Received: 28 March 2017 | Revised: 27 April 2017 | Accepted: 10 May 2017
DOI: 10.1002/ajmg.c.31565

RESEARCH ARTICLE

WILEY AMERICAN JOURNAL OF medical genetics Part C Seminars in Medical Genetics

Similar reliability and equivalent performance of female and male mice in the open field and water-maze place navigation task

Ann-Kristina Fritz[1,2] | Irmgard Amrein[1,2] | David P. Wolfer[1,2]

[1] Institute of Anatomy, University of Zurich, Zurich, Switzerland

[2] Institute of Human Movement Sciences and Sport, ETH Zurich, Zurich, Switzerland

Correspondence
David P. Wolfer, Institute of Anatomy, University of Zurich, Winterthurerstrasse 190/42J6, CH-8057 Zurich, Switzerland.
Email: davidp.wolfer@uzh.ch

Funding information
Swiss National Science Foundation; 6th and 7th Framework Programmes of the European Union

Although most nervous system diseases affect women and men differentially, most behavioral studies using mouse models do not include subjects of both sexes. Many researchers worry that data of female mice may be unreliable due to the estrous cycle. Here, we retrospectively evaluated sex effects on coefficient of variation (CV) in 5,311 mice which had performed the same place navigation protocol in the water-maze and in 4,554 mice tested in the same open field arena. Confidence intervals for Cohen's d as measure of effect size were computed and tested for equivalence with 0.2 as equivalence margin. Despite the large sample size, only few behavioral parameters showed a significant sex effect on CV. Confidence intervals of effect size indicated that CV was either equivalent or showed a small sex difference at most, accounting for less than 2% of total group to group variation of CV. While female mice were potentially slightly more variable in water-maze acquisition and in the open field, males tended to perform less reliably in the water-maze probe trial. In addition to evaluating variability, we also directly compared mean performance of female and male mice and found them to be equivalent in both water-maze place navigation and open field exploration. Our data confirm and extend other large scale studies in demonstrating that including female mice in experiments does not cause a relevant increase of data variability. Our results make a strong case for including mice of both sexes whenever open field or water-maze are used in preclinical research.

KEYWORDS
anxiety, exploration, learning and memory, mouse model, sex differences

1 | INTRODUCTION

Since 1993, the US National Institutes of Health require the inclusion of women in clinical research funded by them. Since 2014, they implement policies that oblige applicants to report their plans for the balance of male and female cells and animals in preclinical studies (Clayton & Collins, 2014). Most nervous system diseases, including multiple sclerosis, Parkinson's disease, schizophrenia, autism, depression, and dementia affect women and men differentially with respect to prevalence, severity, or disease course (Christensen et al., 2016; Golden & Voskuhl, 2017; Haaxma et al., 2007; Kokras & Dalla, 2017; Leger & Neill, 2016; Tschanz et al., 2011). So, it would seem adequate for preclinical research using mouse models of these diseases to include subjects of both sexes in all experiments. However, this is still

not common practice (Kokras & Dalla, 2017; Leger & Neill, 2016; Zucker & Beery, 2010). Even in 2015, of the 71 research articles that used rodents published in the journal Pain, only 3 affirmed the use of both sexes (Mogil, 2016). Our own informal PubMed survey of 100 studies published in 2016 that used the water-maze place navigation task to study spatial learning and memory of mice revealed a similar result. Only 12 studies had included subjects of both sexes with 6 stating that they had added sex as a factor to the statistical model. A total of 8 studies had used only female subjects and 69 only males. There are several reasons why researchers are so reluctant to include female mice in their studies (Mogil, 2016; Prendergast, Onishi, & Zucker, 2014), the most important being the worry that due to their estrous cycle, inclusion of female mice in a study will increase variability of results and necessitate the testing of more subjects. It is often believed that females must be tested at each stage of the estrous cycle to generate reliable data. For tests of pain sensitivity, this has been disproved by a retrospective analysis in 7,875 mice (Mogil & Chanda, 2005), clearly demonstrating that data are equally reliable in female and male mice—also if females are tested at random stages of the estrous cycle. In addition, the study showed that, reproducing human sex differences in pain sensitivity, female mice overall had significantly shorter tail-withdrawal latencies. A recent meta-analysis extracted variability measures from 293 published articles, in which female mice tested at random stages of the estrous cycle were compared with males with regard to behavioral, physiological, morphological, and molecular traits (Prendergast et al., 2014). The study concluded that variability was not significantly greater in females than males for any endpoint and was substantially greater in males for several traits.

Even in 2015, of the 71 research articles that used rodents published in the journal Pain, only 3 affirmed the use of both sexes. Our own informal PubMed survey of 100 studies published in 2016 that used the water-maze place navigation task to study spatial learning and memory of mice revealed a similar result.

To our knowledge, retrospective behavioral studies as rigorous as Mogil and Chanda (2005) have not been conducted in other domains of behavior. In particular, comparable analyses are lacking for the important domains of learning and memory as well as for exploration and anxiety. The water-maze place navigation task (Morris, 1981; Morris, Garrud, Rawlins, & O'Keefe, 1982) is one of the most common behavioral tests for assessing spatial learning and memory in mice (D'Hooge & De Deyn, 2001; Schoenfeld, Schiffelholz, Beyer, Leplow, & Foreman, 2017; Wolfer, Colacicco, & Welzl, 2013). The open field test is very often used to measure activity, exploration, and anxiety-related responses in a novel environment. Both tests have been used during several decades in our laboratory with standardized protocols. This has accumulated a large body of data which we use in this report for a systematic retrospective analysis.

Laboratory mouse strains were once derived mostly from mus musculus domesticus (Frazer et al., 2007; Yang, Bell, Churchill, & Pardo-Manuel de Villena, 2007; Yang et al., 2011) through selective breeding for morphological and behavioral traits. Given that free living male mus musculus have larger territories and venture farther away than females (Chambers, Singleton, & Krebs, 2000; Pocock, Hauffe, & Searle, 2005; Pocock, Searle, & White, 2004), one would predict sex differences in spatial learning and exploratory behavior also in laboratory tests. But while literature reports indicate large reliable male advantages for rats in radial-maze and water-maze protocols (Jonasson, 2005), findings have remained contradictory in laboratory mice (Frick, Burlingame, Arters, & Berger-Sweeney, 2000; Ge, Qi, Qiao, Wang, & Zhou, 2013; Hendershott, Cronin, Langella, McGuinness, & Basu, 2016; Jonasson, 2005; Voikar, Koks, Vasar, & Rauvala, 2001). Consistent reports on sex differences are also lacking for open field exploration (Ge et al., 2013; Voikar et al., 2001).

The large number of subjects available for the present study provides not only sufficient power to detect small differences but also permits to test for equivalence. Therefore, in our retrospective analysis we addressed two questions: (i) is the behavior of female mice in the water-maze and open field equally reliable or more variable than that of males? (ii) Do male mice perform better in the place navigation task and are they more explorative in the open field?

2 | METHODS

2.1 | Animals

This study is a retrospective analysis of behavioral data collected at the Institute of Anatomy of the University of Zurich during the years 1991–2016 using standardized protocols by laboratory technicians, scientists, and students. The mice used for the experiments were either bred at the Institute of Anatomy or brought there at least one week before behavioral testing. Generally, mice were housed under a 12/12 hr light–dark cycle (lights on at 20:00) in groups of 2–5, unless individual housing was required by experimental protocols or to prevent fighting. Testing occurred during the dark phase under dim light (~12 lux). Mice were transferred to the testing room at least

30 min before testing. All procedures were approved by the Veterinary Office of the Canton of Zurich. Most of the animals had also been tested in other behavioral tests, but water-maze and open field were typically performed at the beginning of the test battery. Part of the data have already been published elsewhere but with different analyses not focusing on sex differences.

Studies included in the analysis were performed with more than five mice of each sex and experimental condition tested concurrently in the same standard water-maze and open field protocols. Female mice were not tracked for estrus cyclicity, and thus, tested at random points of their cycle. Of 185 water-maze place navigation experiments performed with the standard protocol, 125 had a sex composition suitable for the present study. A total of 26 experiments had tested only females, 15 only male mice, 19 comprised subject of both sexes but in numbers unsuitable for the planned analyses. Of 166 open field experiments, 104 included balanced numbers of both sexes and were thus included in the present analysis. Of 20 open field experiments had included only female, 28 only male mice, 14 had tested individuals of both sexes but in insufficient numbers and were excluded as well.

Reflecting the scientific focus of the lab, most animals included in this retrospective analysis were from experiments using genetically modified mouse models to study normal function and diseases of the brain. A number of strain comparison experiments were included as well. Table 1 gives an overview of the study population. About 70% of the animals had a heterogeneous genetic background, mostly F2-3 or partial backcross generations of crosses between substrains of C57BL/6 and 129. Thirty percent were either inbred (mostly C57BL/6) or F1 hybrids. Median age of the animals was 3.4 months with an interquartile range of 2.5–4.7 months. Six percent of the mice had an age of 12 months or older.

2.2 | Behavioral procedures

2.2.1 | Water-maze place navigation

The round white poly-propylene pool had a diameter of 150 cm with 68 cm high walls. It was filled with water (24–26°C, depth 15 cm) which was rendered opaque by addition of 1 L of milk (UHT whole milk 3.5% fat, Coop., Switzerland). The white quadratic goal platform (14 × 14 cm) was made of metallic wire mesh and painted white. It was hidden 0.5 cm below the water surface in the center of one of the four quadrants, approximately 30 cm from the side wall. Salient extra-maze cues were placed on the walls of the testing room. Animals performed 30 training trials (max. duration 120 s), 6 per day with intertrial intervals of 30–60 min and varying starting positions. During the first 18 trials the hidden platform was held in the same position (acquisition phase) and then moved to the opposite quadrant for the remaining 12 trials (reversal phase). The first 60 s of the first reversal trial served as probe trial to test for spatial retention.

2.2.2 | Open field

The round arena had a diameter of 150 and 35 cm high sidewalls made of white polypropylene. Each subject was released near the wall and observed for 10 min on 2 subsequent days.

2.2.3 | Video-tracking

During all experiments, the path of moving mice was tracked using a video-tracking system. Due to the extended data collection period, different generations of systems were used: ASBA Wild & Leitz (Basel, Switzerland) 1991–1996, Noldus EthoVision 1996–2016 (Wageningen, Netherlands, versions 1.96 through XT11.5). For analysis all data were imported in

TABLE 1 Description of the study population

	Water-maze			Open field		
Type of study	Female	Male	Studies	Female	Male	Studies
GM mouse models with mutations affecting						
βAPP and other proteins related to dementia (gain and loss of function)	660	732	32	719	700	33
Prion protein (loss of function)	143	144	4	164	121	5
Kinases, calcium-binding proteins, guanine nucleotide exchange factors	358	320	19	289	261	15
N-CAM and related proteins, ephrins and receptors	208	192	12	180	171	9
Glutamate and GABA receptors; transmitter synthesis, release, uptake	136	113	8	66	68	4
Glucocorticoid receptors, growth factors and receptors (BDNF, CNTF, VEGF)	224	201	10	260	257	9
Extracellular proteases and their inhibitors	287	308	12	265	297	10
Circadian clock, cytoskeleton, metabolis, ribosomal proteins	233	251	13	198	228	12
Other studies						
Lesions, toxicological studies, hypoxia	131	136	6	76	87	4
Lab strains, crosses, selective breeding	278	256	9	86	61	3
Total N	2,658	2,653	125	2,303	2,251	104
Median N per study	17	17		18.5	18	

custom programmed software Wintrack (www.wintrack.ch) (Wolfer, Madani, Valenti, & Lipp, 2001) and converted to a consistent format.

2.3 | Behavioral measures

2.3.1 | Water-maze place navigation, training

Escape performance during training was assessed by computing the standard measures escape latency, swim path length, and swim speed (excluding floating episodes). Spatial orientation was further evaluated using cumulative search error (sum of distances to target measured at 1 s intervals minus value that would be obtained for an ideal direct swim) (Gallagher, Burwell, & Burchinal, 1993), Whishaw's error adapted to mice and pool size (% path outside a 0.1856 m wide corridor connecting release point and goal) (Whishaw, 1985), path efficiency (% path during which speed vector component toward goal is 75% or more) (Poirier et al., 2007), and average distance from target (proximity). Further, we evaluated time spent floating (episodes of immobility or decelerations with speed minimum <0.06 m/s), the number of times mice jumped off the goal platform instead of climbing and remaining on it, the time they spent within a 7 cm zone from the wall as a measure of thigmotaxis, and the number of times they approached and contacted the wall. Body weight was recorded before and after training in a subset of animals. Body weight is known to be strongly and reliably sex-dependent in mice. Therefore, we included this measure as a reference for comparison with other sex effects. To facilitate analysis of the large study sample, all measures were averaged across training trials to obtain a single value per subject.

2.3.2 | Water-maze, probe trial

Spatial retention was assessed using % time spent swimming in the former goal quadrant and in a circular target zone comprising 12.5% of the pool surface, annulus crossing index (crossings over target minus average of crossings over annuli in adjacent quadrants divided by distance swum), proximity (average distance to trained target) (Gallagher et al., 1993), and polar error (average angle between two lines connecting the pool center to the subject and trained target, respectively).

2.3.3 | Open field

Locomotor activity was evaluated using the standard measure of distance moved. To further characterize locomotion, we segmented the recorded path into bouts of walking (>8.5 cm/s) and episodes of lingering or resting (Madani et al., 2003). This permitted to determine % time walking and walking speed. Anxiety related parameters were % time spent in the center field (50% of area) and average distance to center. To assess the efficiency of exploration, we divided the arena into quadratic tiles of 7 × 7 cm and determined the % of tiles in which the mice had shown lingering activity. When moving toward the center mice typically walk slowly and accelerate considerably when returning toward the wall. Therefore we also calculated the difference between out and inbound speed. To facilitate analysis of the large study sample, all measures were averaged across days to obtain a single value per subject. Counts of fecal boli deposited in the arena were available in a subset of animals.

2.4 | Statistical analysis

Individual measurements were evaluated using a factorial two-way ANOVA model including the between-subject factors experimental group and sex. The experimental group factor captured treatment and genotype effects as well as baseline variation between experiments. It was only included to reduce unexplained variance and was not further evaluated nor shown in the figures. An experimental group was defined as a subset of mice tested in the same experiment and sharing the same experimental condition (mutation genotype/treatment, genetic background, age). Each experimental group contained two similarly sized subgroups of male and female mice. Most experiments had two experimental groups, a treatment and a control group. Some experiments included additional treatment groups. Numbers of subjects and groups are detailed in Table 1 and Figure 1. To evaluate size and significance of the sex effect in the model in a comparable way across different variables, we computed the 95% confidence interval for Cohen's d (CId, mean difference male-female divided by the pooled standard deviation). Effects were considered significant if CId did not spread over zero. Group differences are generally considered negligible if $|d| < 0.2$. Thus, we tested for equivalence with 0.2 as equivalence margin, considering equivalence of female and male scores established if CId of the sex effect was within the interval (−0.2, 0.2), that is differences between males and females amounted to less than one fifth of the standard deviation. According to common practice, effects with $|d| > 0.2$ were classified as small, medium, and large as shown in Figure 1. Effects with $0.20 < |d| < 0.25$ were considered marginal.

To compare the variability of measurements between sexes, we calculated the coefficient of variation (CV, standard deviation divided by mean) separately for females and males in each group. Sex differences of CV were then tested using a factorial one-way ANOVA model with female and male subgroups as the unit of observation. To determine size and significance of the sex effect on CV, we again computed the 95% confidence interval for Cohen's d and tested for equivalence with 0.2 as equivalence margin.

As to be expected given the very large sample sizes, Kolmogorov–Smirnov tests detected highly significant deviations from normality ($p < 0.001$) for all measures. Therefore we applied Box-Cox transformations to all variables with λ parameters optimized by the R boxcox function and computed the entire statistical analysis with both raw and transformed data. Because both analyses yielded very similar results and led to identical conclusions, we present only the analysis based on untransformed data. All statistical analyses and graphs were produced using R version 3.2.3 (R Development Core Team, 2008),

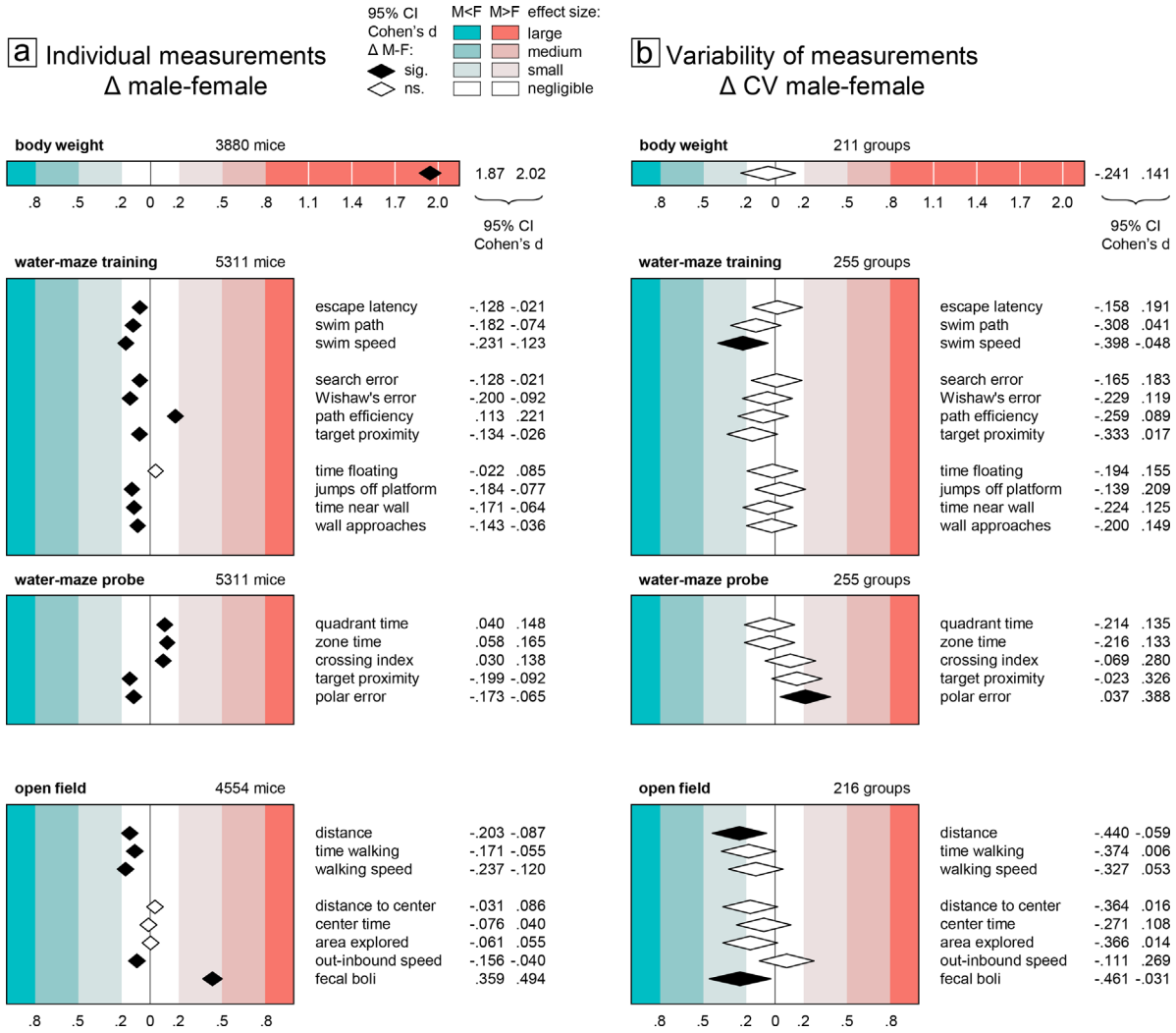

FIGURE 1 Sex differences of body weight, water-maze, and open field parameters and their coefficient of variation, visualized using the 95% confidence interval of Cohen's d as standardized measure of effect size

complemented with the packages effsize and ggplot2 (Wickham, 2009).

3 | RESULTS

3.1 | Body weight

Body weight measurements were available for 211 of the 255 experimental groups (3,880 of 5,311 mice) that had been tested in the water-maze place navigation task. Even though the distributions of male and female body weight values showed some overlap (Figure 2a), statistical analysis as expected revealed a significant and very large sex difference with higher body weight in males (Figure 1a top). In all but four experimental groups, male mice were heavier than females (Figure 2b), making this effect highly reliable across experiments. By contrast, analysis of CV detected no significant sex difference with regard to variability of body weight (Figure 1b top). There was a potential bias toward more variability in females with a lower CId boundary of $d > -0.25$ indicating that if present at all the difference in variability was marginal at most. Confirming this, body weight CV values showed virtually identical distributions across male and female subgroups (Figure 2c). Male minus female CV differences of experimental groups were mostly near 0 and symmetrically distributed with larger variability in females in about half of the groups (Figure 2d).

3.2 | Water-maze place navigation, training

As judged by escape latency and swim path length, male mice showed significantly better overall escape performance than females during training in the water-maze place navigation task. But despite the statistical significance of the effect, evaluation of CId established equivalence of the sexes for both measures (Figure 1a middle). Accordingly, the distributions of individual male and female escape latencies were nearly congruent (Figure 3a) and the effect was highly unreliable across experiments with female mice having longer escape latencies only in 56% of experimental groups (Figure 3b). Female mice

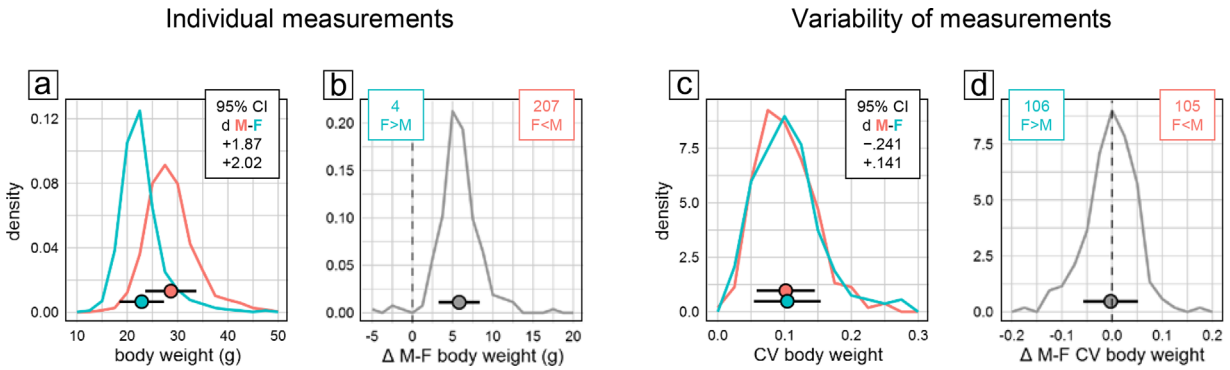

FIGURE 2 Frequency polygons illustrating sex differences of body weight values (a and b) and their coefficient of variation CV as a measure of variability (c and d). Points and horizontal bars represent mean and standard deviation. (a) Distribution of individual values of 1,965 male (red) and 1,954 female (cyan) mice. Inset shows the 95% confidence interval for Cohen's d as an effect size estimate for the overall male–female difference (absolute value: <0.2 negligible, 0.5 medium, >0.8 large). (b) Distribution of male–female mean differences in 215 pairs of groups. Pairs with larger mean in males are to the right of the dashed line, those with larger mean in females to the left. (c) CV distribution in 215 male (red) and 215 female (cyan) groups. Inset shows the 95% confidence interval for Cohen's d as an effect size estimate for the overall male–female CV difference. (d) Distribution of the male–female CV difference in 215 pairs of groups. Pairs with higher variability in males are to the right of the dashed line, those with higher variability in females to the left

swam significantly faster than males (Figure 1a middle), however, with a lower CId boundary of $d > -0.24$ indicating that the effect was marginal at most. The distributions of female and male speed values were indeed very similar and the effect was still rather unreliable across experiments with female mice swimming faster than males in 62% of experimental groups (Figure 3e,f). Search error, Wishaw's error, path efficiency, and target proximity, are measures designed to quantify spatial selectivity and orientation during place navigation training. According to all three, male mice were significantly more selective than females. But despite the statistical significance of the effect, evaluation of CId established equivalence of the sexes for three of the measures (Figure 1a middle). For path efficiency, an upper CId boundary of $d < 0.23$ indicated that the male advantage was marginal at most. In accordance with this, Whishaw's error showed broadly overlapping individual scores of male and female mice (Figure 3i) with males earning better scores only in 61% of experimental groups (Figure 3j).

Time floating, jumps off the platform, time near wall, and wall approaches are measures that help to assess how well mice adapt emotionally to the stressful test situation in the water-maze and how flexible they are in exploring different escape strategies. While passive floating showed no significant sex difference, female mice jumped off the platform significantly more often and were also significantly more oriented toward the wall (Figure 1a middle). However, evaluation of CId established equivalence of the sexes for all four measures (Figure 1a middle), indicating that female and male mice adapted equally well to the test situation.

Evaluation of CV as a measure of variability within experimental groups provided no evidence for significant sex differences in any measure of training performance. Only swim speed was significantly more variable in female mice (Figure 1b middle), with the sex effect accounting for less than 2% of total variance ($\eta^2 = 0.01234$) and a lower CId boundary of $d > -0.40$ indicating that the effect was small at most. Based on CId, equivalence of variability could be established for four measures (escape latency, search error, time floating, wall approaches). For jumps off the platform, there was a potential bias toward more variability in males with an upper CId boundary of $d < 0.21$ indicating that if present the difference was marginal at most. The remaining five measures showed a potential bias toward more variability in females with lower CId boundaries between $d > -0.22$ and $d > -0.33$ indicating that if present at all the differences were marginal (Whishaw's error, time near wall) or small (swim path, path efficiency, target proximity) at most. Accordingly, CV values for measures of training performance in female and male mice showed largely overlapping distributions across experimental groups (Figure 3c,g,k). The distributions of male–female CV differences were close to symmetrical (Figure 3d,h,l) indicating that differences were poorly predictable at the level of single experimental groups. Despite swim speed being overall significantly more variable in female mice (Figure 1b middle), there were still 43% of groups in which males actually had a higher CV (Figure 3f).

3.3 | Water-maze place navigation, probe trial

Quantification of searching in the former goal area during the probe trial serves to assess spatial memory and the precision of navigation. In the present study, we evaluated time spent in the former goal quadrant, time spent in a narrow circular zone around the former goal, annulus crossing index, target proximity, as well as polar error. According to all five measures males were significantly better (Figure 1a middle). But despite the statistical significance of the effect, evaluation of CId established equivalence of the sexes in all measures (Figure 1a middle). In line with this, individual probe trial scores had very similar distributions (Figure 4a,e). The overall effect was also reproduced poorly across experiments with male

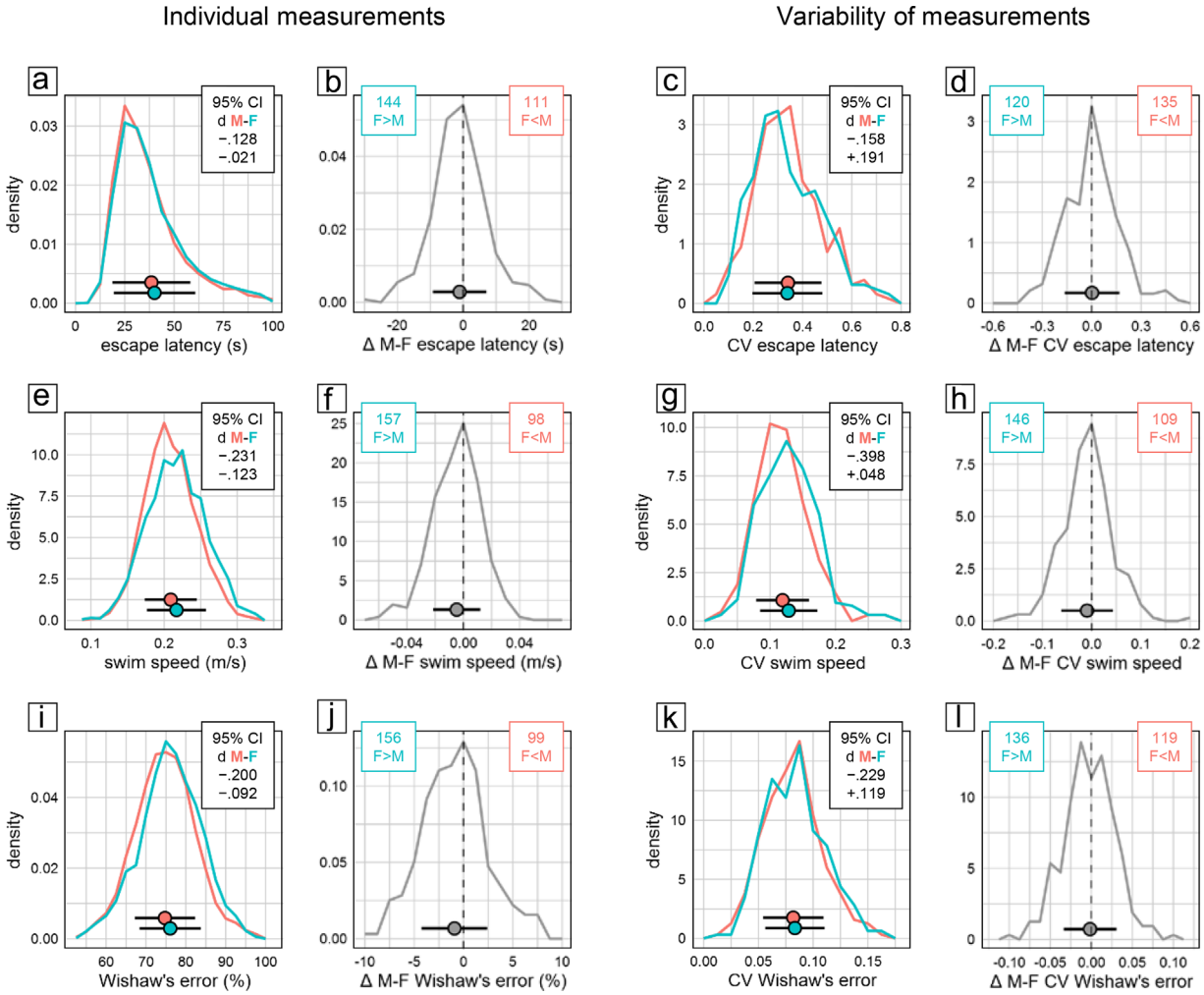

FIGURE 3 Selected measures characterizing training performance in the water-maze place navigation task. Frequency polygons illustrate sex differences of individual values (a and b, e and f, i and j) and their coefficient of variation CV as a measure of variability (c and d, g and h, k and l) for escape latency (a–d), swim speed (e–h), and Whishaw's error (i–l). Data are analyzed and presented as in Figure 2. There were 2,658 male (red) and 2,653 female (cyan) mice tested in 255 male (red) and 255 female (cyan) groups

mice showing superior proximity scores only in 62% of experimental groups (Figure 4b) and smaller polar error only in 66% (Figure 4f).

Analysis of CV detected a significant sex difference only for 1 of 4 measures (Figure 1a middle). Polar error was significantly more variable in males, with the sex effect accounting for little more than 1% of total variance ($\eta^2 = 0.01110$) and an upper CId boundary of $d < 0.39$ confirming that the effect was small at most. For quadrant and zone time there was a potential bias toward more variability in females with a lower CId boundary of $d > -0.22$ indicating that if present at all the difference was marginal at most. Crossing index and target proximity showed a potential bias toward more variability in males, with an upper boundary of $d < 0.33$ indicating that if present the difference was small at most. The distribution of CV values across female and male subgroups was very similar (Figure 4c,g). Proximity scores were more variable in males in only 53% of experimental groups (Figure 4d), polar error varied more only in 57% (Figure 4h).

3.4 | Open field

The open field test is primarily used to assess locomotor activity in a novel environment. In the present study, we assessed three measures of activity: distance moved, time spent walking, and walking speed. In all three, females has significantly higher scores than males (Figure 1a bottom). But despite the statistical significance of the effect, evaluation of CId established equivalence of the sexes for time walking. For distance moved and walking speed, a lower CId boundary of $d > -0.24$ indicated that the effect was marginal at most. This was supported by the highly congruent distribution of individual male and female scores (Figure 5a,e). Despite overall statistical significance, the sex difference was also unreliable across experiments, with females moving a longer distance than males only in 58% of experimental groups (Figure 5b) and moving faster than males only in 62% of experimental groups (Figure 5f).

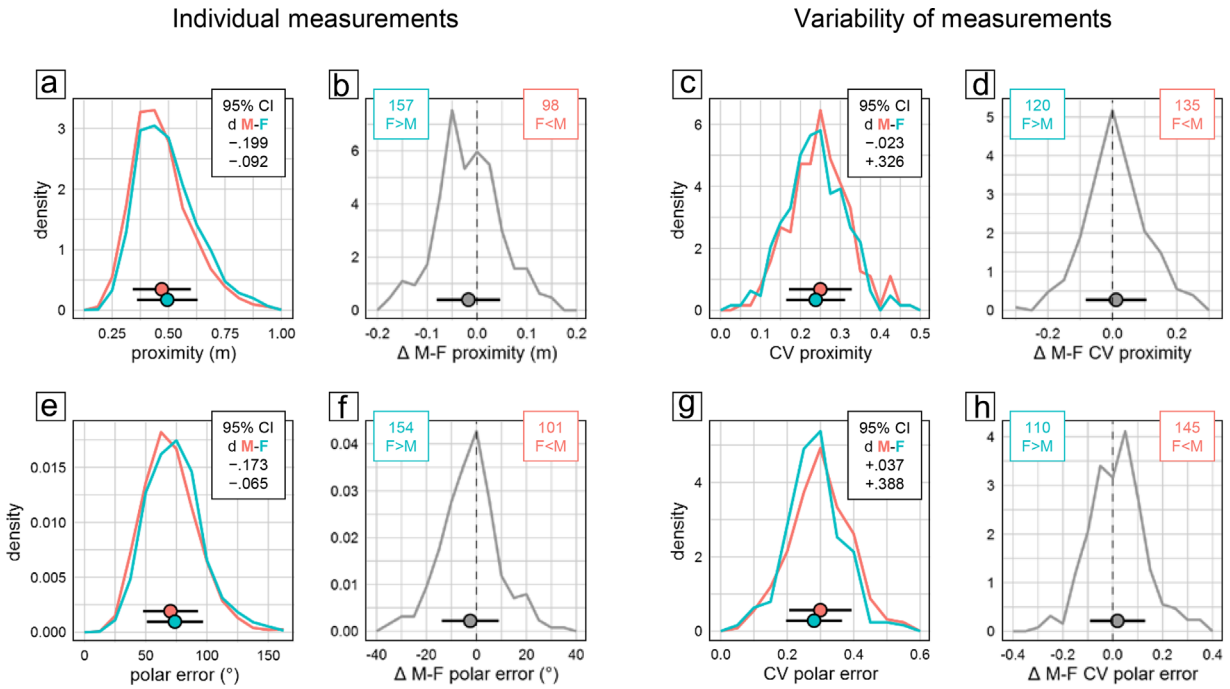

FIGURE 4 Selected measures characterizing probe trial retention in the water-maze place navigation task. Frequency polygons depict sex differences of individual values (a and b, e and f) and their coefficient of variation CV as a measure of variability (c and d, g and h) for average distance to trained target (a–d) and polar error relative to trained target (e–h). Data are analyzed and presented as in Figure 2. There were 2,658 male (red) and 2,653 female (cyan) mice tested in 255 male (red) and 255 female (cyan) groups

In addition to measuring activity, the open field is also used to assess anxiety related responses. We evaluated average distance to center and time in center as measures of center avoidance. In addition, we examined the percentage of arena surface explored and the speed difference between movements toward and away from the center. Of these four measures only speed difference showed a significant sex difference (Figure 1a bottom) and evaluation of CId established equivalence for all four measures. In agreement with this, the distribution of individual measures of female and male mice was nearly congruent (Figure 5i) and the distribution of male–female differences within experimental groups highly symmetrical (Figure 5j). Fecal boli deposited in the open field are sometimes counted as an indirect measure of emotionality. In our data, counts were available in 3,451 mice from 170 experimental groups. Male mice deposited significantly more boli than females, with a CId indicating an effect of small size.

Analysis of CV as measure of variability detected significant sex effects in only two of eight open field measures (Figure 1b bottom). Variability was larger in female mice for distance moved ($\eta^2 = 0.01547$) and number of fecal boli ($\eta^2 = 0.01500$), with the sex effect accounting for less than 2% of total variance and a lower CId boundary of $d > -0.46$ indicating that the effect was small at most. Five measures (time walking, walking speed, distance to center, center time, area explored) showed a potential bias toward more variability in females with lower CId boundaries of $d > -38$ indicating that if present at all differences were small at most. Out-inbound speed difference showed a potential bias toward higher variability in males, with an upper CId boundary of $d < 0.27$ indicating that if present the difference was small at most. In line with this, the distribution of male and female CV values in experimental groups showed large overlap (Figure 5c,g,k), and the number of female subgroups effectively showing larger CV than corresponding male subgroups did not exceed 58% (Figure 5d,h,l).

4 | DISCUSSION

4.1 | Inclusion of female mice has a negligible or very small impact on data variability

To compare variability in female and male mice, we calculated the coefficient of variation (CV = standard deviation/mean) for female and male subgroups and determined the 95% confidence interval for the size (CId) of the sex difference in order to test for equivalence and to judge the maximal impact in case of non-equivalence. Of the 16 measures evaluated in the water-maze place navigation task (Figure 1b middle), only two showed a significant sex effect. Swim speed was more variable in females, probe trial polar error was more variable in males. Female–male equivalence of CV could be established for four measures. For the rest CId evaluation showed that differences, if present at all, were marginal to small at most, with females potentially more variable in eight, males in two measures. Of the eight measures evaluated in the open field (Figure 1b bottom), only two showed a significant sex effect, with slightly larger CV in females. For the rest CId

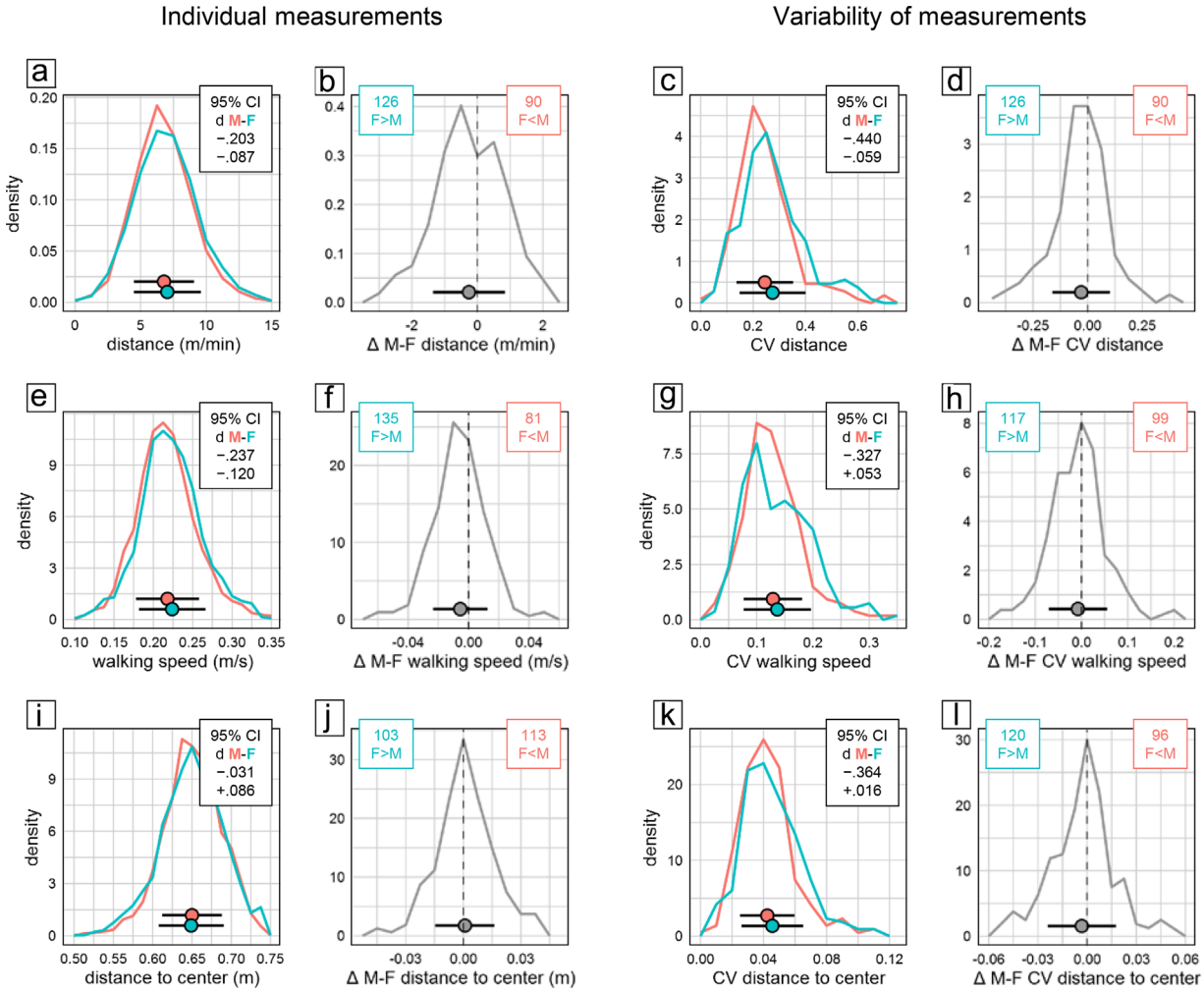

FIGURE 5 Selected measures characterizing behavior in the open field. Frequency polygons illustrate sex differences of individual values (a and b, e and f, i and j) and their coefficient of variation CV as a measure of variability (c and d, g and h, k and l) for distance moved per min observation time (a–d), average speed during bouts of walking (e–h), and distance to center (i–l). Data are analyzed and presented as in Figure 2. There were 2,303 male (red) and 2,251 female (cyan) mice tested in 216 male (red) and 216 female (cyan) groups

evaluation showed that differences, if present at all, were marginal to small at most, with females tending to be more variable in five and males in one measure.

Even though slightly larger CV in female mice was found or could not be ruled out for part of the measures, the results of this study clearly demonstrate that inclusion of female mice in water-maze place navigation experiments and in open-field tests does not lead to a relevant increase of data variability. (i) Despite the large data set only few and small significant sex effects on CV were detected. (ii) For the rest of the measures anything more than a marginal or small sex effect on CV could be ruled out. (iii) Some measurements showed larger CV in males, indicating that there are male-specific sources of variability, such as differences in social hierarchy. (iv) Even the largest sex effects on CV accounted for <2% of total CV variance, meaning that >98% of group to group variation in CV is due to factors unrelated to the sex, such as genetic background, test environment, or history of the animals. (v) Due to the small size of sex effects, it was impossible in our data to predict reliably whether CV would be larger in males or females at the level of the single experiment.

Even though slightly larger CV in female mice was found or could not be ruled out for part of the measures, the results of this study clearly demonstrate that inclusion of female

mice in water-maze place navigation experiments and in open-field tests does not lead to a relevant increase of data variability.

Our study makes a strong case for the inclusion—together with male mice—of females without testing for estrous cycle in studies of spatial memory, anxiety-related behavior, and locomotor activity with mouse models of nervous system disease. This will render studies more representative (Clayton, 2016; Miller et al., 2017; Shansky & Woolley, 2016) and provide opportunities to study sex-specific aspects of disease, provided that they are biologically determined (Eliot & Richardson, 2016). Using only male mice will not make results more reliable. Rather, efforts should be maximized to control more relevant sources of variability, such as genetic background (Crusio, Goldowitz, Holmes, & Wolfer, 2009; Lipp & Wolfer, 2003; Magara et al., 1999; Mohajeri et al., 2004; Wolfer, Muller, Stagliar-Bozizevic, & Lipp, 1997) and laboratory environment (Crabbe, Wahlsten, & Dudek, 1999).

4.2 | Equivalent learning performance and exploration in female and male mice

In addition to evaluating CV as a measure of data variability, we also directly compared mean scores of female and male mice. The analysis of water-maze place navigation performance revealed statistically significant sex effects in the majority of measures suggesting superior training and probe trial performance in male compared to female mice. However, the CId of nearly all these sex effects did not extend beyond the equivalence margin of 0.2, meaning that the differences were negligible and scores in fact equivalent in both sexes. Evaluation of open field activity measures revealed statistically significant sex differences as well, suggesting higher activity in female mice. But again, they were negligibly small. Analysis of anxiety-related measures in the open field provided no evidence for a significant sex difference, with female and male scores being statistically equivalent. Even for those measures that showed a statistically significant sex effect, the distributions of male–female differences across experimental groups were near symmetrical, making it impossible to predict in our data whether in a single experimental group females or males would have higher scores. The only measurements to show sex differences that were both statistically significant and of relevant size were body weight and the number of fecal boli deposited in the open field, both higher in male mice.

The only measurements to show sex differences that were both statistically significant and of relevant size were body weight and the number of fecal boli deposited in the open field, both higher in male mice.

While the observation of a significant advantage of male mice in the place-navigation task confirmed our expectations, the small, in fact negligible, size of this effect may appear surprising. We speculate that the observed sex effects are the vestiges of a former biologically relevant sexual dimorphism in spatial navigation during dispersal. In wild mice, dispersal is mainly driven by aggressive behavior, which is more intense among males than females. However, whether or not males have to venture new territories depends on complex interactions of factors such as age and hierarchical status of the individual, population density, food availability and abiotic environmental factors (Latham & Mason, 2004). Under laboratory conditions, these factors are largely non-existent and sex specific variations in spatial navigation are present but do not reach significant effect size. The lack of sex differences in anxiety- and exploration related open field measures found here is in line with a recent study in wild Mus musculus performing the same task (Bimova, Mikula, Macholan, Janotova, & Hiadlovska, 2016), which found neither sex differences nor an effect of estrus phase on female exploratory behavior. Alternatively, the negligible effect size in sex-specific spatial navigation in laboratory mice could be the consequence of generally reduced aggressive behavior in males by the lack of selective pressure on this trait over many generations of breeding in a laboratory environment.

The observation of a significant advantage of male mice in the water-maze place navigation task is of interest also in view of the ample evidence in human psychology documenting superior performance of men in spatial tasks, while women perform better in other cognitive domains. This effect is particularly well known and reliable in tasks involving mental rotation of objects (Voyer, 2011; Voyer, Voyer, & Bryden, 1995), but a male advantage has also been observed repeatedly in studies involving navigation of a virtual water-maze (Astur et al.,1998; Astur, Purton, Zaniewski, Cimadevilla, & Markus, 2016; Astur, Tropp, Sava, Constable, & Markus, 2004; Korthauer, Nowak, Frahmand, & Driscoll, 2017; Newhouse, Newhouse, & Astur, 2007). Performance differences between women and men in spatial tasks may be strongly confounded with socialization and gender-biased expectations (Eliot & Richardson, 2016; Estes & Felker, 2012), but the

observation of congruent sex differences in water-maze navigation between mice and human subjects is nevertheless remarkable.

5 | CONCLUSION

Our retrospective analysis of sex differences in a large number of mice tested in the water-maze place navigation task and in the open field confirms and extends other large scale studies in demonstrating that including female mice in experiments does not cause a relevant increase of data variability. Further, performance of female and male mice in these tests is equivalent. Our results make a strong case for including mice of both sexes whenever these tests are used in preclinical research. This will be necessary for adequate modeling of the human population and to capture sex-dependent aspects of disease mechanisms.

ACKNOWLEDGMENTS

We would like to express our gratitude—without naming them individually—to the many staff members and students who over the years have contributed to collecting behavioral data. We thank Gene Fisch for helpful advice with statistical analysis and Lutz Slomianka for critical reading of the manuscript. The studies have been supported by numerous grants of the Swiss National Science Foundation as well as by the 6th and 7th Framework Programmes of the European Union. IA and DPW are members of the Neuroscience Center Zurich (ZNZ). DPW is also a member of the Zurich Center for Integrative Human Physiology (ZIHP).

ORCID

David P. Wolfer http://orcid.org/0000-0002-5957-1401

REFERENCES

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*(1–2), 185–190.

Astur, R. S., Purton, A. J., Zaniewski, M. J., Cimadevilla, J., & Markus, E. J. (2016). Human sex differences in solving a virtual navigation problem. *Behavioural Brain Research*, *308*, 236–243.

Astur, R. S., Tropp, J., Sava, S., Constable, R. T., & Markus, E. J. (2004). Sex differences and correlations in a virtual Morris water task, a virtual radial arm maze, and mental rotation. *Behavioural Brain Research*, *151*(1–2), 103–115.

Bimova, B. V., Mikula, O., Macholan, M., Janotova, K., & Hiadlovska, Z. (2016). Female house mice do not differ in their exploratory behaviour from males. *Ethology*, *122*(4), 298–307.

Chambers, L. K., Singleton, G. R., & Krebs, C. J. (2000). Movements and social organization of wild house mice (Mus domesticus) in the wheatlands of northwestern Victoria, Australia. *Journal of Mammalogy*, *81*(1), 59–69.

Christensen, D. L., Baio, J., Van Naarden Braun, K., Bilder, D., Charles, J., Constantino, J. N. ... Centers for Disease C, Prevention. (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveillance Summaries*, *65*(3), 1–23.

Clayton, J. A. (2016). Studying both sexes: A guiding principle for biomedicine. *FASEB Journal*, *30*(2), 519–524.

Clayton, J. A., & Collins, F. S. (2014). Policy: NIH to balance sex in cell and animal studies. *Nature*, *509*(7500), 282–283.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science*, *284*(5420), 1670–1672.

Crusio, W. E., Goldowitz, D., Holmes, A., & Wolfer, D. P. (2009). Standards for the publication of mouse mutant studies. *Genes Brain and Behavior*, *8*, 1–4.

D'Hooge, R., & De Deyn, P. P. (2001). Applications of the Morris water maze in the study of learning and memory. *Brain Research Reviews*, *36*(1), 60–90.

Eliot, L., & Richardson, S. S. (2016). Sex in context: Limitations of animal studies for addressing human sex/gender neurobehavioral health disparities. *The Journal of Neuroscience*, *36*(47), 11823–11830.

Estes, Z., & Felker, S. (2012). Confidence mediates the sex difference in mental rotation performance. *Archives of Sexual Behavior*, *41*(3), 557–570.

Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., ... Cox, D. R. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, *448*(7157), 1050–1053.

Frick, K. M., Burlingame, L. A., Arters, J. A., & Berger-Sweeney, J. (2000). Reference memory, anxiety and estrous cyclicity in C57BL/6NIA mice are affected by age and sex. *Neuroscience*, *95*(1), 293–307.

Gallagher, M., Burwell, R., & Burchinal, M. (1993). Severity of spatial learning impairment in aging: Development of a learning index for performance in the Morris water maze. *Behavioral Neuroscience*, *107*(4), 618–626.

Ge, J. F., Qi, C. C., Qiao, J. P., Wang, C. W., & Zhou, N. J. (2013). Sex differences in ICR mice in the Morris water maze task. *Physiological Research*, *62*(1), 107–117.

Golden, L. C., & Voskuhl, R. (2017). The importance of studying sex differences in disease: The example of multiple sclerosis. *Journal of Neuroscience Research*, *95*(1–2), 633–643.

Haaxma, C. A., Bloem, B. R., Borm, G. F., Oyen, W. J., Leenders, K. L., Eshuis, S., ... Horstink, M. W. (2007). Gender differences in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, *78*(8), 819–824.

Hendershott, T. R., Cronin, M. E., Langella, S., McGuinness, P. S., & Basu, A. C. (2016). Effects of environmental enrichment on anxiety-like behavior, sociability, sensory gating, and spatial learning in male and female C57BL/6J mice. *Behavioural Brain Research*, *314*, 215–225.

Jonasson, Z. (2005). Meta-analysis of sex differences in rodent models of learning and memory: A review of behavioral and biological data. *Neuroscience and Biobehavioral Reviews*, *28*(8), 811–825.

Kokras, N., & Dalla, C. (2017). Preclinical sex differences in depression and antidepressant response: Implications for clinical research. *Journal of Neuroscience Research*, *95*(1–2), 731–736.

Korthauer, L. E., Nowak, N. T., Frahmand, M., & Driscoll, I. (2017). Cognitive correlates of spatial navigation: Associations between executive functioning and the virtual Morris Water Task. *Behavioural Brain Research*, *317*, 470–478.

Latham, N., & Mason, G. (2004). From house mouse to mouse house: The behavioural biology of free-living Mus musculus and its implications in the laboratory. *Applied Animal Behaviour Science*, *86*, 261–289.

Leger, M., & Neill, J. C. (2016). A systematic review comparing sex differences in cognitive function in schizophrenia and in rodent models for schizophrenia, implications for improved therapeutic strategies. *Neuroscience and Biobehavioral Reviews*, *68*, 979–1000.

Lipp, H. P., & Wolfer, D. P. (2003). Genetic background problems in the analysis of cognitive and neuronal changes in genetically modified mice. *Clinical Neuroscience Research*, *3*, 223–231.

Madani, R., Kozlov, S., Akhmedov, A., Cinelli, P., Kinter, J., Lipp, H. P., … Wolfer, D. P. (2003). Impaired explorative behavior and neophobia in genetically modified mice lacking or overexpressing the extracellular serine protease inhibitor neuroserpin. *Molecular and Cellular Neurosciences*, *23*(3), 473–494.

Magara, F., Muller, U., Li, Z. W., Lipp, H. P., Weissman, C., Stagliar-Bozizevic, M., & Wolfer, D. P. (1999). Genetic background changes the pattern of forebrain commissure defects in transgenic mice underexpressing the beta-amyloid-precursor protein. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 4656–4661.

Miller, L. R., Marks, C., Becker, J. B., Hurn, P. D., Chen, W. J., Woodruff, T., … Clayton, J. A. (2017). Considering sex as a biological variable in preclinical research. *FASEB Journal*, *31*(1), 29–34.

Mogil, J. S. (2016). Perspective: Equality need not be painful. *Nature*, *535*(7611), S7.

Mogil, J. S., & Chanda, M. L. (2005). The case for the inclusion of female subjects in basic science studies of pain. *Pain*, *117*(1–2), 1–5.

Mohajeri, M. H., Madani, R., Saini, K., Lipp, H. P., Nitsch, R. M., & Wolfer, D. P. (2004). The impact of genetic background on neurodegeneration and behavior in seizured mice. *Genes Brain and Behavior*, *3*(4), 228–239.

Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation*, *12*, 239–260.

Morris, R. G., Garrud, P., Rawlins, J. N. P., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*(5868), 681–683.

Newhouse, P., Newhouse, C., & Astur, R. S. (2007). Sex differences in visual-spatial learning using a virtual water maze in pre-pubertal children. *Behavioural Brain Research*, *183*(1), 1–7.

Pocock, M. J. O., Hauffe, H. C., & Searle, J. B. (2005). Dispersal in house mice. *Biological Journal of the Linnean Society of London*, *84*(3), 565–583.

Pocock, M. J. O., Searle, J. B., & White, C. L. (2004). Adaptations of animals to commensal habitats: Population dynamics of house mice Mus musculus domesticus on farms. *Journal of Animal Ecology*, *73*, 878–888.

Poirier, R., Jacquot, S., Vaillend, C., Soutthiphong, A. A., Libbey, M., Davis, S., … Wolfer, D. P. (2007). Deletion of the Coffin–Lowry syndrome gene rsk2 in mice is associated with impaired spatial learning and reduced control of exploratory behavior. *Behavior Genetics*, *37*(1), 31–50.

Prendergast, B. J., Onishi, K. G., & Zucker, I. (2014). Female mice liberated for inclusion in neuroscience and biomedical research. *Neuroscience and Biobehavioral Reviews*, *40*, 1–5.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Compuring.

Schoenfeld, R., Schiffelholz, T., Beyer, C., Leplow, B., & Foreman, N. (2017). Variants of the Morris water maze task to comparatively assess human and rodent place navigation. *Neurobiology of Learning and Memory*, *139*, 117–127.

Shansky, R. M., & Woolley, C. S. (2016). Considering sex as a biological variable will be valuable for neuroscience research. *The Journal of Neuroscience*, *36*(47), 11817–11822.

Tschanz, J. T., Corcoran, C. D., Schwartz, S., Treiber, K., Green, R. C., Norton, M. C., … Lyketsos, C. G. (2011). Progression of cognitive, functional, and neuropsychiatric symptom domains in a population cohort with Alzheimer dementia: The Cache County Dementia Progression study. *American Journal of Geriatric Psychiatry*, *19*(6), 532–542.

Voikar, V., Koks, S., Vasar, E., & Rauvala, H. (2001). Strain and gender differences in the behavior of mouse lines commonly used in transgenic studies. *Physiology & Behavior*, *72*(1–2), 271–281.

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin and Review*, *18*(2), 267–277.

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*(2), 250–270.

Whishaw, I. Q., (1985). Evidence for two types of place navigation in the rat. In G. Buzsaki (Ed.), *Electrical activity of the archicortex* (pp. 233–253). Budapest: Hungarian Academy of Sciences.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, USA: Springer.

Wolfer, D. P., Colacicco, G., & Welzl, H., (2013). Learning and memory: Water navigation tasks. In W. E. Crusio, F. Sluyter, R. T. Gerlai, & S. Pietropaolo (Eds.), *Behavioral genetics of the mouse*. Cambridge: Cambridge University Press.

Wolfer, D. P., Madani, R., Valenti, P., & Lipp, H. P. (2001). Extended analysis of path data from mutant mice using the public domain software Wintrack. *Physiology & Behavior*, *73*(5), 745–753.

Wolfer, D. P., Muller, U., Stagliar-Bozizevic, M., & Lipp, H. P. (1997). Assessing the effects of the 129/Sv genetic background on swimming navigation learning in transgenic mutants: A study using mice with a modified beta-amyloid precursor protein gene. *Brain Research*, *771*, 1–13.

Yang, H., Bell, T. A., Churchill, G. A., & Pardo-Manuel de Villena, F. (2007). On the subspecific origin of the laboratory mouse. *Nature Genetics*, *39*(9), 1100–1107.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., … Pardo-Manuel de Villena, F. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics*, *43*(7), 648–655.

Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, *465*(7299), 690.

How to cite this article: Fritz A-K, Amrein I, Wolfer DP. Similar reliability and equivalent performance of female and male mice in the open field and water-maze place navigation task. *Am J Med Genet Part C Semin Med Genet*. 2017;175C:380–391. https://doi.org/10.1002/ajmg.c.31565