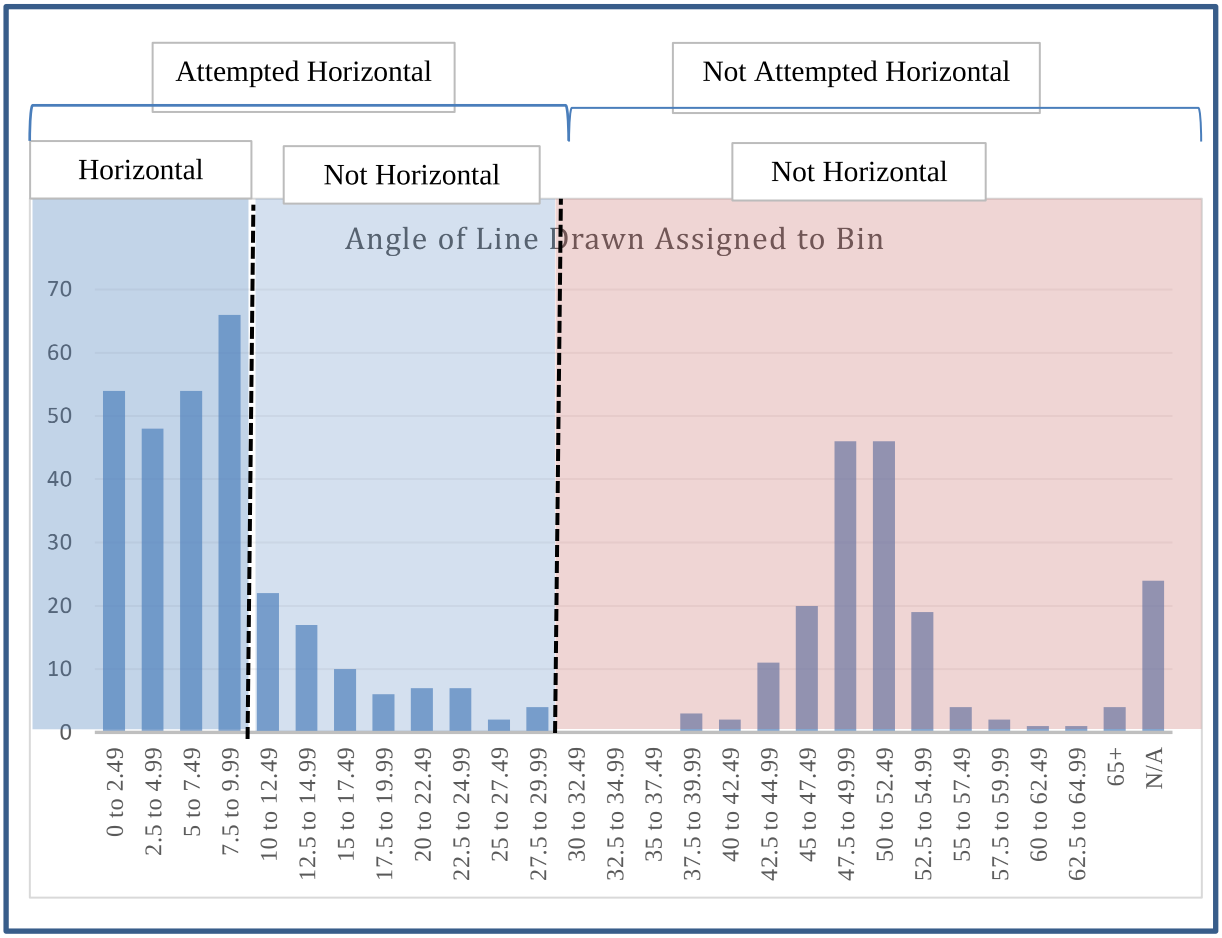


Figure 2 Angles grouped in bins by sex, horizontal versus not horizontal, and attempted horizontal versus not attempted horizontal. The criterion for horizontal is 0 – 9.99 degrees. The criterion for not horizontal is ≥ 10 degrees. The criterion for not attempting horizontal is ≥ 30 degrees. The criterion for attempted horizontal is 0 – 29.99 degrees.

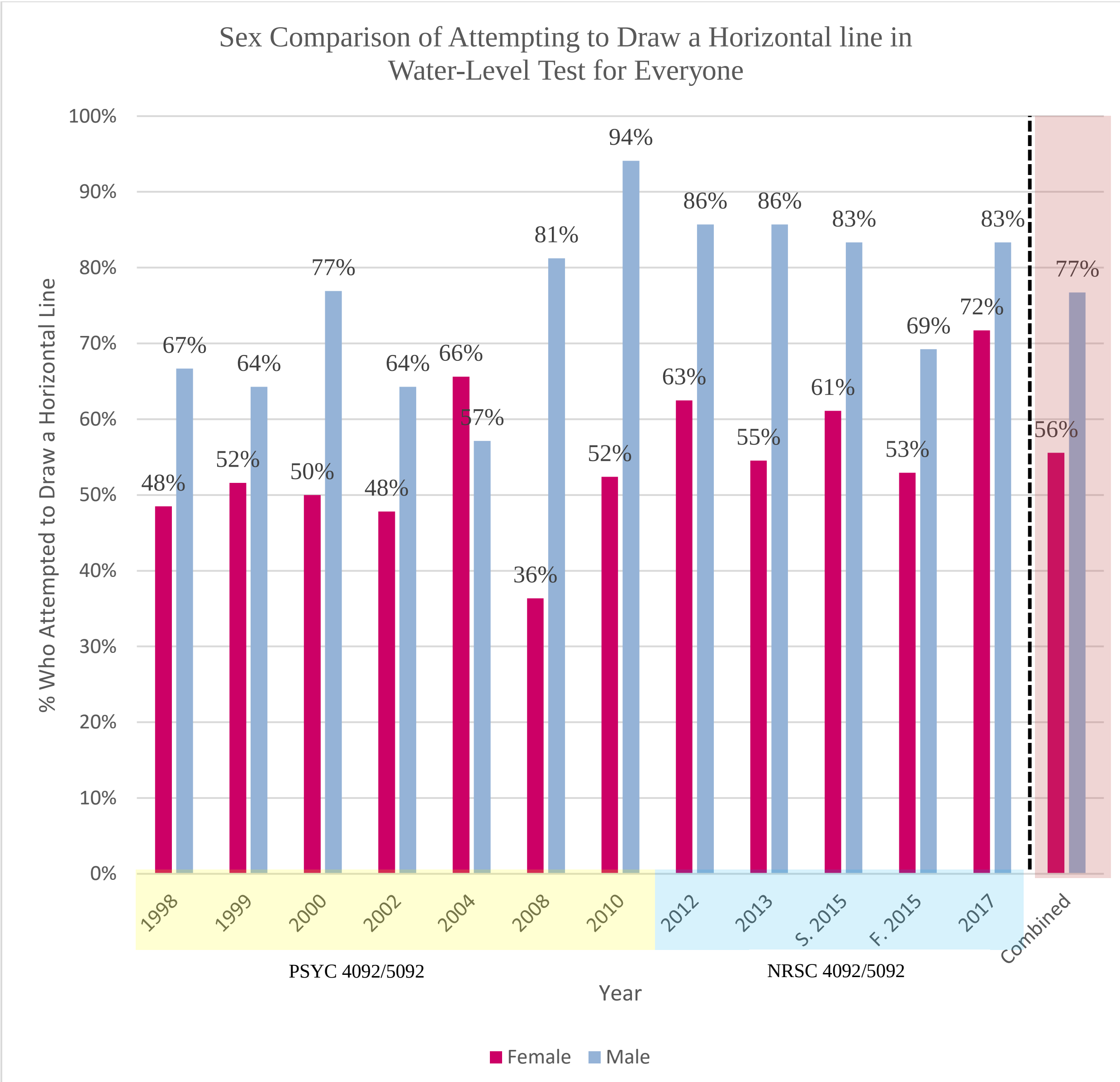
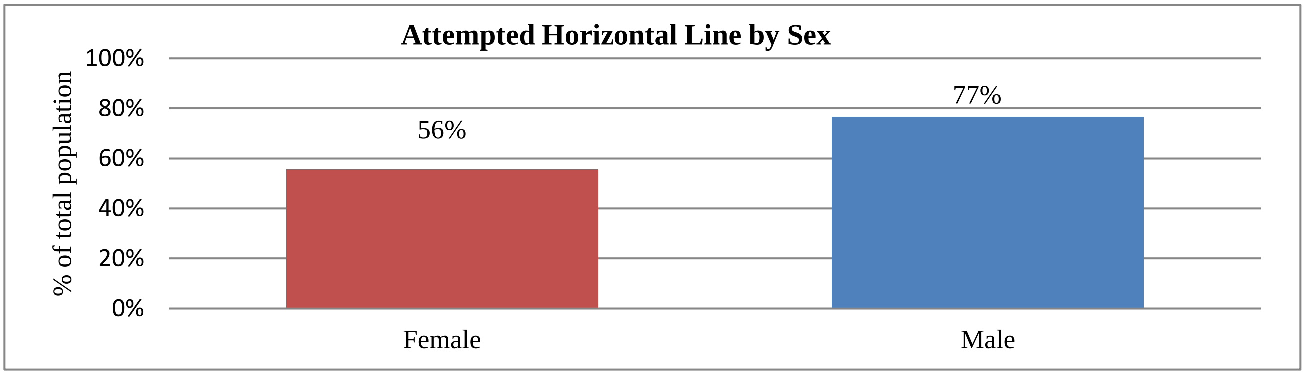
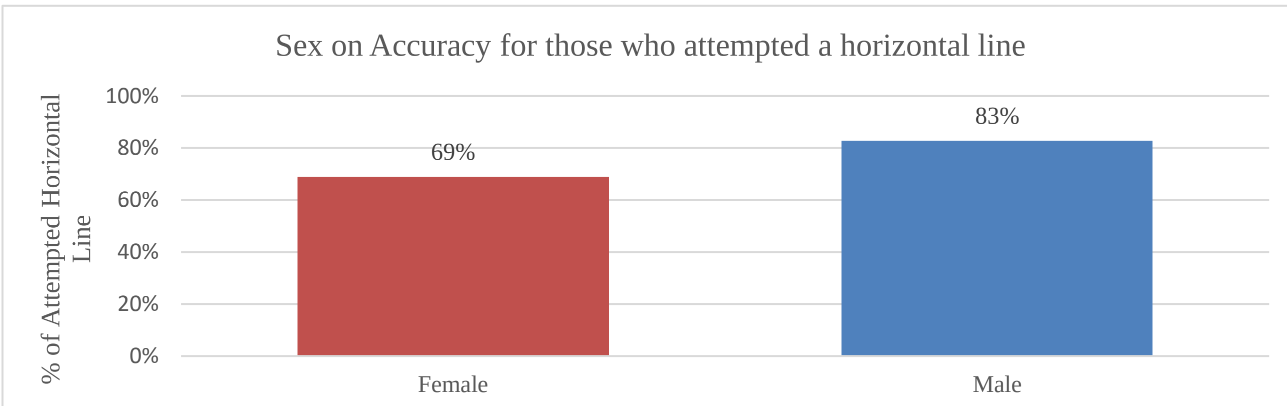


Figure 3 Males in each of the 12 semesters attempted to draw a horizontal line more frequently than females. The last set of numbers looks at all semesters combined together**.** The course was taught as PSYC 4092/5092 through 2010 (in yellow) and taught as NRSC 4092/5092 starting in 2012.

Panel A:



Panel B:



Panel C:

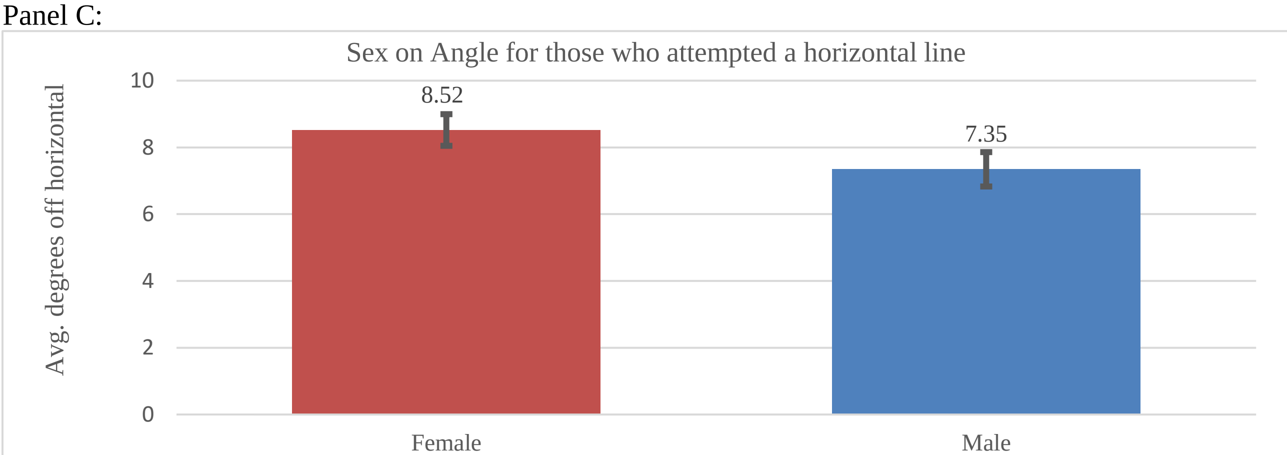


Figure 4 Panel A. A greater percentage of males (n=122, 77%) than females (n=175, 56%) attempted to draw a horizontal line (Chi Square test, $p<0.0005$; Table 2). **Panel B.** Of those who attempted to draw a horizontal line (angle < 10°), a greater percentage of males (n=101, 83%) than females (n=120, 69%) drew a horizontal line (Chi Square test, $p=0.007$; Table 2). **Panel C.** For those who attempted to draw a horizontal line, the average angle males drew (7.35 degrees off horizontal, n=122) was more accurate than that of females (8.52 degrees off horizontal, n=175) (t-test, $p=0.10$; Table 2).

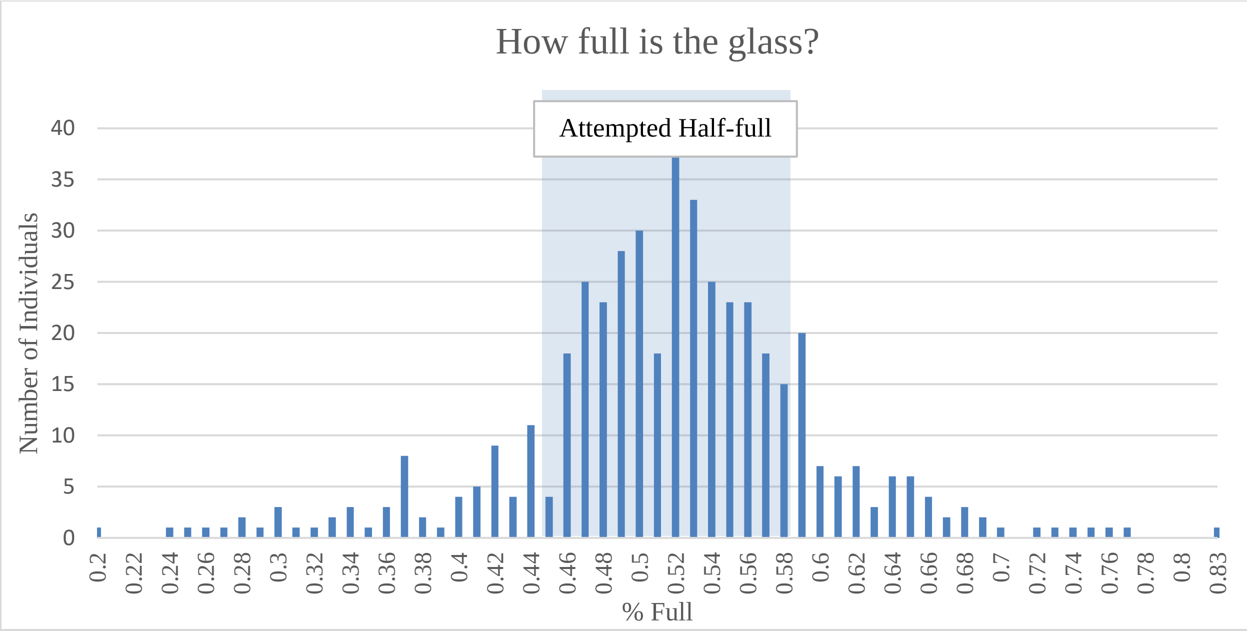


Figure 5 This graph shows how full (based on percentage of full glass) each individual drew the line in the glass. The mean is 52%. The area +/- one standard deviation is considered attempting half-full (between 45% and 58% full).

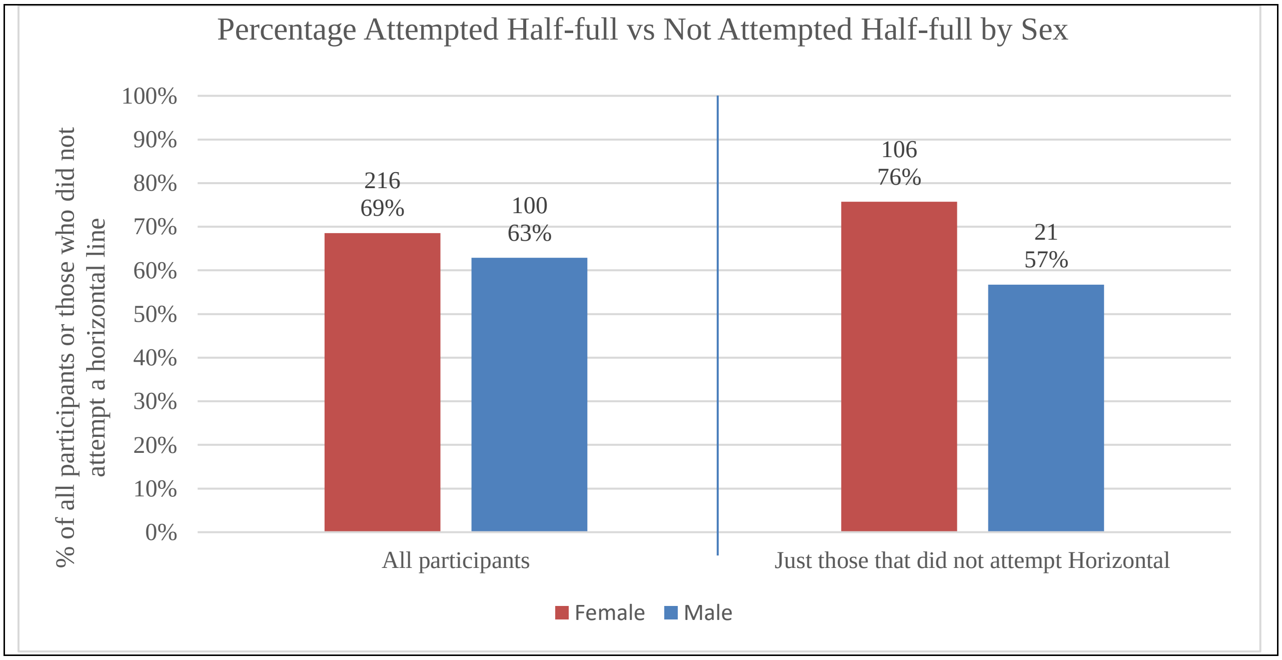


Figure 6 For all participants, a greater percentage of females (n=216, 69%) attempted to draw a glass half-full than males (n= 100, 63%). Sex (male/female), though, did not have a significant effect (Chi Square test, p=0.216; Table 2). For those not attempting to draw a horizontal line, there was a significant effect (Chi Square test, p=0.023; Table 2) of sex on whether they attempted to draw a glass half-full. Of those who did not attempt a horizontal line, a greater percentage of females (n=106, 76%) were instead attempting to draw a glass half-full than males (n=21, 57%).

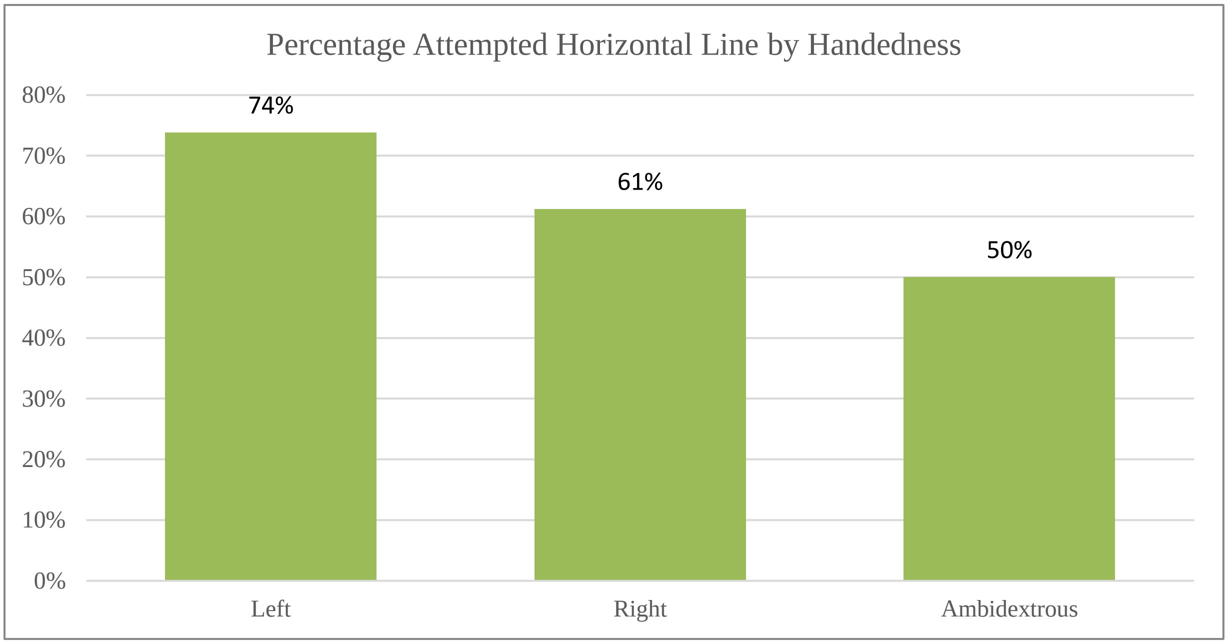


Figure 7 Left-handed people (n=31, 74%) outperform right-handed people (n=259, 61%) at attempting to draw a horizontal line. The ambidextrous people who attempted horizontal and did not attempt to draw a horizontal line is split down the middle at 50% (n=2). (Chi square test, p>0.05, Table 2).

Panel A

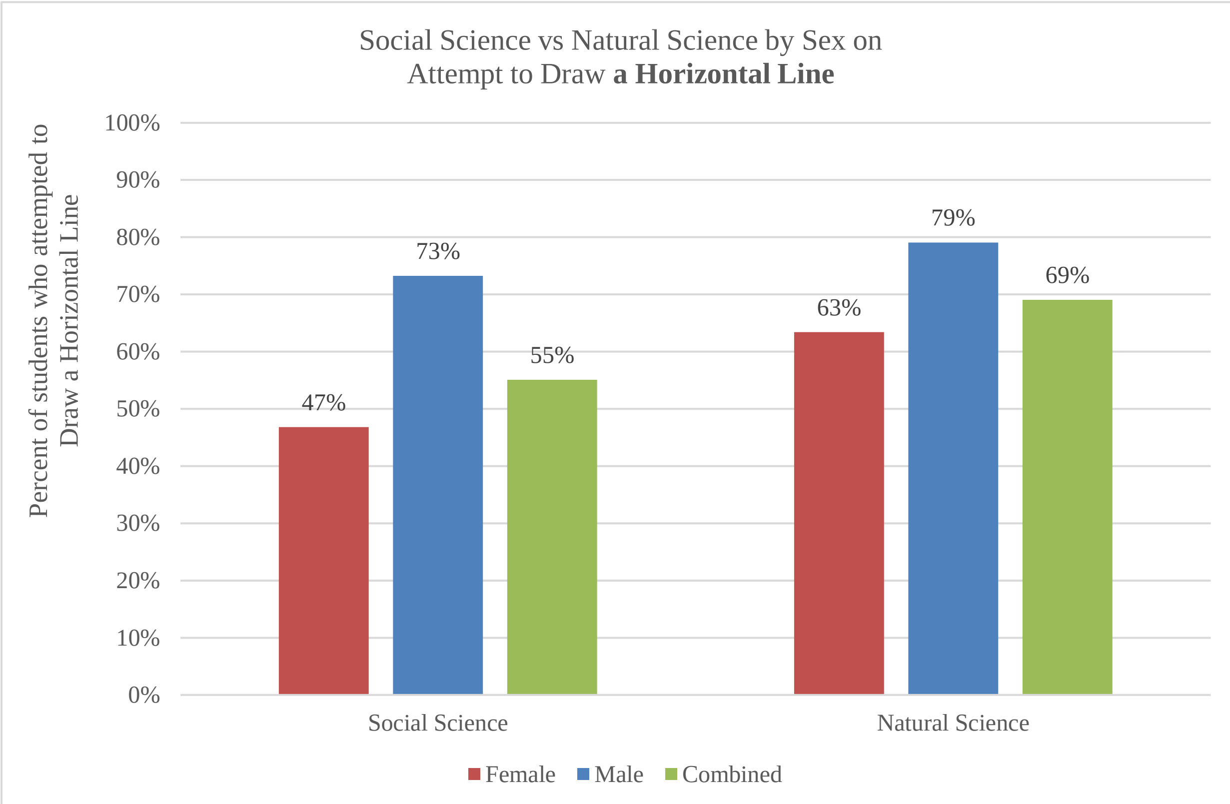


Figure 8 Panel A. When looking at all combined participants, a higher percentage of natural science majors (n=239, 69%) attempted to draw a horizontal line compared to social science majors (n=227, 55%) (Chi Square test, p=0.001; Table 2). Likewise, when looking at just females, a higher percentage of natural science majors attempted to draw a horizontal line (n=97, 63%) compared to social science majors (n=73, 47%). (Chi Square test, p=0.003; Table 2). Also, when looking at just males, a higher percentage of natural science majors attempted to draw a horizontal line (n=68, 79%) compared to social science majors (n=52, 73%). But this result is not significant (Chi Square test, p=0.392; Table 2). Within the social science majors, a higher percentage of males (n=52, 73%) attempted to draw a horizontal line compared to females (n=73, 47%) (Chi Square test, p<0.0005; Table 2). Within the natural science majors, a higher percentage of males attempted to draw a horizontal line (n=68, 79%) compared to females (n=97, 63%) (Chi Square test, p=0.012; Table 2).

(Figure 8 Continued on next page)

Panel B

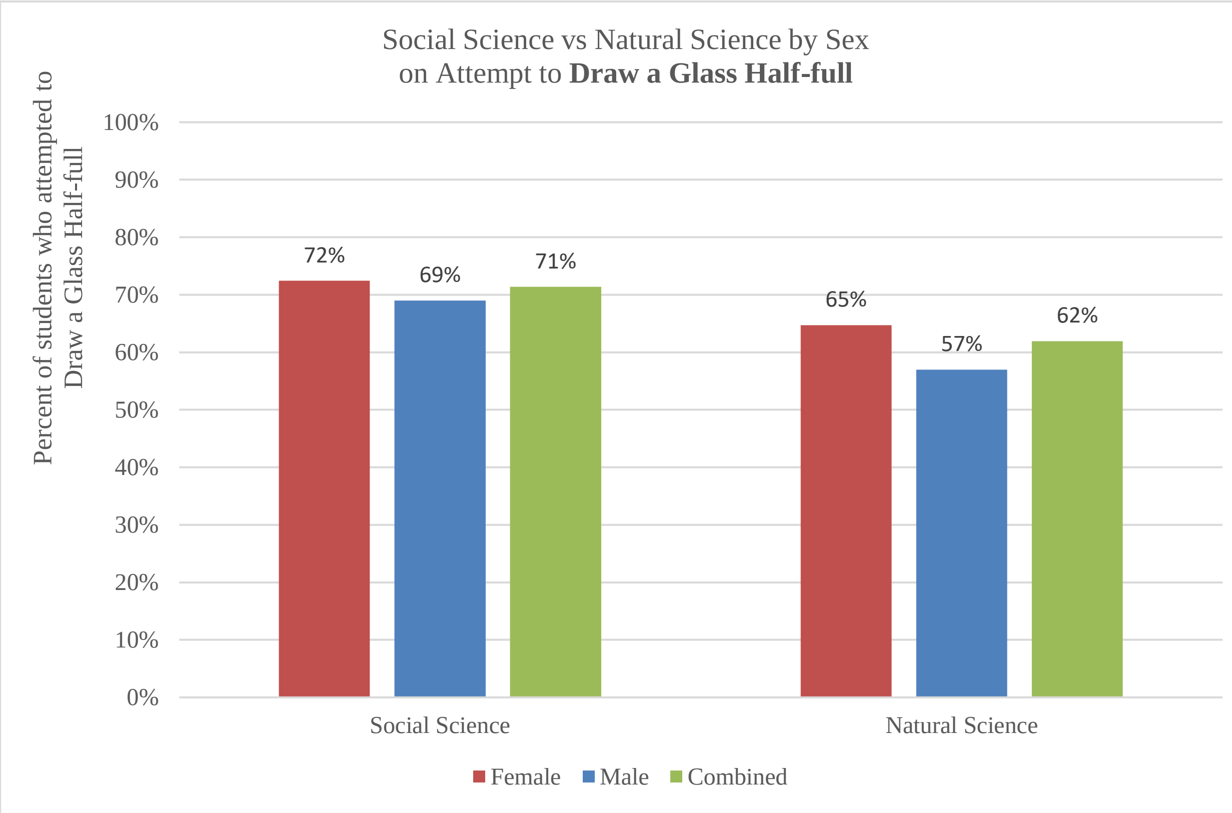


Figure 8 (continued from previous page) Panel B. When looking at **all combined** participants (in green), a higher percentage of social science majors (n=162, 71%) attempted to draw a glass half-full compared to natural science majors (n=148, 62%) **(Chi Square test, p=0.024; Table 2).** Similarly, when looking at **just females** (in red), a higher percentage of social science majors (n=113, 72%) attempted to draw a glass half-full compared to natural science majors (n=99, 65%), but the results were not significant (Chi Square test, p=0.14; Table 2). Likewise, when looking at **just males** (in blue)**,** a higher percentage of social science majors attempted to draw a glass half-full (69%) compared to natural science majors (n=49, 57%). This result was not significant (Chi Square test, p=0.12; Table 2). Within the **social science majors,** a higher percentage of females (n=113, 72%) attempted to draw a glass half-full compared to males (n=49, 69%) but it was not a significant effect (Chi Square test, p=0.597; Table 2). Within the **natural science majors,** a higher percentage of females attempted to draw a glass half-full (n=99, 65%) compared to males (n=49, 57%) but again it was not significant (Chi Square test, p=0.238; Table 2).

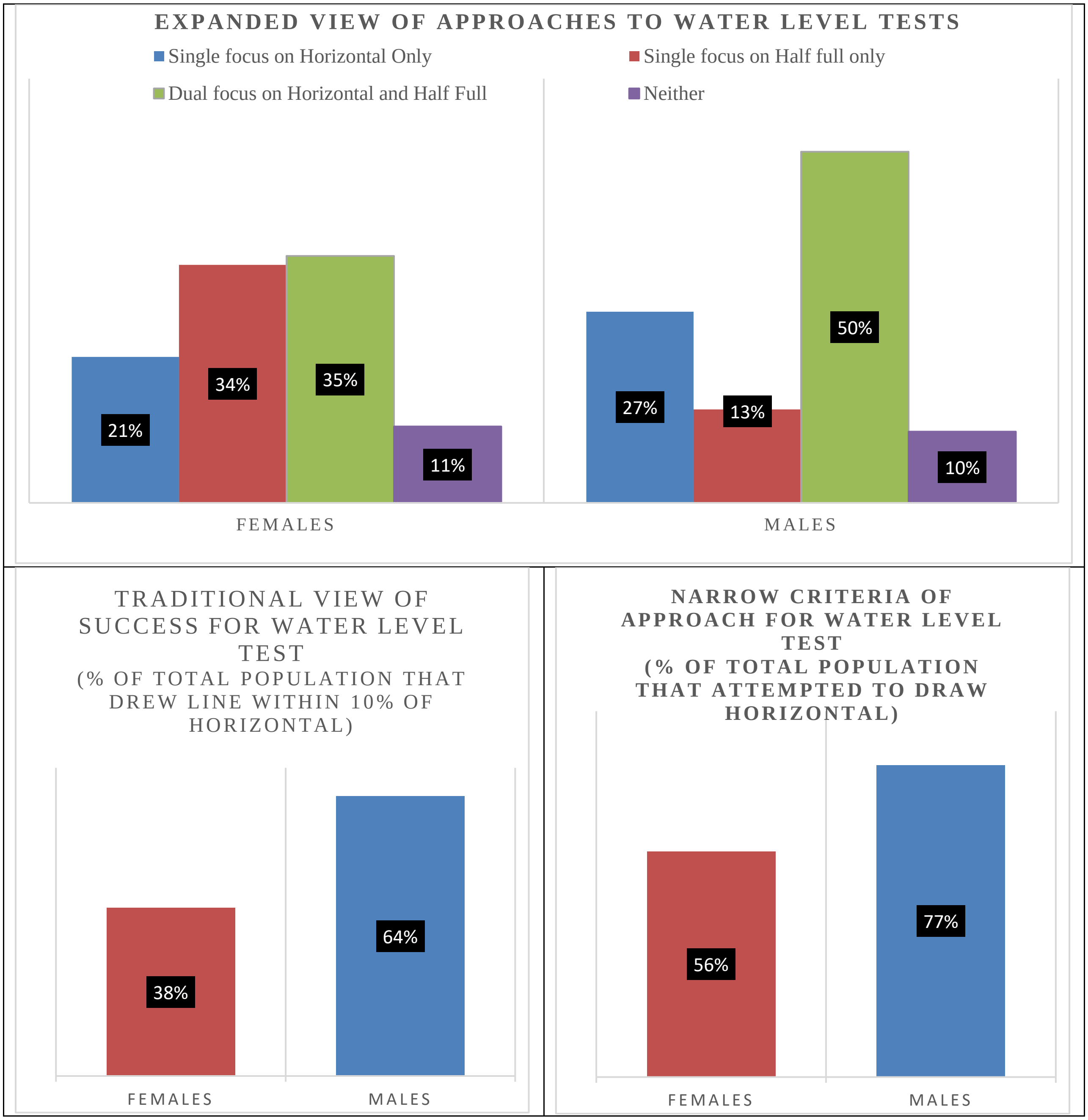


Figure 9 Panel A. Comparison of differences in approach by sex to water level task. Females and males both chose a dual approach more often than any other approach. **Panel B** Traditional Water Level Test results are reported as success and fail based on whether the line was horizontal, These results are significant ((Chi Square test, $p = \leq 0.0001$; Table 2). Panel C This is the same data as found in Figure 4A. This study initially looked at a narrower criteria of approaching the water level test. The effect of sex on approaching the task by attempting to draw a horizontal line is significant (Chi Square test, $p = \leq 0.00005$; Table 2).

List of self reported major/minor fields of study contained in the data	
Social Sciences	**Natural Sciences**
Advertising Behavioral Psychology Clinical Psychology Clinical Science Psychology	Biochemistry Biology Integrated Physiology Kinesiology Linguistics Molecular and Cellular Developmental Biology Neuroscience Pre-dental Pre-med Pre-nursing

Table 1 List of self reported major/minor fields of study contained in the data. If the participant had at least one of the Natural Sciences listed as either a major or minor, they were considered to be a Natural Science major, regardless of whether they also designated a social science field as an area of study.

Is there a significant effect of (independent variable):	On (dependent variable):	For population:	Inferential Statistical Test:	Statistic:	Probability Level:	Conclusion:
Sex	Accuracy (horizontal/not horizontal)	Everyone	Chi-Square	No statistic in excel	$p<.0001$	Significant
Sex	Accuracy (horizontal/not horizontal)	Natural Science Majors	Chi-Square	No statistic in excel	$p=0.003$	Significant
Sex	Accuracy (horizontal/not horizontal)	Natural Science Majors who attempted horizontal	Chi-Square	$X^2_1=2.596$	$p=0.107$	Not Significant
Sex	Accuracy (horizontal/not horizontal)	Social Science Majors	Chi-Square	No statistic in excel	$p<.0001$	Significant
Sex	Accuracy (horizontal/not horizontal)	Social Science Majors who attempted horizontal	Chi-Square	$X^2_1=3.577$	$p=0.059$	Not Significant
Sex	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	$X^2_1=7.242$	$p=0.007$	Significant
Sex	Angle	Natural Science Majors who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	$t(163)=1.194$ $t(150.810)=1.210$	$p=0.234$ $p=0.228$	Not Significant Not Significant
Sex	Angle	Social Science Majors who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	$t(122)=0.935$ $t(116.654)=0.953$	$p=0.352$ $p=0.343$	Not Significant Not Significant
Sex	Angle	Those who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	$t(294)=1.652$ $t(275.38)=1.681$	$p=0.100$ $p=0.094$	Not Significant Not Significant
Sex	Four Approaches	Everyone	Chi-Square	No statistic in excel	$p<0.0001$	Significant
Sex	Four Approaches	Natural Science Majors	Chi-Square	No statistic in excel	$p=0.002$	Significant
Sex	Four Approaches	Social Science Majors	Chi-Square	No statistic in excel	$p<0.0001$	Significant
Sex	Approach: (Horizontal and/or Half full) vs Neither	Everyone	Chi-Square	No statistic in excel	$p=0.176$	Not Significant
Sex	Whether they attempt a glass half-full	Everyone	Chi-Square	No statistic in excel	$p=0.216$	Not Significant
Sex	Whether they attempt a glass half-full	Natural Science Majors	Chi-Square	No statistic in excel	$p=0.238$	Not Significant
Sex	Whether they attempt a glass half-full	Natural Science Majors that did not attempt horizontal	Chi-Square	No statistic in excel	$p=0.016$	Significant
Sex	Whether they attempt a glass half-full	Social Science Majors	Chi-Square	No statistic in excel	$p=0.597$	Not Significant
Sex	Whether they attempt a glass half-full	Social Science Majors that did not attempt horizontal	Chi-Square	No statistic in excel	$p=0.500$	Not Significant
Sex	Whether they attempt a glass half-full	Those who did not attempt horizontal	Chi-Square	No statistic in excel	$p=0.023$	Significant
Sex	Whether they attempt a horizontal line	(Majorcode=2 or 3 or 4) or (Majorcode=1 and second major = 2)	Chi-Square	$X^2_2=7.292$	$p=0.026$	Significant
Sex	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_1=20.247$	$p<0.0005$	Significant
Sex	Whether they attempt a horizontal line	Majorcode=1 AND (Second major = 0 or 3)	Chi-Square	$X^2_2=13.546$	$p=0.001$	Significant
Sex	Whether they attempt a horizontal line	Natural Science Majors	Chi-Square	$X^2_1=6.325$	$p=0.012$	Significant
Sex	Whether they attempt a horizontal line	Social Science Majors	Chi-Square	$X^2_1=13.790$	$p<0.0005$	Significant
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Everyone	Chi-Square	No statistic in excel	$p=0.002$	Significant
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Females	Chi-Square	No statistic in excel	$p=0.004$	Significant
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Females who attempted horizontal	Chi-Square	$X^2_1=1.230$	$p=0.267$	Not Significant

Table 2 Results from inferential statistical tests (continued on next page)

Is there a significant effect of (independent variable):	**On (dependent variable):**	**For population:**	**Inferential Statistical Test:**	**Statistic:**	**Probability Level:**	**Conclusion:**
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Males	Chi-Square	No statistic in excel	p=0.357	Not Significant
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Males who attempted horizontal	Chi-Square	X^2_1=0.190	p=0.663	Not Significant
Majorcode 4 (Social science vs natural science)	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	X^2_1=1.340	p=0.247	Not Significant
Majorcode 4 (Social science vs natural science)	Angle	Females who attempted horizontal	ANOVA	F(1,167)=0.669	p=0.414	Not Significant
Majorcode 4 (Social science vs natural science)	Angle	Males who attempted horizontal	ANOVA	F(1,118)=0.908	p=0.343	Not Significant
Majorcode 4 (Social science vs natural science)	Angle	Those who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	t(288)=1.256 t(276.282)=1.272	p=0.210 p=0.205	Not Significant Not Significant
Majorcode 4 (Social science vs natural science)	Four Approaches	Everyone	Chi-Square	No statistic in excel	p<0.0002	Significant
Majorcode 4 (Social science vs natural science)	Four Approaches	Females	Chi-Square	No statistic in excel	p=0.003	Significant
Majorcode 4 (Social science vs natural science)	Four Approaches	Males	Chi-Square	No statistic in excel	p=0.062	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	All Females	Chi-Square	No statistic in excel	p=0.143	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	All Males	Chi-Square	No statistic in excel	p=0.121	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	Everyone	Chi-Square	No statistic in excel	p=0.024	Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	Everyone that did not attempt horizontal	Chi-Square	No statistic in excel	p=0.295	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	Females that did not attempt horizontal	Chi-Square	No statistic in excel	p=0.903	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a glass half-full	Males that did not attempt horizontal	Chi-Square	No statistic in excel	p=0.141	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a horizontal line	All females	Chi-Square	X^2_1=8.605	p=0.003	Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a horizontal line	All males	Chi-Square	X^2_1=0.734	p=0.392	Not Significant
Majorcode 4 (Social science vs natural science)	Whether they attempt a horizontal line	Everyone	Chi-Square	X^2_1=10.227	p=0.001	Significant
Handedness	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	X^2_1=0.786	p=0.375	Not Significant
Handedness	Angle	Everyone	t-test(Equal Variances) t-test (Unequal Variances)	t(440)=2.017 t(49.18)=2.206	p=0.044 p=0.032	Significant Significant

Table 3 (continued) Results from inferential statistical tests (continued on next page)

Is there a significant effect of (independent variable):	On (dependent variable):	For population:	Inferential Statistical Test:	Statistic:	Probability Level:	Conclusion:
Handedness	Angle	Those who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	$t(287)=1.489$ $t(41.70)=1.749$	$p=0.138$ $p=0.088$	Not Significant Not Significant
Handedness	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_1=2.576$	$p=0.108$	Not Significant
Handedness	Whether they attempt a glass half-full	Everyone	Chi-Square	No statistic in excel	$p=0.975$	Not Significant
Age	Accuracy (horizontal/not horizontal)	Females only for attempted horizontal only	Chi-Square	$X^2_5=5.188$	$p=0.393$	Not Significant
Age	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	$X^2_5=5.799$	$p=0.326$	Not Significant
Age	Angle	Those who attempted horizontal	ANOVA	$F(5,289)=0.182$	$p=0.969$	Not Significant
Age	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_5=1.695$	$p=0.890$	Not Significant
Class in School	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	$X^2_6=7.624$	$p=0.267$	Not Significant
Class in School	Angle	Those who attempted horizontal	ANOVA	$F(3,289)=0.358$	$p=0.783$	Not Significant
Class in School	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_3=3.837$	$p=0.280$	Not Significant
Class in School	Whether they attempt a horizontal line	Females Only	Chi-Square	$X^2_3=3.157$	$p=0.368$	Not significant
Class Prefix	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	$X^2_1=0.024$	$p=0.877$	Not Significant
Class Prefix	Angle	Those who attempted horizontal	t-test(Equal Variances) t-test (Unequal Variances)	$t(295)=0.762$ $t(240.259)=0.740$	$p=0.447$ $p=0.460$	Not Significant Not Significant
Class Prefix	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_1=4.797$	$p=0.029$	Significant
Major	Accuracy (horizontal/not horizontal)	Those who attempted horizontal for majorcode 1-4	Chi-Square	$X^2_3=1.706$	$p=0.636$	Not Significant
Major	Angle	Those who attempted horizontal	ANOVA	$F(3,281)=1.055$	$p=0.369$	Not Significant
Major	Whether they attempt a horizontal line	Majorcode groups 1-2 only	Chi-Square	$X^2_1=4.861$	$p=0.027$	Significant
Major	Whether they attempt a horizontal line	Majorcode groups 1-4	Chi-Square	$X^2_3=7.331$	$p=0.062$	Not Significant
Year	Accuracy (horizontal/not horizontal)	Those who attempted horizontal	Chi-Square	$X^2_{11}=19.231$	$p=0.057$	Not Significant
Year	Angle	Those who attempted horizontal	ANOVA	$F(11,285)=3.712$	$p<0.0005$	Significant
Year	Whether they attempt a horizontal line	Everyone	Chi-Square	$X^2_{11}=13.010$	$p=0.293$	Not Significant

Table 2 (continued) Results from inferential statistical tests

References:

Cahill, L. (2014). Equal ≠ the same: Sex differences in the human brain. *Cerebrum*, Mar-Apr; 2014: 5.

Carothers, B. J., & Reis, H. T. (2013). Men and women are from Earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology*, Vol. 104, No. 2, 385–407.

Collaer, M. L., & Hines, M. (1995). Human Behavioral Sex Differences: A Role for Gonadal Hormones During Early Development? *Psychological bulletin*, Vol. 118, No. 1,55-107.

Cunningham, B. C., Hoyer, K. M., & Sparks, D. (2015). *Gender Differences in Science, Technology, Engineering and Mathematics (STEM): Interest, Credits Earned, and NAEP Performance in the 12th Grade.* Washington, D.C.: U.S. Department of Education.

Fine, C. (2017). *Testosterone Rex: Myths of Sex, Science and Society.* New York, New York: W. W. Norton and Company.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest*, 8(1): 1–51.

Hecht, H., & Proffitt, D. R. (1995). THE PRICE OF EXPERTISE: Effects of Experience on the Water-Level !ask. *PSYCHOLOGICAL SCIENCE*, Vol. 6, No. 2, March 1995 90-95.

Hyde, J. S. (2005). The Gender Similarities Hypothesis. *American Psychologist*, Vol. 60, No. 6, 581–592.

Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., . . . Urchs, S. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50): 15468–15473.

Kalichman, S. C. (1986). HORIZONTALITY AS A FUNCTION OF SEX AND ACADEMIC MAJOR. *Perceptual and Motor Skills*, vol. 63, 2: pp. 903-906.

Kalichman, S. C. (1988). Individual Differences in Water-Level Task Performance: A Component-Skills Analysis. *DEVELOPMENTAL REVIEW*, 8, 273-295 (1988).

Liben, L. (1978). PERFORMANCE ON PIAGETIAN SPATIAL TASKS AS A FUNCTION OF SEX, FIELD DEPENDENCE, AND TRAINING. *Merrill-Palmer Quarterly of Behavior and Development*, Vol. 24, No. 2, pp. 97-110.

Liben, L. (1991). The Piagetian water-level task: Looking beneath the surface. *Annals of child development: Vol. 8*, 81-144.

Nelson, R. J. (2011). Sex Differences: Animal Models and Humans. In R. J. Nelson, *An Introduction to Behavioral Endocrinology* (p. 174). Sunderland, MA: Sinauer Associates, Inc. doi:10.1002/dev.10039

Papadatou-Pastou, M., Martin, M., Munafò, M., & Jones, G. (2008). Sex differences in left-handedness: a meta-analysis of 144 studies. *Psycholgy Bulletin*, Sep 134(5):677-699.

Papadatou-Pastou, M., Ntolka, E., Schmitz, J., Martin, M., Munafò, M. R., Ocklenburg, S., & Paracchini, S. (2020). Human handedness: A meta-analysis. *Psycholgy Bulletin*, 146(6), 481–524.

Pascual-Leone, J., & Morra, S. (1991). Horizontality of Water level: A neo-Piagetian developmental review. *Advances in child development and behavior*, Vol 23 (pp. 231-276).

Piaget, J., & Inhelder, B. (1956). *The child's conception of space.* London: W. W. Norton.

Robert, M., & Morin, P. (1993). Gender differences in horizontality and verticality representation in relation to initial position of the stimuli. *Canadian Journal of Experimental Psychology,* (Sep 1993): 507-522.

Roberts, J. E., & Bell, M. A. (2002). The Effects of Age and Sex on Mental Rotation Performance, Verbal Performance, and Brain Electrical Activity. *Developmental Psychobiology*, 391-407.

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of Assessing Spatial Ability in Intellectually Talented Young Adolescents: A 20-Year Longitudinal Study. *Journal of Educational Psychology*, 604-614. doi:10.1037/0022-0663.93.3.604

Sholl, M. J., & Liben, L. S. (1995). Illusory Tilt and Euclidean Schemes as Factors in Performance on the Water-Level Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 21, No. 6,1624-1638.

Swirsky-Sacchetti, T., & Genetta-Wadley, A. (1990). Sex Differences and Handedness in Hemispheric Lateralization of Tactile-Spatial Functions. *Perceptual and Motor Skills*, 70, 579-590.

Vasta, R., & Liben, L. S. (1996). The Water-Level Task: An Intriguing Puzzle. *Current directions in psychological science : a journal of the American Psychological Society*, Volume 5, Issue 6, 171-177.

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*, Vol. 117, No. 2,250-270.