Not All Mice Are Created Equal: Interval Timing Accuracy and Scalar Timing in 129, Swiss-Webster, and C57BL/6 Mice

Catalin V. Buhusi1, Abby E. Meyer2, Sorinel A. Oprisan2 and Mona Buhusi1

1 Neuroscience Program, Department of Psychology, Utah State University, Logan, UT 84322, USA
2 Department of Physics and Astronomy, College of Charleston, Charleston, SC 29424, USA

Corresponding author; e-mail: Catalin.Buhusi@USU.edu

ORCID iDs: C.V. Buhusi: 0000-0002-9895-9687; Oprisan: 0000-0001-6585-3985

Received 1 March 2022; accepted 6 June 2022; published online 15 July 2022; published in issue 10 April 2023

Abstract

Many species, including humans, show both accurate timing — appropriate time estimation in the seconds to minutes range — and scalar timing — time estimation error varies linearly with estimated duration. Behavioral paradigms aimed at investigating interval timing are expected to evaluate these dissociable characteristics of timing. However, when evaluating interval timing in models of neuropsychiatric disease, researchers are confronted with a lack of adequate studies about the parent (background) strains, since accuracy and scalar timing have only been demonstrated for the C57BL/6 strain of mice (Buhusi, Aziz, Winslow, Carter, Swearingen, & Buhusi (2009) Behav. Neurosci., 123, 1102–1113). We used a peak-interval (PI) procedure with three intervals — a protocol in which other species, including humans, demonstrate accurate, scalar timing — to evaluate timing accuracy and scalar timing in three strains of mice frequently used in genetic and behavioral studies: 129, Swiss-Webster (SW), and C57BL/6. C57BL/6 mice showed accurate, scalar timing, while 129 and SW mice showed departures from accuracy and/or scalar timing. Results suggest that the genetic background/strain of the mouse is a critical variable for studies investigating interval timing in genetically engineered mice. Our study validates the PI procedure with multiple intervals as a proper technique, and the C57BL/6 strain as the most suitable genetic background to date for behavioral investigations of interval timing in genetically engineered mice modeling human disorders. In contrast, studies using mice in 129, SW, or mixed-background strains should be interpreted with caution, and thorough investigations of accuracy and scalar timing should be conducted before a less studied strain of mouse is considered for use in timing studies.

Timing & Time Perception 11 (2023) 242–262
Published with license by Koninklijke Brill nv | DOI: 10.1163/22134468-Bja10052
© Catalin V. Buhusi et al., 2022 | ISSN: 2213-445X (print) 2213-4468 (online)
This is an open access article distributed under the terms of the CC BY 4.0 license.
Keywords
Interval timing, mouse, C57BL/6, 129, Swiss-Webster, scalar property, scalar timing, peak-interval procedure

1. Introduction

Many species, including humans, show both accurate timing — appropriate time estimation in the seconds to minutes range — and scalar timing — time estimation error varies linearly with estimated duration, reflecting Weber’s Law applied to interval timing (Gibbon, 1977). A number of neuropsychiatric disorders, such as Parkinson’s disease (Malapani et al., 1998), Huntington’s disease (Paulsen et al., 2004), schizophrenia (Snowden and Buhusi, 2019), attention deficit hyperactivity disorder (Noreika et al., 2013), and Alzheimer’s disease (Nichelli et al., 1993) affect patients’ accuracy and/or their scalar timing. To elucidate the genetic and molecular mechanisms underlying disruptions of interval timing in human disorders, studies have been carried out in animal models, including wild-type mice (Mus musculus) and genetically modified mice (e.g., Abner et al., 2001; Buhusi et al., 2013; Gallistel et al., 2004; Gür et al., 2019a; Meck, 2001). However, to date, only one strain of mice — the C57BL/6 strain — has been thoroughly investigated regarding its accurate and scalar timing (Buhusi et al., 2009). In contrast, many mouse models of disease are generated using parental (background) strains such as 129 or Swiss-Webster (SW) whose timing performance is relatively unknown (e.g., C3H mice used in Balci et al., 2009a), raising questions on the results, interpretation, and translational value of interval timing data from studies using these models.

Cutting-edge behavioral techniques evaluating interval timing in animal subjects should allow for the independent evaluation of both accurate timing of multiple intervals and of scalar timing in the same paradigm. In contrast, some studies in mice investigate accuracy but not scalar timing (Gür et al., 2019a), others evaluate scalar timing but not accuracy (Gallistel et al., 2004), while others fail to report the strain of mouse used as background (Meck, 2001). A behavioral task fully demonstrated to allow the simultaneous evaluation of both accurate and scalar timing is the peak-interval (PI) procedure (Catania, 1970) with multiple, for example three, intervals. The peak-interval procedure with three intervals (3PI) has been successfully used in both humans (Rakitin et al., 1998) and mice (Buhusi et al., 2009) to investigate in the same study whether subjects use accurate, scalar timing. In the 3PI procedure accurate timing is reflected by maximal response at each of the three criterion durations, while scalar timing is reflected by the superposition of the three individual response functions when normalized relative to the estimated durations.

Due to the important homology between the mouse and human genomes, mice offer a powerful tool for elucidating both major genes underlying human disorders
— through generation of genetically engineered mice (transgenics, knockout, knockins) — and the genetic architecture of complex traits, including behavioral traits. Traditionally, mouse models have been generated by introducing foreign genes into the nucleus of a fertilized egg or generating chimaeric mice using homologous recombination in embryonic stem (ES) cells (Capecchi, 2022; Doetschman et al., 1987; Thomas and Capecchi, 1987). For example, because 129-derived ES cells have been used the most to create genetically engineered mice, many studies investigating the effect of genetic manipulations on various behaviors (reviewed in Crawley et al., 1997), including interval timing (reviewed in Karson and Balci, 2021) were performed in inbred 129, or 129 mixed-backgrounds.

However, one main drawback with the generation of engineered mice using ES cells is the presence of a mixed background (Carstea et al., 2009), leading to the necessity for an extended period of backcrossing to the target genetic background, such as the inbred C57BL/6 strain. C57BL/6 mice are one of the best characterized inbred strains genetically (Mouse Genome Sequencing Consortium et al., 2002; Seong et al., 2004), neuroanatomically (Paxinos and Franklin, 2013), and behaviorally (Crawley et al., 1997; Gerlai, 1996; Owen et al., 1997). Recent utilization of the CRISPR/Cas9 system has permitted researchers to develop methods for direct genome editing in fertilized eggs of animals within the target genetic background, saving considerable time for backcrossing (Mashiko et al., 2013, 2014; Yang et al., 2013).

Another approach to investigating gene–behavior links is to perform gene-wide association studies (GWAS) in both humans and rodents. This approach aims to identify expressed quantitative trait loci (eQTL) that co-map with behavioral QTLs in an effort to reveal the most likely candidate genes (Parker et al., 2016; Rice and O’Brien, 1980; Zou et al., 2022). Although not developed for genetic research, SW mice have been maintained as an outbred population for more than 100 generations, which degraded the linkage disequilibrium between nearby alleles making this strain more valuable (when compared to other commercially available strains) for GWAS.

Here we evaluated timing accuracy and scalar timing in the 3PI procedure with three strains of mice frequently used in genetic and behavioral studies: 129, C57BL/6, and SW mice. C57BL/6 mice showed accurate, scalar timing, while 129 and SW mice showed departures from accuracy and/or scalar timing, thus calling into question the interpretation of interval timing studies using mice in the 129 and SW strains, in mixed strains, or in strains in which timing was not fully characterized (Agostino et al., 2013; Balci et al., 2009b, 2010; Balzani et al., 2016; Brunner et al., 2015; Cordes and Gallistel, 2008; Drew et al., 2007; Gür et al., 2019a, b; Meck, 2001; Morè et al., 2017; Tucci et al., 2014).
2. Materials and Methods

2.1. Animals

Six-month old naïve male mice of the following strains: inbred C57BL/6NCrl (BL6, \( n = 7 \)), inbred 129S2/SvPasCrl (129, \( n = 7 \)), and outbred Crl:CFW(SW) Swiss-Webster (SW, \( n = 7 \)) (Charles River Labs, Raleigh, NC, USA) were housed in groups of two or three in a temperature-controlled room, under a 12/12-h light/dark cycle. Mice were tested during the light period of the cycle. Water was given ad lib in the home cages. Mice were maintained at 85% of their ad lib weight by restricting their access to food (Rodent Diet 5001, PMI Nutrition International, Brentwood, MO, USA). All manipulations were performed in compliance with guidelines of the National Institutes of Health regarding the care and use of animals for experimental procedures.

2.2. Apparatus

The apparatus consisted of 12 standard mouse operant boxes housed in sound-attenuating cubicles (MED Associates, St. Albans, VT, USA). Boxes were equipped with three nosepokes situated two on the front wall and one on the back wall. For each subject one criterion duration (10-s, 20-s, or 40-s) was randomly assigned to each nosepoke, counterbalanced by location. The lights inside the nosepokes were used to signal the to-be-timed durations. According to the schedule, 20-mg precision food pellets (Research Diets, Inc., New Brunswick, NJ, USA) were delivered in a food cup situated on the front wall, in between the nosepokes, by a pellet dispenser. A 66-dB sound produced by a fan was present throughout all procedures. The intensity of the fan was measured with a sound level meter (Realistic Sound Level Meter 33–2050, RadioShack, Fort Worth, TX, USA) from the center of the (silent) box. The experimental procedures were controlled through a MED Associates interface (MED-Associates, 1999).

2.3. Behavioral Procedure

Procedures closely followed Buhusi et al. (2009). Briefly, after mice were accustomed to nose-poking for food in the boxes, they received nine fixed-interval (FI) sessions in which FI trials with the criterion durations associated to the nosepokes were randomly intermixed. At the onset of each FI trial, the light inside the appropriate nosepoke was turned on. A nosepoke at or after the criterion time associated with that nosepoke was reinforced with one food pellet irrespective of responding on the other nosepokes, and the light was turned off for the duration of the intertrial interval (ITI), which lasted for three times the criterion duration plus a variable 2–16-s duration (uniform distribution). In the next 34 sessions, FI trials were randomly interspersed with non-reinforced PI trials for each criterion duration. During each PI trial, the light in the corresponding nosepoke was turned on for three times the corresponding criterion duration, and then turned...
off irrespective of any responses for the duration of the ITI. Subjects received on average 24 FI trials and 12 trials of each type of PI trial per 3.5-h session. Data from the last four sessions (about 50 PI trials for each criterion duration) were submitted to data analyses.

2.4. Data Analyses

The start time, end time, and duration of each nosepoke response were recorded in real time. A time-in-nosepoke (TIN) was computed as the average proportion of each time unit the mouse nosepoked. TIN functions were normalized in the interval 0–300% criterion in 10% criterion time bins: 1-s bins for the 10-s criterion, 2-s bins for the 20-s criterion, and 4-s bins for the 40-s criterion. Three analyses were conducted. First, the individual normalized TIN distributions were fit (least-squares minimization) using the Marquardt–Levenberg iterative algorithm (Marquardt, 1963) to the following Gaussian + linear equation:

\[ TIN(t) = \text{peak} \times \exp \left\{ -0.5 \times \left[ \frac{(t - \mu)}{\sigma} \right]^2 \right\} + \text{tail} \times (t - \mu) \]  

(see Buhusi and Meck, 2000; Buhusi et al., 2009). The algorithm provided individual parameters \( \mu \), peak, \( \sigma \), and tail, estimating respectively the accuracy of timing (peak time of TIN function), maximum TIN (amplitude of TIN function), precision of interval timing (width of TIN function), and skewness of the TIN function (right tail of TIN function). The goodness of fit (the variance in responding explained by equation (1)) was estimated. One SW mouse was very variable, and its estimated goodness of fit was under 50%; this mouse was eliminated from analyses. For the remaining mice the estimated goodness of fit was 95.43 ± 0.01% for BL6 mice \( (n = 7) \), 90.10 ± 0.03% for 129 mice \( (n = 7) \) and 86.67 ± 0.05% for SW mice \( (n = 6) \). A second, independent measure of response peak time was derived using a modified interquartile range (IQR) procedure. The area under the individual normalized TIN curve was divided into four quartiles; the estimated median peak time \( (M) \) was taken to be the average of the first and third quartiles. Third, the degree of superposition of the individual TIN functions was estimated using the \( \eta^2 \) superposition index, which measures the degree with which one variable accounts for variability in another variable. In summary, timing accuracy was estimated using two methods: a shape-dependent fitting method which provided a lower-bound estimate of the peak time, and a shape-independent IQR method which provided an upper-bound estimate for it. Finally, the degree to which the subjects use scalar timing was estimated by two methods: a shape-dependent fitting method which provided the precision parameter \( \sigma \) of the normalized TIN function, and a shape-independent method which provided the degree \( \eta^2 \) with which the response functions for the three criteria superimpose: \( \eta^2_{10–20}, \eta^2_{10–40}, \)
and $\eta^2_{20-40}$. These measures were submitted to statistical analyses to evaluate differences between strains.

2.5. Statistical Analyses

The estimated peak times, $\mu$ and $M$, were submitted to repeated-measures ANOVAs with between-subjects factor strain (BL6, 129, SW) and within-subjects factor duration (10, 20, 40 s) to evaluate differences between strains in accuracy, followed by $t$ tests to evaluate differences from the to-be-timed durations. The precision $\sigma$ of the normalized TIN function was submitted to a repeated-measures ANOVA with between-subjects factor strain (BL6, 129, SW) and within-subjects factor duration (10, 20, 40 s) to evaluate scalar timing; lack of reliable differences in $\sigma$ over criterion durations would indicate scalar timing. The degree of superposition of response functions $\eta^2$ was submitted to a repeated-measures ANOVA with between-subjects factor strain (BL6, 129, SW) and within-subjects factor pair of response functions ($\eta^2_{10-20}$, $\eta^2_{10-40}$, and $\eta^2_{20-40}$) to provide an independent evaluation of scalar timing; superposition (scalar timing) was taken to be ‘very good’ for $\eta^2 > 0.80$, ‘good’ for $\eta^2$ between 0.6 and 0.8, ‘fair’ for $\eta^2$ between 0.4 and 0.6, while values under 0.4 are taken to reflect ‘poor’ and ‘very poor’ superposition. Statistical analyses were followed by post-hoc least significant difference (LSD) tests and planned comparisons. All statistical tests were evaluated at a significance level of 0.05.

3. Results

3.1. Timing Functions in Absolute Time Units

Figure 1 shows average TIN functions for the three criterion durations in 129, BL6, and SW. Figure 1 indicates that all mice showed temporal regulation of behavior at the three intervals, maximal poking about the criterion durations, and timing functions which seem to increase in width with the to-be-timed criterion, possibly indicating scalar timing. Yet, Fig. 1 also indicates differences in responding between strains: while BL6 mice seem accurate at all three criteria, 129 and SW mice seem to peak later than the 10-s criterion. SW mice also seem earlier at the 40-s criterion, and to have wider functions (reduced precision) than the other

![Figure 1. Average time-in-nosepoke (TIN) functions for the three criterion durations in 129, BL6, and SW mice.](image-url)
strains. Finally, all strains seem to respond more at the shorter durations than at the longer duration, although this effect is not considered to be related to timing (for a similar observation see Buhusi et al., 2009). These suggestions were confirmed by further analyses.

3.2. Peak TIN Amplitude

The estimated peak amplitudes of the TIN function for each strain and for each criterion interval were submitted to repeated-measures ANOVA with between-subjects variable strain (129, BL6, and SW) and within-subjects variable criterion duration (10 s, 20 s, and 40 s). Analyses indicated a main effect of criterion duration ($F_{2,34} = 22.64, p < 0.0001$, partial $\eta^2 (\eta^2_p) = 0.57$), but no other main effects or interactions ($Fs < 2.54, p > 0.1086$), suggesting that mice responded more at the short durations than at the longer duration, similarly to Buhusi et al. (2009). As traditionally the measure on the y-axis is not considered relevant for timing, but may indicate other effects possibly related to motivation and so on, and no significant differences were found between strains, the TIN functions in Fig. 1 were normalized in amplitude (relative to the peak TIN value) and analyzed regarding their temporal distribution (x-axis).

3.3. Timing Functions in Relative Time Units

The TIN functions in Fig. 1 were normalized in amplitude (relative to the peak TIN value), and in 10% criterion (relative time) units, as shown in Fig. 2. Figure 2 indicates that while the normalized response functions superimpose well in BL6 mice, the degree of superposition is reduced in both 129 and SW mice, which seem to peak later at the 10-s criterion, and earlier at the 40-s criterion. To evaluate these suggestions, the individual normalized TIN functions were fitted with a Gaussian function whose parameters were submitted to statistical analyses.

3.4. Timing Accuracy by the Curve-Fitting Method

The estimated normalized peak time, $\mu$, estimated by a shape-dependent curve-fitting method, for each strain and for each criterion interval, is shown as filled symbols in Fig. 3. A repeated-measures ANOVA with between-subjects variable strain (129, BL6, and SW) and within-subjects variable criterion (10 s, 20 s, and 40 s)
40 s) indicated a main effect of criterion ($F_{2,34} = 8.07, p = 0.0014, \eta^2_{p} = 0.32$), but no other main effects or interactions ($Fs < 1.83, p > 0.1468$), suggesting that mice were not equally accurate at the three intervals, as indicated in Fig. 3. LSD post-hoc analyses indicated that BL6 mice were not only accurate (not significantly different from 100% criterion for all three criteria, all $ts < 1.87, p > 0.1101$), but also similarly accurate at the three criteria (not significantly different between criteria, all $ps > 0.2879$). In contrast, 129 mice responded earlier at the 40-s criterion than at the 10-s criterion ($p = 0.0346$); similarly, SW mice were earlier at the 20-s criterion than at the 10-s criterion ($p = 0.0365$), and also much earlier at the 40-s criterion than at the 10-s criterion ($p = 0.0003$). The latter can be observed in Fig. 2 as well: the TIN (red broken) curve for the 10-s criterion peaks later than (at the right of) the 40-s (thick green) curve in both 129 and SW mice. Taken together, results indicate that of the three strains, the C57BL/6 mice are consistently accurate at the three intervals, while both 129 and SW mice show deviations from accuracy. Both show later timing of the shorter interval than of the longer interval.

### 3.5. Timing Accuracy by the IQR Method

The estimated median peak time, $M$, estimated by a shape-independent IQR method, for each strain and for each criterion interval, is shown as open symbols in Fig. 3. One notes that while the shape-dependent curve-fitting method provided a
lower-bound estimate for accuracy (Fig. 3 filled symbols), the shape-independent IQR method provided an upper-bound estimate for accuracy. Indeed, the IQR method provided higher estimates for accuracy than the curve-fitting method because it is heavily influenced by the skewness of the response function (the tail of the function), while the curve-fitting method accounts for skewness as a separate parameter from accuracy and precision, and thus it is not influenced by the tail. As shown in Fig. 3, the IQR method estimated that all strains differ from 100% accuracy. However, although the IQR method provided higher estimates for the peak time, the pattern of accuracy produced by the two methods is consistent: both methods indicated that BL6 mice were more accurate than both 129 and SW mice.

A repeated-measures ANOVA of the median peak time, $M$, estimated by the shape-independent IQR method, with between-subjects variable strain (129, BL6, and SW) and within-subjects variable criterion duration (10 s, 20 s, and 40 s) indicated a main effect of strain ($F_{2,17} = 8.06, p = 0.0035, \eta^2_p = 0.49$), a main effect of criterion ($F_{2,34} = 4.77, p = 0.0149, \eta^2_p = 0.32$), and a marginally significant interaction ($F_{2,34} = 2.33, p = 0.0761, \eta^2_p = 0.21$), suggesting that mice differed in their accuracy at the three intervals, as indicated in Fig. 3. While all mice differed significantly from 100% accuracy (all $t$s > 3.79, $ps < 0.01$), planned comparisons indicated that over all durations BL6 mice were significantly more accurate than both 129 mice ($p = 0.0096$) and SW mice ($p = 0.0018$). LSD post-hoc analyses indicated that both BL6 mice and 129 mice were similarly accurate at the three criteria (not significantly different between criteria, all $ps > 0.1787$), while SW mice peaked later at the 10-s criterion than both at the 20-s criterion ($p = 0.0192$) or 40-s criterion ($p = 0.0009$). Taken together, results using both estimation methods indicate that of the three strains, the C57BL/6 mice are consistently accurate (or more accurate) at the three intervals, while both 129 and SW mice show deviations from accuracy: Both show later timing at the shorter interval than at the longer ones.

### 3.6. Scalar Timing by the Curve-Fitting Method

Scalar timing was evaluated by estimating precision $\sigma$ of the normalized TIN function, which was submitted to repeated-measures ANOVA with between-subjects variable strain and within-subjects variable criterion duration. Scalar timing would be indicated by lack of differences in $\sigma$ between the criteria. Analyses indicated a main effect of strain ($F_{2,17} = 9.89, p < 0.0014, \eta^2_p = 0.54$), and a marginally significant main effect of criterion duration ($F_{2,34} = 3.14, p = 0.0561, \eta^2_p = 0.16$), but no significant interaction, suggesting that three strains have different timing precisions, as indicated in Fig. 4. LSD post-hoc analyses provided details of the differences in precision and scalar timing between strains: BL6 were consistently precise (scalar), they had small $\sigma$ values, indicating small widths of the TIN function (increased precision) and no differences in precision between criteria.
(all \( p > 0.1866 \)), indicating scalar timing. SW mice were consistently less precise but scalar: they had large \( \sigma \) values, indicating large widths of the TIN function (decreased precision) but no differences in precision between criteria (all \( p > 0.22 \)), indicating scalar timing. In contrast, precision varied in 129 mice between the three criteria, with significantly higher precision (small \( \sigma \)) at the 20-s criterion relative to both the 10-s and 40-s criteria (all \( p < 0.0155 \)), indicating departures from scalar timing. Taken together, the shape-dependent curve-fitting method indicated that interval timing is precise and scalar in BL6 mice, it is imprecise but scalar in SW mice, and it departs from scalar in 129 mice in which relative precision varies between the three intervals.

### 3.7. Scalar Timing by Curve Superposition

To further evaluate scalar timing we estimated the degree of superposition between the three timing functions using the \( \eta^2 \) superposition index, as shown in Fig. 5. Figure 5 indicates very good superposition in BL6 mice, good superposition in 129 mice, and poor superposition in SW mice. A repeated-measures ANOVA with between-subjects variable strain and within-subjects variable the pair of function (\( \eta^2_{10-20}, \eta^2_{10-40}, \) and \( \eta^2_{20-40} \)) indicated a main effect of strain (\( F_{2,17} = 47.77, p < 0.0001, \eta^2_p = 0.85 \)), and a main effect of pair (\( F_{2,34} = 3.29, p = 0.0493, \eta^2_p = 0.16 \)), but no interaction, suggesting that three strains have different degrees of superposition between timing functions. Planned comparisons indicated that over all pairs of functions, BL6 mice showed significantly higher superposition over 129 mice (\( p = 0.0103 \)), and that both BL6 and 129 mice showed significantly higher superposition than SW mice (\( p < 0.0001 \) for both comparisons). Taken together, the two analyses of scalar timing show consistently that BL6 mice have

**Figure 4.** Average precision \( \sigma \) of the normalized TIN function (± SEM) using the shape-dependent curve-fitting method in 129 (left, blue, circles), BL6 (center, red, squares), and SW mice (right, green, triangles) for the three criterion durations: 10 s, 20 s, and 40s. Interval timing is precise and scalar in BL6 mice, it is imprecise but scalar in SW mice, and it is non-scalar in 129 mice. *, \( p < 0.05 \).
better scalar timing than both 129 and SW mice: BL6 mice have equal precision at, and better superposition between, timing function than both 129 and SW mice, which show smaller or different precision at, or superposition between, timing functions.

4. Discussion

We evaluated timing accuracy and scalar timing in 129S2 (129), C57BL/6N (BL6), and CFW/SW mice trained to time three intervals (3PI procedure): 10s, 20s, and 40s. Accurate and scalar timing were each evaluated by two independent methods, a shape-dependent, Gaussian curve-fitting method, and a shape-independent method (either the IQR method or the superposition index $\eta^2$). Analyses consistently indicated significant differences in interval timing behavior between the three strains of mice. Interval timing in BL6 mice was both accurate (Figs 2 and 3), precise (small width of timing functions, Fig. 3), and scalar: relative precision did not differ significantly between criteria (Fig. 4), and timing curves superimposed very well (Fig. 5). In contrast, interval timing in SW mice was inaccurate, with differences in accuracy between criteria (e.g., later responses for short duration and earlier responses for longer durations, Figs 2 and 3), imprecise (large width of timing functions, Fig. 4), but scalar (relative precision did not differ between criteria, Fig. 3); their timing functions superimposed very poorly (Fig. 5), due in part by their poor accuracy (Fig. 2). Finally, 129 mice showed departures both in accuracy and scalar timing: they showed later responses for short durations and earlier responses for longer durations (Figs 2 and 3), their relative timing precision
was not similar between the three criteria (Fig. 4); although their functions superimposed better than SW, and their superposition index was lower than in BL6 (Fig. 5). These results suggest that timing accuracy and scalar timing reflect separable aspects of interval timing. In our study the three strains showed varying degrees of accuracy, and varying degrees of scalar timing (or lack thereof), at different criterion intervals.

4.1. Implications for Behavioral Procedures

These results suggest that appropriate techniques are required to estimate both accuracy and scalar timing in the same setting in various strains of mice at multiple intervals. In contrast, current studies in mice investigate accuracy but not scalar timing (Gür et al., 2019a), others evaluate scalar timing but not accuracy (Gallistel et al., 2004), others used procedures that engage additional behaviors besides interval timing (Balci et al., 2008), while other simply fail to report the strain of mouse or background (Meck, 2001). For example, while in the present study we found scalar timing in BL6 mice (see also Buhusi et al., 2009), studies using ‘switching’ – a behavior that relies on both spatial and temporal control – fail to do so (Balci et al., 2008). Differences may stem at the neurobiological level, since interval timing depends on the fronto-striatal circuits (Buhusi and Meck, 2005), while response switching depends on the limbic system (Goto and Grace, 2005; Sleezer et al., 2017). In summary, evaluation of timing abilities in mice should be pursued in a suitable behavioral task, fully demonstrated to allow the simultaneous evaluation of both accurate and scalar timing, such as the peak-interval procedure (Catania, 1970).

4.2. Strain Differences in Interval Timing Behavior

Various strains of mice have vastly different performance characteristics in many behavioral paradigms including the Morris water task (Klapdor and van der Staay, 1996), fear conditioning (Smith et al., 2007), radial arm maze (Ammassari-Teule et al., 1993), conditional spatial alternation (Crawley et al., 1997), social behavior (Champagne et al., 2007), prepulse inhibition (Paylor and Crawley, 1997), and various tests of analgesia (Smith, 2019) or depression (Andolina et al., 2014). Therefore, it is not surprising that strain differences are also found in interval timing behavior. For example, while outbred SW mice respond earlier at longer durations than at shorter ones (our findings, Fig. 3; see also Gallistel et al., 2004), inbred BL6 mice are equally accurate in the range of 10–40 s. Moreover, while BL6 were highly scalar in our study, both 129 and SW mice showed varying degrees of deficits in scalar timing.

Interstrain differences may be the result of genetic, neuroanatomical and neurophysiological differences in the three strains. For example, genetic studies (SNP characterization through GBS; Parker et al., 2016) indicate that outbred SW mice are most genetically similar to FVB/NJ, and less similar to 129 substrains or BL6
mice. In turn, while the genome of 129 and SW mice is not fully characterized, the BL6 genome has been almost entirely sequenced and analyzed (Mouse Genome Sequencing Consortium et al., 2002) and represents the laboratory mouse reference genome. The latter is continuously updated by the Genome Reference Consortium and other groups (Sarsani et al., 2019). Similarly, the neuroanatomy of BL6 mice is also best characterized through numerous studies and systematic analyses (see Allen Institute for Brain Science, 2004, a genome-wide, cellular-resolution atlas of gene expression throughout the adult mouse brain; Paxinos and Franklin, 2013). In contrast, while the neuroanatomy of SW mice is poorly known, 129 substrains exhibit known differences in brain structure compared to BL6: smaller cortical and striatal, but increased cerebellar and hippocampal volumes (Chen et al., 2006), and many 129 mice exhibit major abnormalities, such as reduction or absence of corpus callosum and hippocampal commissure (Bohlen et al., 2012; Wahlsten, 1982; Wahlsten et al., 2003).

Neuropharmacological studies also reveal differences between these mouse strains: BL6 and SW strains are differentially sensitive to D1-like and D2-like agonists (Ralph and Caine, 2005). Moreover, relative to other strains, BL6 mice are more sensitive to the effects of direct and indirect dopaminergic (DA) agonists (Puglisi-Allegra and Cabib, 1997), possibly due to increased DA transmission in the mesostriatal projection, a DA circuit critical for interval timing (Buhusi and Meck, 2005; Meck, 1996). These features explain BL6 mice’s greater novelty-induced locomotor activity and impulsivity (Crawley et al., 1997; Gerlai, 1996; Owen et al., 1997). In contrast, 129 mice show reduced basal striatal DA uptake compared to BL6 and SW mice, but show enhanced DA responses to cocaine (He and Shippenberg, 2000). Differences in the DA system between these three strains are most likely responsible for the different results in timing experiments. The present results validate the BL6 mice as the most suitable genetic background to date to model the accurate and scalar timing in normal human populations (Rakitin et al., 1998).

It is interesting to note that differences may occasionally be observed even between variants of the same inbred strain, e.g., when maintained for many generations by different suppliers. For example, genetic differences exist between mice supplied by Charles River Labs (C57BL/6N, also used in this study) and those supplied by Jackson Labs (C57BL/6J) (Mekada et al., 2009). Behavioral differences in acoustic startle, open-field locomotion, rotarod, Morris water maze, light/dark test, and even morphologic and function differences in the eye were noted between these two C57BL/6 substrains (Simon et al., 2013). To further investigate timing endophenotypes in C57BL/6N and C57BL/6J, Maggi et al. (2014) measured their behavior in an automated apparatus over several days, and noted no major behavioral differences in time-based switching (Balci et al., 2008). Although the latter study does not speak to whether the two substrains are similar with respect to their accuracy and scalar timing, it suggests that the two substrains may have very
similar timing behaviors. To compare and contrast their accurate and scalar timing, the two substrains should be tested side-by-side in a PI timing task, as done in this study. For example, using a PI procedure with gaps and distracters we noted sex differences in attention to time in C57BL/6J mice (Buhusi et al., 2017); replicating this results in C57BL/6N mice would provide further confidence these two substrains are similar not only regarding timing, but also regarding attentional processing of temporal information.

4.3. Implications for Genetic Models of Human Interval Timing

To model human disorders, and to allow for the dissociation of the effect of the genetic manipulation at hand from the contribution of the genetic background of the model (strain), careful controls are required both at the level of phenotype, genotype, and behavioral technique. The timing phenotype of an animal is the result of a triple interaction between a new Gene mutation × the Genetic Background of the animal (other genes) × Environment. To evaluate the effect of a new gene mutation, one needs to control both for background and environmental effects. First, to control for environmental effects, it is recommended one uses as controls littermate animals, bred and raised in identical environmental conditions. Second, to control for genetic background effects, and to model human interval timing, the background strain should demonstrate both accurate and scalar timing similar with the target population, humans (Rakitin et al., 1998). To date, this has been demonstrated only in C57BL/6 mice (this study; see also Buhusi et al., 2009). In contrast, current studies of interval timing in mice use a variety of mixed strains (Agostino et al., 2013; Balci et al., 2009b, 2010; Balzani et al., 2016; Brunner et al., 2015; Cordes and Gallistel, 2008; Drew et al., 2007; Gür et al., 2019a, b; Meck, 2001; Moré et al., 2017; Tucci et al., 2014), or simply fail to report the strain used (Meck, 2001), thus failing to provide adequate controls for possible gene–background interactions. Results from these interval timing studies may be equally due to the genetic manipulation or to the poor timing abilities of the specific background strain used. Optimally, to control for the genetic background, the mouse model should be selected from an inbred (e.g., BL6, see this study), rather than from an outbred, strain (e.g., SW, see this study). This report extends our previous study validating the C57BL/6 strain as the only suitable genetic background to date for exploring the genetic and molecular mechanisms of interval timing with relevance to human conditions and disorders (Buhusi et al., 2009). This finding is consistent with C57BL/6 strain having good learning abilities in associative studies, and being the background of choice for behavioral research with genetically modified mice (Crawley et al., 1997; Gerlai, 1996; Owen et al., 1997).

Here, it is important to note that simply using ‘wild-type’ control animals does not automatically control for either genetic background or environmental effects, as oftentimes an argument is made. Wild-type animals control for the effect of the
new gene mutation only in the context of a given genetic background. Instead, if the wild-type animals are not littermates with the animals carrying the mutated gene, they do not control for environmental effects. Similarly, if the genetic background alters accuracy and/or scalar timing in the first place (such as 129 and SW in this study), wild-type animals of said genetic background are not adequate controls. In summary, when assessing gene mutations on the timing phenotype, one should use wild-type littermate controls, and also check whether the genetic background itself alters timing behavior. The latter requires that the genetic background of the animal be validated before use in timing studies, as done in this study (see also Buhusi et al., 2009, 2017). Alternatively, animals carrying the mutated gene should be backcrossed for at least 10 generations into an already validated genetic background (such as C57BL/6), which would allow studying the timing phenotype of the mutated gene in a “control background” (Buhusi et al., 2013).

Finally, when modeling human disorders in which interval timing is disrupted, it is critical to evaluate both the accurate and scalar timing in the same study, as these two features of timing may not covary in wild-type and mutant mice, and may reflect differential effects of the mutation on different brain mechanisms. For example, in the present study BL6 mice showed accurate and scalar timing, SW mice showed inaccurate but scalar timing, while 129 mice showed both inaccurate and non-scalar timing, suggesting the two properties do not covary. The current protocol — the 3PI procedure — allows for independent evaluation of the individual timing accuracy and scalar timing. The PI procedure was also successfully used both in human participants and in rodent models to evaluate the effect of various pharmacological manipulations on the speed of an internal clock (Maricq and Church, 1983), on memory for criterion durations (Buhusi and Meck, 2002; Meck, 1996), and on attention to time (Buhusi and Meck, 2002, 2007; Buhusi et al., 2017). Therefore, the PI procedure seems an ideal technique for investigations of interval timing in genetically engineered mice. The use of similar tasks — like the PI procedure — in which the accurate, scalar timing was demonstrated by both human participants (Rakitin et al., 1998) and mice (this study), is important for establishing an effective translational value of a finding in genetically engineered mice.

Acknowledgements

This work is dedicated to the memory of Warren H. Meck, teacher, mentor, colleague and friend. His work continues to be an inspiration for future generations of interval timing enthusiasts. This manuscript was supported by NIH grants NS123824 to CVB, and NS109723, NS090283 to MB.
References


