Evaluation of two automated metrics for analyzing partner preference tests

Todd H. Ahern[a,*], **Meera E. Modi**[a,1], **James P. Burkett**[a,1], and **Larry J. Young**[a,b,2]

[a] Center for Behavioral Neuroscience and Yerkes National Primate Research Center, Emory University, Atlanta, GA 30322, USA

[b] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322, USA

Abstract

The partner preference test (PPT) is commonly used to examine sexual and social preferences in rodents. The test offers experimental subjects a choice between two stimulus animals, and time spent with each is used to calculate a preference score. In monogamous prairie voles (*Microtus ochrogaster*), the PPT has been paramount to the study of pair bonding. Although powerful, use of the PPT in voles has depended primarily on human manual scoring. Manual scoring is time-consuming and is susceptible to bias and fatigue, limiting the use of the PPT in high-throughput studies. Here we compared manual scoring (real-time and 16x) and two automated scoring metrics: "social proximity" and "immobile social proximity" We hypothesized that "immobile social contact" would provide data most comparable to manually scored "huddling", and thus be the most sensitive measure of partner preference in prairie voles. Each automated metric produced data that highly correlated with manual scoring (R > 0.90); however, "immobile social contact" more closely reflected manually scored huddling (R = 0.99; P < 0.001). "Social proximity" and "immobile social contact" were then used to detect group partner preferences in four data sets that varied by cohabitation length and sex. "Immobile social contact" revealed a significant partner preference in each data set; "social proximity" detected partner preferences in only three of the four. Our results demonstrate the utility of automated systems in high-throughput PPTs, and further confirm that automated systems capable of scoring "immobile social contact" yield results indistinguishable from manual scoring.

Keywords

Behavior; Automation; Monogamy; Vole

---

*Corresponding author: Todd H. Ahern, 954 Gatewood Rd., Yerkes National Primate Research Center, Emory University, Atlanta GA 30322, Phone: 404 727-8269, Fax: 404 727-8070, Email: todd.h.ahern@gmail.com.
[1]Tel.: 1 404 727 8269; fax: +1 404 727 8070.
[2]Tel.: 1 404 727 8272; fax: +1 404 727 8070.
Todd H. Ahern, 954 Gatewood Rd., Yerkes National Primate Research Center, Emory University, Atlanta GA 30322, Phone: 404 727-8269, Fax: 404 727-8070, Email: todd.h.ahern@gmail.com
Meera E. Modi, 954 Gatewood Rd., Yerkes National Primate Research Center, Emory University, Atlanta GA 30322, Phone: 404 727-8269, Fax: 404 727-8070, Email: meera.modi@gmail.com
James P. Burkett, 954 Gatewood Rd., Yerkes National Primate Research Center, Emory University, Atlanta GA 30322, Phone: 404 727-8269, Fax: 404 727-8070, Email: james.p.burkett@gmail.com
Larry J. Young, 954 Gatewood Rd., Yerkes National Primate Research Center, Emory University, Atlanta GA 30322, Phone: 404 727-8272, Fax: 404 727-8070, Email: lyoun03@emory.edu

Introduction

The partner preference test (PPT) is a commonly used behavioral assay for the evaluation of social and sexual behavior. In the PPT, test subjects are offered a choice of two stimulus animals of differing social or sexual valence. Subjects are then observed and the number of approaches toward and time spent in close proximity to, or in contact with, each stimulus animal is used to calculate a preference score (Slob et al., 1981; Slob et al., 1987; Baum et al., 1990; Williams et al., 1992; Crawley, 2004; Nadler et al., 2004). Use of the PPT has varied widely and has helped identify an array of biological factors regulating sexual preference (Johnson and Tiefer, 1972; Slob et al., 1987; Baum et al., 1990), social approach and investigation (Crawley, 2004; Nadler et al., 2004), and social attachment (Williams et al., 1992; Winslow et al., 1993; Carter et al., 1995; Insel and Hulihan, 1995; Young and Wang, 2004). Each laboratory has modified the PPT to suit its own needs, but one commonality has been a dependence on human scoring for one or more of the behavioral metrics. Unfortunately, manual scoring is extraordinarily time and labor intensive and has prevented, in some cases, the use of the PPT in high-throughput studies. Manual scoring also suffers from questions of bias, fatigue, and inter-rater reliability (Noldus et al., 2001).

One superlative example illustrating the issue of time investment is the use of the PPT to study pair bonding in monogamous prairie voles (*Microtus ochrogaster*). Prairie voles are socially monogamous rodents that form selective, enduring attachments between mates after a period of cohabitation and/or mating. This emerging model organism has proven useful for investigating the neurobiological and genetic mechanisms underlying a variety of social phenomena, including bond formation (Carter et al., 1995; Young and Wang, 2004), bond maintenance (Aragona et al., 2006; Bosch et al., 2009), and the consequences of social loss and isolation (Grippo et al., 2007; Bosch et al., 2009). Central to nearly all of these investigations is the ability to measure a selective social bond using the PPT. Adoption of the PPT to study voles, however, has required several modifications. Before testing, prairie voles undergo a period of cohabitation with a conspecific. The "partner" is then loosely tethered in one stimulus chamber while a novel "stranger" is loosely tethered in another. The test animal is placed in the central, socially neutral chamber and allowed to roam for 3 hours. The amount of time spent "huddling" (which is variably denoted as contact, side-by-side contact, or affiliative contact) with each stimulus animal is then scored, since it is the most sensitive indicator of a selective partner preference in this species and its congeners (Williams et al., 1992; Winslow et al., 1993; Carter et al., 1995; Insel and Hulihan, 1995; Lim et al., 2004; Young and Wang, 2004). While the PPT has been highly effective in identifying socio-active manipulations in both male and female prairie voles, the near exclusive reliance on manual scoring has stalled the deployment of this animal model in large-scale drug discovery studies. This has remained true even with the use of time-lapse recording techniques.

In an attempt to increase the efficiency, scalability, and reliability of the PPT, we objectively compared two manual scoring methods (real-time and time-lapse) and two different automated metrics: "social proximity" and "immobile social contact." In this study, we scored "social proximity" using a relatively old release of Noldus EthoVision, version 3.0 (release date November, 2002), which we had been using for elevated plus maze and open-field assays (Ahern & Young, unpublished data). We used this older version of EthoVision 3.0 as just one, representative example of an automated system that can easily and automatically processes the location of a single animal over time within defined zones. It is important to note that several other video- and beam-break-based systems now have similar capabilities and could likely produce a similar "social proximity" metric under partner preference testing conditions in which the stimulus animals are tethered.

To process "immobile social contact," we used Clever Sys Inc.'s newest version of SocialScan, 2.0 (January, 2007), with minor custom-modifications. We used this software as one, representative example of an automated system that can distinguish the contours of three unmarked animals in a single arena, process social interactions based on distinct bodily features and movements, and could be adapted to provide a measure of "immobile social contact." We again note that, while SocialScan 2.0 was the first system to meet our specific needs, other software packages, including Noldus EthoVision XT 6.0, claim similar capabilities.

Ultimately, the goal of this study was to test the hypothesis that the automated metric of "immobile social contact" would better reflect "huddling" time as determined by a human rater than the metric of "social proximity," where "huddling" is predominately immobile, affiliative social contact. After comparing the accuracy of each method in relation to manual scoring, we then tested the ability of the "social proximity" and "immobile social contact" metrics to detect significant group partner preferences in four data sets from subjects that differed by sex and cohabitation time. We again hypothesized that the ability to detect "immobile social contact" would produce data more consistent with manual scoring methods than the "social proximity" metric.

Our results demonstrate that an automated approach to scoring the PPT yields data that are highly correlated with manual scoring methods. Systems that provide a "social proximity" metric provide good approximations in most situations and may be a cost-effective automated method, if used cautiously. Alternatively, software capable of detecting "immobile social contact," appears to be a more viable option for high-throughput studies of social bond formation, since it more accurately approximates manual scoring and is able to detect partner preferences across several key variables.

Differences in scoring between the two automated methods were apparent, but not easily explained. The most obvious differences occurred when scoring male prairie voles that had cohabitated for a week. Males exhibit increasing levels of selective aggression toward strangers after extended periods of cohabitation with a partner (Aragona et al., 2006). Aggression therefore offered one plausible explanation for the discrepancy between the two metrics: aggression would likely increase stranger "social proximity" time without increasing stranger "immobile social contact" time. Aggression and other qualitative explanations are discussed.

Most importantly, both metrics reached a respectable level of accuracy and substantially reduced time and labor investments. Automation thus appears to be a highly efficient and generally accurate method of obtaining data from the PPT.

Methods

The partner preference testing setup

Each PPT apparatus (Fig. 1a) was constructed in-house out of 0.6 cm thick acrylic sheets and heat-resistant cement (Weld-On #16 Acrylic Solvent Cement, IPS Corporation, Gardena, CA). The apparatus consists of one long box 75 cm × 20 cm × 30 cm (length × width × height), divided into three chambers by placing two opaque slats, measuring 25 cm tall × 6 cm wide × 0.6 cm thick, opposite each other at 25 cm and 50 cm along the length of the walls. At each end of the box, a 20 cm deep × 1 cm wide slot was carved out to allow tethering anchors to slide in and out.

The anchoring apparatus was also made in-house. A small loop was made out of metal wire (1/16th inch thick) and then cinched (3/32 in. Aluminum Ferrule & Stop Set, The Lehigh Group, Macungie, PA). The free end was passed through three 1/4 in. nuts and two 3/8 by 1(1/2) in. zinc plated washers in alternation. The wire was then cinched again. To this loop-nut-washer

combination, three interlinked Brass Snap Swivel fishing leaders (Size 10; Eagle Claw, Wright & McGill Co., Denver, CO) were attached. Through the most distal leader, an 8 in. cable tie (Commercial Electric, 826 843, USA) was looped to make a neck leash. The tethering anchor allows the researcher to tether the animal outside the apparatus then slide the anchor into the 20 cm deep slot, securing the tethered animal in the appropriate area.

Before each test day, the cages were steam washed and new bedding (Bed-O-Cob, Maumee, OH) was added to cover the floor. 50 mL conical tubes with water and sipper tubes were attached to aluminum frames and hung at each end over the anchoring slots; there was no water-bottle in the center chamber.

Recording equipment

Our setup includes 12 PPT boxes. Video recording was performed with four digital video cameras (WV-CP284, Panasonic), each monitoring three adjacent testing boxes (Fig. 1b and 2a, c). Cameras were situated vertically 91 cm from the floor of the testing chambers, and all four cameras fed into a QUAD video box (4CQ, EverPlex); the QUAD collapsed the video streams into a single output. Exiting the QUAD, the video was simultaneously sent to a digital video disk (DVD) recorder (LQ-DRM200, Panasonic) and to a WinTV-PVR-350 video-card (model 990) installed in a PC (Dell Precision T3400) running Windows XP (Fig. 1b). The DVD recorder outputs to a television for real-time viewing of the cameras' perspectives (not shown). By collapsing all four video streams into one output, there is some loss of resolution. The use of a vertical viewing perspective and our move to real-time rather than compressed video-recording, however, easily compensated for any resolution loss, while simultaneously allowing high-throughput testing.

Experimental animals

All subjects were either female or male prairie voles from our colony at Emory University (Atlanta, GA), which were weaned at 21 days post-birth, housed in same sex pairs or trios, and maintained on a 14:10 light:dark schedule (lights ON at 06:00 h) with *ad libitum* access to food and water. Animals were tested as adults (70–100 days old). All procedures were approved by Emory University's Institutional Animal Care and Use Committee and were in accordance with national guidelines.

Partner preference testing

Twelve sexually naïve, gonadally intact experimental females and twelve sexually naïve, gonadally intact experimental males were paired at 07:00 h with an equal number of sexually-naïve, gonadally-intact stimulus animals of the opposite sex and allowed to cohabitate for 24 hours. At 07:00 h the following morning, all pairs were moved to a behavioral testing room and allowed to acclimate for >30 min. Each cohabitation "partner" was then tethered and anchored to one end of the PPT box, while a "stranger" was tethered and anchored at the other end. Water-bottles were placed at each end to cover the anchor slots. Finally, the experimental test animal was placed in the center, socially neutral chamber and allowed to roam freely for 3 hours.

On each day of testing, a total of 12 experimental test animals were tested in the morning and 12 more were tested in the afternoon. The partners from the morning session served as strangers in the afternoon and vice versa. This procedure has been used previously and there are no measureable test order effects, which would confound the results (Lim et al., 2007). It also substantially reduces animal use. Our 24 experimental animals (12 males + 12 females) underwent PPT after 24 hours of cohabitation and again after 6 more days (a total 1 week of cohabitation). All 48 PPTs were video-recorded to DVD and compressed to MPEG files using the WinTV-PVR-350 video-card. All behavioral scoring occurred *post hoc*.

Automated video-tracking of "social proximity"

DVD recorded test sessions were analyzed for "social proximity" *post hoc* using an early version of Noldus EthoVision, version 3.0 (Noldus, The Netherlands). A DVD player (D-R410, Toshiba) fed video directly into a PC (Dell Dimension 8200), which contained a Picolo Frame grabber and video receiver. EthoVision 3.0 processed all forty-eight 3-h PPTs as they passed through the Picolo system.

For each DVD, at least 10 seconds of video containing clean PPT boxes with water-bottles had been recorded, providing a background image. This image served as a template for the virtual arenas (Fig. 2a). A virtual arena defines where the program can track an animal; thus each PPT apparatus was defined by a separate arena. In each case, the virtual arena outlined the floor of the test box, but ended at a position in the stimulus chambers beyond which the tethered animals could not reach (Fig. 2a). This prevented the presence of two animals in the arena simultaneously, which causes errors in the program.

Virtual zones define regions of interest within the virtual arena. Zone lines were therefore drawn across the base of each pair of slats and again a few centimeters from the ends of the arena definition. Six zones were defined: (1) LeftSocial, (2) LeftTransition, (3) CenterChamber, (4) RightTransition, (5) RightSocial, and (6) NonSocial (using the cumulative zone feature) (Fig. 2a). LeftSocial and RightSocial were further defined as hidden zones. Hidden zones are typically used in assays where an animal can escape into an opaque house and thus out of sight of the camera. If an animal crosses the hidden zone border and disappears, EthoVision 3.0 allocates the interim time to the hidden zone. We modified the use of this feature by simply truncating the virtual arena. When the test animal crosses the marker for the left or right social zone and disappears out of the virtual arena, the program allocates time to the appropriate hidden social zone. Time in the hidden social zones thus served as a measure of time in "social proximity." With this software, "social proximity" is the best proxy of "huddling" without dyeing or marking the animals in some way—a potential confound for studies of social behavior. Regardless of zone, the test animal's location was measured by its center-of-gravity (Fig. 2c).

After defining the arena, zones, and distance calibration, several other features were altered to optimize detection of the free-roaming, experimental animal. Trial protocol was set to 3:00 hours, with a sample rate of 15 samples/s. Hidden zones were set to 40 mm for entering and 0 mm for exiting. Detection was accomplished by background subtraction, but only when the object was darker than the background and larger than 40 pixels. To eliminate background noise, contrast and brightness were adjusted to +236 and −14, respectively (see Fig. 2c). Each detectable object underwent a two-pixel erosion and then a three-pixel dilation. When the test animal entered the hidden zone and thus was not found, the program used the last measured position. Detection thresholds were set to −255 to −25 and the reference image was updated before proceeding to the introduction of test animals.

With our setup, EthoVision 3.0 is able to process 12 PPT boxes simultaneously at a processor load of between 15 and 45, well below a load of 100, which can yield slow and inaccurate image processing. EthoVision 3.0 processed the entire 3-h test for each animal. Test animals were assessed for "social proximity" time in relation to the partner and stranger, entries into the center zone, and time spent in the center zone.

Automated video-tracking of "immobile social contact"

To process "immobile social contact," we used SocialScan 2.0. For our study, all forty-eight 3-h PPTs were compressed to MPEG files and then batch processed *post hoc*. Before processing, arenas and zones were defined in a manner similar to EthoVision, except that the

arena encompassed the entire floor of each PPT box and there were only four zones: (1) LeftSocial, (2) Center, (3) RightSocial, and (4) NonSocial (Fig. 2b). The free-moving animal is assigned subject ID #1, the tethered animal in the "LeftSocial" zone is ID #2, and the "RightSocial" zone contains subject ID #3. The RightSocial and LeftSocial zones serve as markers for social proximity and allow for identification reassignment. Adjacent animals will occasionally switch IDs; to correct for this, Clever Sys Inc. has adapted the program such that only animal ID #1 (the test animal) can enter the NonSocial zone (Fig. 2b, red). If IDs become swapped in one of the social zones, they immediately swap back once the test animal enters the NonSocial area. This process ensures that only one pair of IDs can be in social contact in the LeftSocial (#1 and #2) and RightSocial (#1 and #3) zones. Any social contact in either LeftSocial or RightSocial zones will therefore be logged as occurring between the test animal and the appropriate tethered animal (Fig. 2d).

Animal detection was accomplished using the software's default parameters, and there was no need to change contrast or brightness.

SocialScan 2.0 automatically assessed time in "immobile social contact" with each tethered animal, entries into the center zone, and time in the center. Since we are the first group to use this program to analyze partner preference in voles, the ability of the program to simultaneously detect and log "immobile social contact" was added as a custom feature and we varied the "immobility value" from 0.01–1.00. The immobility value represents a percent movement criterion beyond which adjacent animals would be considered mobile and therefore not "huddling." For example, when animals are fighting one another, they are in contact, but not huddling or immobile, and consequently this time should not be counted as prosocial contact. Lower immobility values result in a more stringent definition of "immobile social contact;" higher values, a more liberal definition. As with EthoVision 3.0, SocialScan 2.0 processed the entire 3-h test for each animal. Data were exported and analyzed.

Verification of automated scoring

Fully scoring 48 PPTs in real-time (144 h) was prohibitive. Therefore, we used a series of 15-minute-long video segments to validate the accuracy of two manual scoring methods and two automated video-tracking systems. Segments were taken at 1 h and 2.5 h into the PPT. These time points were chosen based on the *a priori* experimental objective to capture data from test animals during periods with high levels of movement within the social zones (1 h) and huddling within the social zones (2.5 h). Tests of six males and six females were scored at each time period.

Each of the 15-minute video segments was scored under the following conditions: manually in real-time by two experimenters, manually at 16x fast-forward (time-lapse) speed by the same two experimenters, once by EthoVision 3.0 in real-time, and once by SocialScan 2.0 in real-time.

Manual scoring was performed using PowerDVD 5 on a PC (Dell Dimension 4600). PowerDVD 5 allows a single jump from real-time (1x) speed to 16x speed at a bookmarked time point. With this setup, each 15-minute segment consistently played in 56.25 sec, exactly 1/16th of 15 minutes. (We note, however, that several other setups resulted in highly variable 16x playback lengths, ranging from 50 to 60 sec. Because 16x behavioral measures must be converted back to real-time for analysis, having a consistent multiplier was paramount to obtaining reliable data.)

Manual raters scored: (1) left "huddling" time, (2) center chamber time, (3) center chamber entries, and (4) right "huddling" time, using Stopwatch+ (Center for Behavioral Neuroscience: http://www.cbn-atl.org/research/stopwatch.shtml). "Huddling" was characterized by close,

physical, predominantly immobile or affiliative (e.g., grooming) contact. Center time and center chamber entries were characterized by the animal completely entering the center, socially neutral chamber (Fig. 1a).

The old version of EthoVision (3.0) and the new version of SocialScan (2.0) each scored the same video, but "huddling" was replaced by either "social proximity" or "immobile social contact," respectively. Each program had already analyzed the full 3-hour test for each animal, so a time-window function within each program was used to allow direct comparisons of the 15-minute segments with the two manual scoring methods.

Correlational analysis

The first set of analyses examined concordances between four different scoring methods: manual real-time, manual 16x speed, "social proximity," and "immobile social contact." By using time "huddling" (or the best proxy thereof), simple correlations were calculated to produce Pearson R-values. Because all the raters were blind to treatment, "huddling" time from both the left and right social zones were included (thus indiscriminately including both partners and strangers). Correlations were calculated using SPSS 15.0 (SPSS Inc., Chicago, IL). All dot plots represent individual data points and a linear regression line.

Data from the 15-minute video segments were compared across: (1) multiple raters in real-time, (2) multiple raters at 16x, (3) real-time scoring versus 16x scoring, within and across individual raters, (4) "social proximity" versus manual real-time and 16x "huddling," and (5) "immobile social contact" versus manual real-time and 16x "huddling."

Partner Preference Analysis

To examine whether differences in scoring between methods can positively or negatively affect our ability to detect a significant partner preference, a complete analysis of 48 PPTs were conducted using a 2 (sex: male vs female) $\times$ 2 (stimulus: partner vs stranger) $\times$ 2 (metric: "social proximity" vs "immobile social contact") $\times$ 2 (cohabitation: 24 h vs 1 week) ANOVA, with cohabitation and metric as repeated measures and social time as the dependent measure. All statistical tests were performed in SPSS. All bar-graphs represent mean + SEM. A $P < 0.05$ was considered significant.

Aggression

Aggression is one important social behavior that could ostensibly increase "social proximity" time without also increasing "immobile social contact" time, particularly when males are the test subjects. Twenty-four different 15-minute video segments containing only male test animals were scored by two trained observers. "Aggression" was characterized by contact that resulted in a rapid withdrawal or sustained paw slapping, nipping, or flailing. Aggression included aggressive bouts initiated by either the test male or the tethered female. Aggression was only scored manually in attempt to explain the discrepancy between "social proximity" and "immobile social contact" PPT data in males after one week of cohabitation. All bar-graphs represent mean + SEM. A $P < 0.05$ was considered significant.

Results

Verification of scoring methods

In an attempt to verify the accuracy of various manual and automated scoring methods, 15-minute video segments from 24 PPTs were scored six times for (1) left social time, (2) center chamber time, (3) center chamber entries, and (4) right social time: twice in real-time by two separate observers, twice at 16x speed by the same two observers, once by Noldus EthoVision

3.0, and once by SocialScan 2.0. Human observers measured time "huddling," EthoVision 3.0 scored time in "social proximity," and SocialScan 2.0 logged time in "immobile social contact."

A summary of the results can be found in Table 1. Overall, correlations were high between the different methods of scoring social time, center time, and center entries. Most importantly, there were high inter-rater correlations for "huddling" in real-time (R = 0.998; Fig. 3a) and at 16x speed (R = 0.982), high intra-rater reliability for "huddling" between observation speeds (1x vs 16x for two different raters: R = 0.980–0.998; Fig. 3b), and high concordance between real-time ratings of "huddling" and both automated scoring metrics: "social proximity" (R = 0.897; Fig. 3c) and "immobile social contact," with the "immobility" criterion set to 0.04 (R = 0.992; Fig. 3d). It is important to note that "immobile social contact" was examined using several different "immobility" values, ranging from 1 (equaling mere "social contact") to 0.01, but the peak concordance occurred around 0.04 and this value was used in all subsequent analyses.

Each bivariate comparison (e.g., real-time "huddling" vs 16x speed "huddling") produced a significant linear model (P < 0.001). To examine if they were equivalent, the residuals for each correlation were calculated (|predicted – observed|) and then averaged. An ANOVA and subsequent *post hoc* analyses revealed that "immobile social contact" was significantly more accurate, in terms of having smaller residuals, than "social proximity" (ANOVA, $F_{2,141}$ = 44.92, P < 0.001: "social proximity"/real-time-"huddling" |residuals| > "immobile social contact"/real-time-"huddling" |residuals|, P < 0.001 [Tukey's]). Interestingly, "immobile social contact"/real-time-"huddling" and 16X-"huddling"/real-time-"huddling" correlation residuals did not differ significantly (P = 0.643 [Tukey's]).

Automated analysis of partner preference

The essential test of an automated behavioral analysis system is whether it can detect a selective partner preference when one occurs. We compared two different automated video-tracking software packages for their ability to detect group partner preferences using four different data sets, varied by the sex of the experimental subject and the duration of cohabitation. We hypothesized that these variables were capable of altering the types of social interactions that occur during the PPT (for example, time huddling, exploring, mating, and being aggressive) and that these alterations would influence our ability to detect significant partner preferences using "social proximity" or "immobile social contact" as the social metric. PPTs for 12 females and 12 males were conducted after 24 h and 1 week of cohabitation; all 48 recorded tests were scored *post hoc* for "social proximity" and "immobile social contact." EthoVision 3.0 measured "social proximity;" SocialScan 2.0 measured "immobile social contact" (immobility criterion = 0.04).

A repeated measures, 2 (sex: male vs female) × 2 (stimulus: partner vs stranger) × 2 (social measure: "social proximity" vs "immobile social contact") × 2 (cohabitation period [repeated measures]: 24 hrs vs 1 week) ANOVA revealed a between subjects effect of stimulus animal ($F_{1,44}$ = 28.7, P < 0.001) and of social metric ($F_{1,44}$ = 97.3, P < 0.001; Fig. 4). There were also significant cohabitation period × social metric × sex ($F_{1,44}$ = 5.9, P = 0.020) and cohabitation period × social metric × stimulus animal ($F_{1,44}$ = 5.5, P = 0.024) interactions.

Planned *post hoc* Student's t-tests revealed that, in females, both "social proximity" and "immobile social contact" detected a significant partner preference during both the 24-h cohabitation PPT ("social proximity": t(17.69) = 2.825, P = 0.011; "immobile social contact": t(15.69) = 2.672, P = 0.017; Fig. 4) and the 1 week cohabitation PPT ("social proximity": t (21.15) = 4.575, P < 0.001; "immobile social contact": t(17.22) = 5.132, P < 0.001).

In males, the data were similar, but ultimately disparate. Planned *post hoc* t-tests revealed that, in males, "immobile social contact" detected a significant partner preference during both the 24-h and 1-week PPTs (24-hr: t(21.20) = 2.207, P = 0.038; 1 week: t(18.59) = 3.588, P = 0.002), while "social proximity" detected a significant partner preference after 24 h (t(21.65) = 2.73, P = 0.012), but not after 1 week (t(21.63) = 1.670, P = 0.109). This discrepancy left us with a question: what were the test males doing at the stranger end during the 1 week PPT, such that "social proximity" was increased, but "immobile social contact" was not?

Aggression

We hypothesized that an increase in aggressive behavior may account for the increase in stranger "social proximity" seen in the males at 1 week, which prevented the detection of a significant partner preference (Fig. 4).

Twenty-four 15-minute video segments (12 from the 24-hr PPT; 12 from the 1-week PPT) were scored manually by two observers in real-time for aggression and exploratory behavior within each social zone. A 2 (stimulus: partner vs stranger) $\times$ 2 (cohabitation period: 24 h vs 1 week) ANOVA, with cohabitation period as a repeated measure and aggression time as the dependent variable, revealed a between subject effect of stimulus animal ($F_{1,22} = 15.6$, P = 0.001; Fig. 5), but no within subjects effects of cohabitation period ($F_{1,22} = 0.66$, P = 0.426), nor a stimulus $\times$ cohabitation interaction (24 h vs 1 week; $F_{1,22} = 0.04$, P = 0.840). Finally, correlations between aggression time and "social proximity" time were examined to determine if the increases in stranger "social proximity" time were accounted for by increases in aggression time. Increases in aggression did not significantly lead to or follow from increases in stranger "social proximity" time (Pearson's R = 0.132; P = 0.540; data not shown). Interestingly, we did find that, at least during the 15-minute episodes we scored, males were involved in significantly more aggressive bouts at the stranger end than at the partner end, after both 24 h (P < 0.05; Fig. 5) and 1 week of cohabitation (P < 0.05).

Discussion

The partner preference test (PPT) has been and continues to be a powerful laboratory test for the study of social and sexual behavior in a wide range of species. Already shown to be sensitive to an array of treatments, the PPT promises to be a key assay in the identification of socially and sexually active manipulations in the future. Wide, extensive adoption of the PPT, however, will require a move away from the time and labor intensive investment of human scoring toward more automated methods. Already, some groups have started using automated systems to assess social behavior (Crawley, 2004; Nadler et al., 2004; Moy et al., 2008; Scearce-Levie et al., 2008). These methods, however, are limited in scope and do not account for all the complexities of the PPT adapted for monogamous prairie voles.

We attempted to broaden the use of automation by analyzing two different approaches to automated behavioral PPT scoring in voles. One approach was to use a commercially available, center-of-gravity based software package that calculates a single animal's location and its distance moved. We used this type of system to automatically when the test animal was in "social proximity" to the tethered animals. Our second approach was to use a more sophisticated software package that calculates the movement and behavior of multiple unmarked animals in a single arena. We used this software to automatically score when the test animal was in "immobile social contact" with the tethered animals. We then tested the hypothesis that "immobile social contact" would provide a more accurate approximation of manually scored "huddling" than a measure of "social proximity."

To verify that these automated metrics could provide behavioral scoring similar to human raters, we compared the results of twenty-four 15-minute video segments across four different

rating methods. We confirmed that human raters generally have high inter- and intra-rater reliability and that the automated methods of scoring "social proximity" and "immobile social contact" correlate highly with manual methods of scoring "huddling" (see Fig. 4; Table 1). Software capable of scoring "immobile social contact," however, approximated manual, real-time scoring of "huddling" to an accuracy of 99%, which was significantly better than software providing "social proximity" data (only 90% accurate). With investigators attempting to achieve at least 95% inter-rater reliability (e.g., Cushing et al., 2001), the metric of "immobile social contact" appears to be the best automated measure we examined for our prairie vole partner preference testing.

The true test of an automated system, however, is whether it can consistently find group partner preferences when they in fact occur. Our results indicate that the two automated methods produce similar, but ultimately disparate, results (see males at 1 week; Fig. 4). With different results of partner preference formation in hand, investigators are likely come to wholly different conclusions. For example, the automated "social proximity" data suggest that males rapidly form selective partner preferences but fail to express them during a PPT after 1 week. Alternatively, the automated "immobile social contact" data suggest that males display selective partner preferences over an extended period of time.

Since these data were obtained from the same set of video-recorded tests, the comparison clearly demonstrates that the move to automated systems cannot proceed indiscriminately. The assumption that "social proximity" is essentially equivalent to "huddling" as rated by humans could lead to erroneous conclusions.

"Social proximity," however, did provide comparable results for females at both time points and for males during the post-24-hr PPT. Thus, automated measures of "social proximity" may be a cost-effective and relatively accurate alternative in many circumstances, whether by video-tracking, beam-breaks, or some other system. For example, with photobeams placed just beyond the reach of the tethered stimulus animals, we found that Vole Tracker (R. Henderson, Florida State University), which has been used to measure locomotor activity in a number of studies (Curtis et al., 2001; Aragona et al., 2003; Curtis and Wang, 2005), provided nearly identical "social proximity" results as our simplified use of Noldus EthoVision, 3.0 (unpublished data). This suggests that "social proximity" can be obtained through multiple methods, and they are likely to have comparable accuracy.

As we predicted, "immobile social contact" provided a measure of "huddling" that was significantly more accurate than "social proximity." (Table 1). Furthermore, this automated metric successfully detected partner preferences after 24 hours and 1 week of cohabitation in both males and females. Both analyses confirmed our central hypothesis that automated "immobile social contact" would better reflect manually scored "huddling" than the measure of "social proximity."

Although we confirmed our hypothesis, we were intrigued by the factors that resulted in inaccuracies or discrepancies between scoring methods. As noted, males after 1 week of cohabitation demonstrated a significant partner preference with "immobile social contact" as the metric, but not with "social proximity"—primarily because of a relative increase in time spent in "social proximity" to the stranger (Fig 4). Based on known behavioral changes in males after extended cohabitation periods (Aragona et al., 2006), we examined whether an increase in stranger-directed aggressive bouts could explain the discrepancy. Our findings suggest it does not. Qualitative reanalysis indicated that different animals increased stranger "social proximity" due to a number of differing factors between animals. For some, it was aggression, for others it was chewing on the stranger's tethering anchor, for others it was pushing the

bedding around at the stranger end, etc. In sum, the test male behaviors that increased "social proximity" time in the social zone with stranger were many and varied.

A similar pattern of variation in behavior was seen in the 15-minute video segments used for our correlational analyses. We had initially thought that part of the reason our automated systems were not 100% accurate is that they might occasionally lose track of animals. However, this was rarely the case since the cages were well-lit, providing high-contrast between the dark animals and the light-colored bedding, and the cages were viewed from above.

Behaviorally, the sources of inaccuracy were more obvious, but difficult to quantify. During the measure of "social proximity," test animals would occasionally cross the boundary into one of the social zones and then proceed to sit, explore, dig, or attempt to climb the wall in a part of the zone removed from the tethered animal. When these behavioral bouts extended for long periods of time, they would substantially increase the amount of "social proximity" time while having no effect on "huddling" time. Disparities such as these ultimately had a deleterious effect on the overall accuracy of "social proximity" as a social metric.

For the measure of "immobile social contact," the reasons for inaccuracy were more subtle. One of the sources was the process of defining the outer edge of each animal. At times animals will sit near each other, but with a space between them. A human will see this space, whereas the software may overestimate the animal outlines enough to create a point of contact. The software would therefore count "immobile social contact," but the human rater would not score "huddling." Another source of inaccuracy likely depends on how long "immobile social contact" bouts need to extend before a human will rate it as "huddling." Occasionally, animals investigate one another and pause briefly. The automated software is likely to add a second or two of "immobile social contact" time here, whereas a human would not. Both sources of inaccuracy could, in theory, be modified by changing specific parameters within the software's interface. At approximately 99% accuracy using SocialScan 2.0's default parameters and our testing conditions, minor adjustments to achieve slightly better accuracy seemed unnecessarily burdensome.

Achieving a respectable level of accuracy was not the only benefit to using these systems. Automated scoring also greatly decreased our time investment. Forty-eight PPTs would have taken 144 hours to score manually in real-time. Even using time-compression equipment, manual scoring would have taken 9–11 hours (requiring three to seven days to score fully, in order to avoid fatigue and drift). Alternatively, once standard virtual arenas were set, only 30–60 minutes of human time were required to obtain complete automated results by the morning following the final tests. This equals more than an 85% reduction in the expenditure of human time and a 67–86% reduction in the time delay to obtain complete data sets. This drastic reduction in time without a substantial reduction in accuracy suggests that the PPT in combination with the automated scoring of "immobile social contact" is ready for high-throughput discovery.

The success of automated video-tracking systems to analyze social behavior in prairie voles provides evidence that these systems could easily be extended to other rodent models, in which stimulus animals have restricted movement. While mating behavior may still have to be scored manually, many other social metrics could be automated. For instance, sophisticated video-based systems like Clever Sys Inc.'s SocialScan 2.0, Noldus' newer versions of EthoVision (e.g., XT 6.0), and potentially others, claim the ability to measure a host of social metrics, including approach latencies, time in social contact, distance from a social object, avoidance, stretch-attend postures, and even sniffing. In the future, it might be possible to couple these behavioral metrics with a measure of "immobile social contact" to provide a more complete picture of social behavior within the PPT, although we have not yet attempted this. In

conclusion, video-based behavioral analysis software capable of tracking multiple animals in a relatively complex social context may provide the most flexible, accurate, comprehensive, and time-efficient approach to studying social behavior on a large scale. Our results demonstrate that, already, automated video-tracking systems can be adapted to score experimentally relevant behavioral metrics, such as "social proximity" and "immobile social contact," critical for the assessment of partner preference and they appear ready to start replacing human scoring methods.

Acknowledgments

References

Aragona BJ, Liu Y, Curtis JT, Stephan FK, Wang Z. A critical role for nucleus accumbens dopamine in partner-preference formation in male prairie voles. J Neurosci 2003;23:3483–3490. [PubMed: 12716957]

Aragona BJ, Liu Y, Yu YJ, Curtis JT, Detwiler JM, Insel TR, Wang Z. Nucleus accumbens dopamine differentially mediates the formation and maintenance of monogamous pair bonds. Nat Neurosci 2006;9:133–139. [PubMed: 16327783]

Baum MJ, Erskine MS, Kornberg E, Weaver CE. Prenatal and neonatal testosterone exposure interact to affect differentiation of sexual behavior and partner preference in female ferrets. Behav Neurosci 1990;104:183–198. [PubMed: 2317276]

Bosch OJ, Nair HP, Ahern TH, Neumann ID, Young LJ. The CRF system mediates increased passive stress-coping behavior following the loss of a bonded partner in a monogamous rodent. Neuropsychopharmacology 2009;34:1406–15. [PubMed: 18923404]

Carter CS, DeVries AC, Getz LL. Physiological substrates of mammalian monogamy: the prairie vole model. Neurosci Biobehav Rev 1995;19:303–314. [PubMed: 7630584]

Crawley JN. Designing mouse behavioral tasks relevant to autistic-like behaviors. Ment Retard Dev Disabil Res Rev 2004;10:248–258. [PubMed: 15666335]

Curtis JT, Wang Z. Ventral tegmental area involvement in pair bonding in male prairie voles. Physiol Behav 2005;86:338–346. [PubMed: 16165168]

Curtis JT, Liu Y, Wang Z. Lesions of the vomeronasal organ disrupt mating-induced pair bonding in female prairie voles (Microtus ochrogaster). Brain Res 2001;901:167–174. [PubMed: 11368964]

Cushing BS, Martin JO, Young LJ, Carter CS. The effects of peptides on partner preference formation are predicted by habitat in prairie voles. Horm Behav 2001;39:48–58. [PubMed: 11161883]

Grippo AJ, Cushing BS, Carter CS. Depression-like behavior and stressor-induced neuroendocrine activation in female prairie voles exposed to chronic social isolation. Psychosom Med 2007;69:149–157. [PubMed: 17289829]

Insel TR, Hulihan TJ. A gender-specific mechanism for pair bonding: oxytocin and partner preference formation in monogamous voles. Behav Neurosci 1995;109:782–789. [PubMed: 7576222]

Johnson WA, Tiefer L. Sexual preferences in neonatally castrated male golden hamsters. Physiol Behav 1972;9:213–217. [PubMed: 4654735]

Lim MM, Wang Z, Olazabal DE, Ren X, Terwilliger EF, Young LJ. Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. Nature 2004;429:754–757. [PubMed: 15201909]

Lim MM, Liu Y, Ryabinin AE, Bai Y, Wang Z, Young LJ. CRF receptors in the nucleus accumbens modulate partner preference in prairie voles. Horm Behav 2007;51:508–515. [PubMed: 17320879]

Moy SS, Nadler JJ, Young NB, Nonneman RJ, Segall SK, Andrade GM, Crawley JN, Magnuson TR. Social approach and repetitive behavior in eleven inbred mouse strains. Behav Brain Res 2008;191:118–129. [PubMed: 18440079]

Nadler JJ, Moy SS, Dold G, Trang D, Simmons N, Perez A, Young NB, Barbaro RP, Piven J, Magnuson TR, Crawley JN. Automated apparatus for quantitation of social approach behaviors in mice. Genes Brain Behav 2004;3:303–314. [PubMed: 15344923]

Noldus LP, Spink AJ, Tegelenbosch RA. EthoVision: a versatile video tracking system for automation of behavioral experiments. Behav Res Methods Instrum Comput 2001;33:398–414. [PubMed: 11591072]

Scearce-Levie K, Roberson ED, Gerstein H, Cholfin JA, Mandiyan VS, Shah NM, Rubenstein JL, Mucke L. Abnormal social behaviors in mice lacking Fgf17. Genes Brain Behav 2008;7:344–354. [PubMed: 17908176]

Slob AK, Bogers H, van Stolk MA. Effects of gonadectomy and exogenous gonadal steroids on sex differences in open field behavior of adult rats. Behav Brain Res 1981;2:347–362. [PubMed: 7225223]

Slob AK, de Klerk LW, Brand T. Homosexual and heterosexual partner preference in ovariectomized female rats: effects of testosterone, estradiol and mating experience. Physiol Behav 1987;41:571–576. [PubMed: 3441528]

Williams JR, Catania KC, Carter CS. Development of partner preferences in female prairie voles (Microtus ochrogaster): the role of social and sexual experience. Horm Behav 1992;26:339–349. [PubMed: 1398553]

Winslow JT, Hastings N, Carter CS, Harbaugh CR, Insel TR. A role for central vasopressin in pair bonding in monogamous prairie voles. Nature 1993;365:545–548. [PubMed: 8413608]

Young LJ, Wang Z. The neurobiology of pair bonding. Nat Neurosci 2004;7:1048–1054. [PubMed: 15452576]
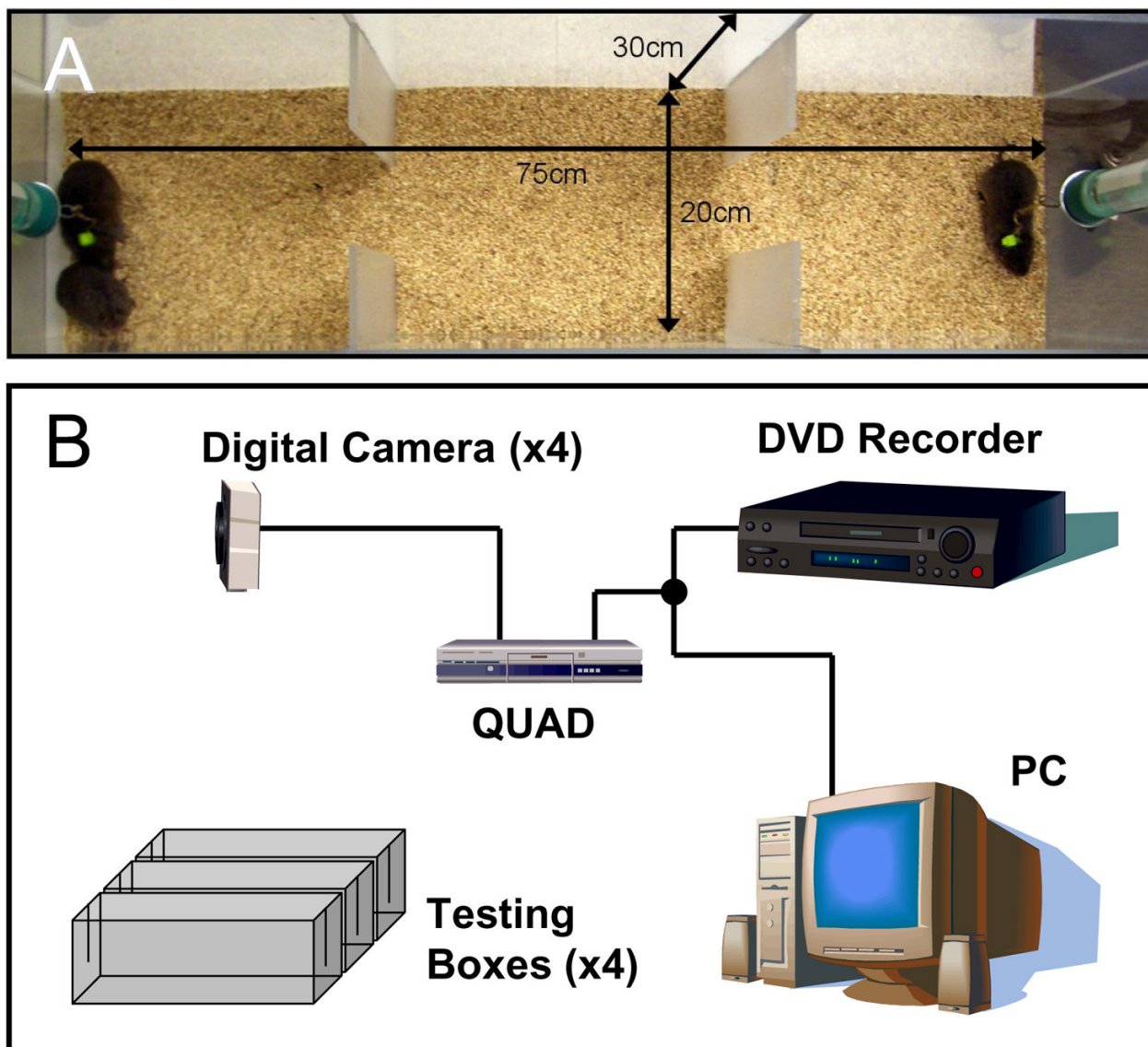
Figure 1.
Partner preference tests (PPTs) were conducted in a three-chambered arena (A). Stimulus animals (one partner and one stranger) were loosely tethered at each end; the test subject roamed freely for 3 hours. 12 PPT boxes were monitored by four cameras (B). The four video streams are compressed to a single output by a QUAD box and sent to both a DVD recorder and PC.
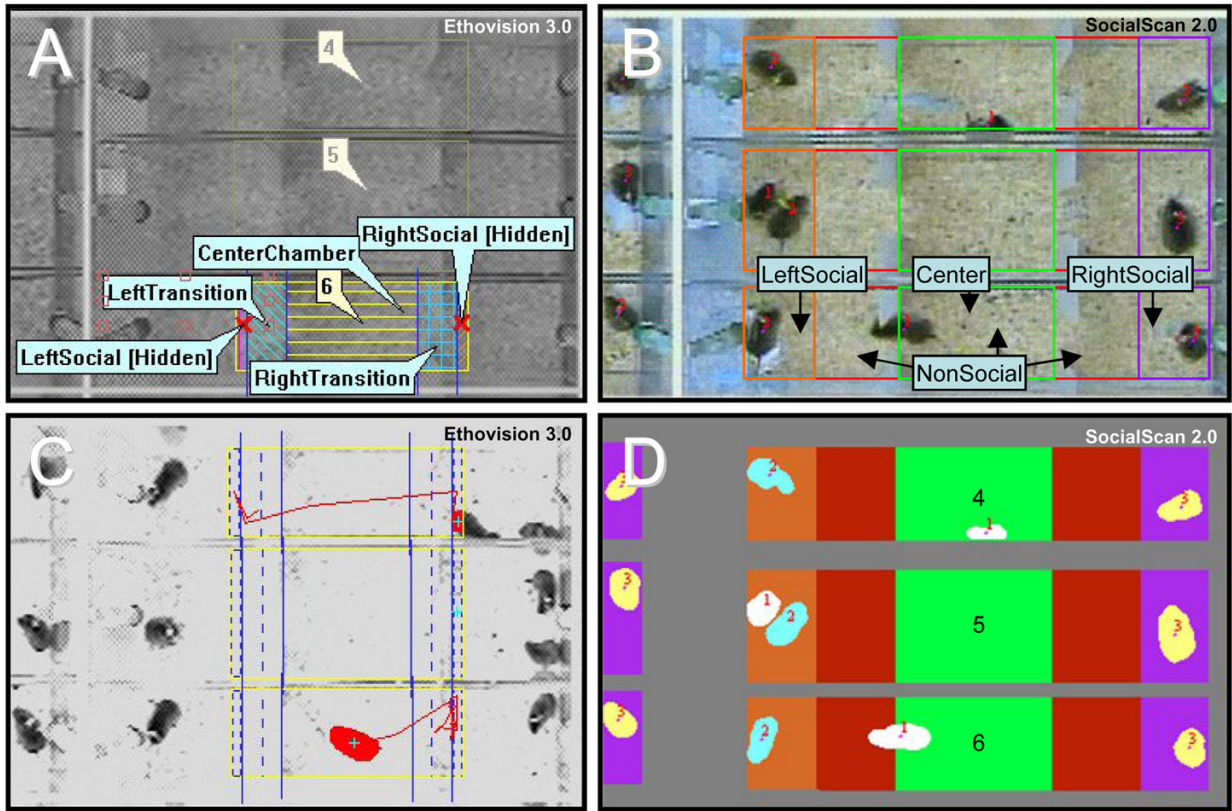
Figure 2.
An early version of EthoVision (version 3.0) and the newest version of SocialScan (version 2.0) were used to automatically process vole behavior in the partner preference test (PPT). Virtual arenas (area in which voles can be tracked) and zones (regions of interest) were outlined for EthoVision (A) and SocialScan (B). For EthoVision 3.0, arenas were truncated at a point beyond which the tethered animals could not reach; zones included LeftSocial, RightSocial, two transitions zones, CenterChamber, and NonSocial (an accumulation of all non-hidden zones); the social zones were designated as hidden, allowing the program to track the amount of time the test subject spent in "social proximity" to the left and right tethered animals. For SocialScan 2.0, arenas included the entire test box floor; zones included LeftSocial, RightSocial, Center, and NonSocial; the NonSocial zone served to correct occasional ID swaps. During analysis, both EthoVision 3.0 and SocialScan 2.0 tracked the animals' location according to center-of-gravity. EthoVision 3.0 logged time in "social proximity" (C, see middle test box); SocialScan 2.0 logged time in "immobile social contact" (D, see middle test box).
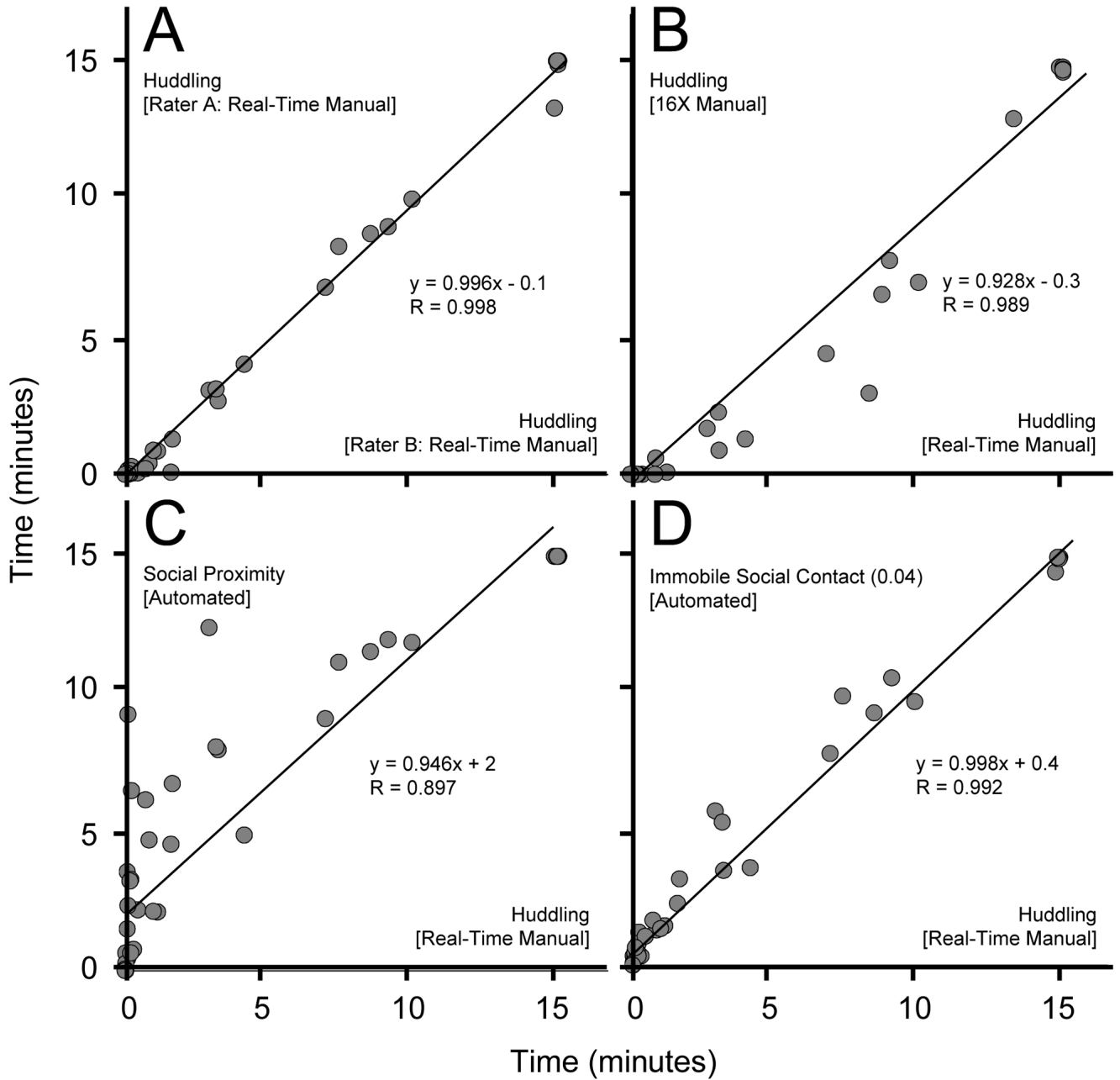
Figure 3.
Dot-plots and linear correlations were generated for key scoring method comparisons to illustrate variation. Plots demonstrate that there is high inter-rater reliability in real-time (A) and intra-rater reliability across scoring speeds (B). The metric of "immobile social contact" (D) approximated manually scored "huddling" better than the metric of "social proximity" (B). Equations (y = mx + b) and Pearson's Rs illustrate how closely the linear models approximate a 1:1 ratio with no scale shift and the degree of concordance, respectively.
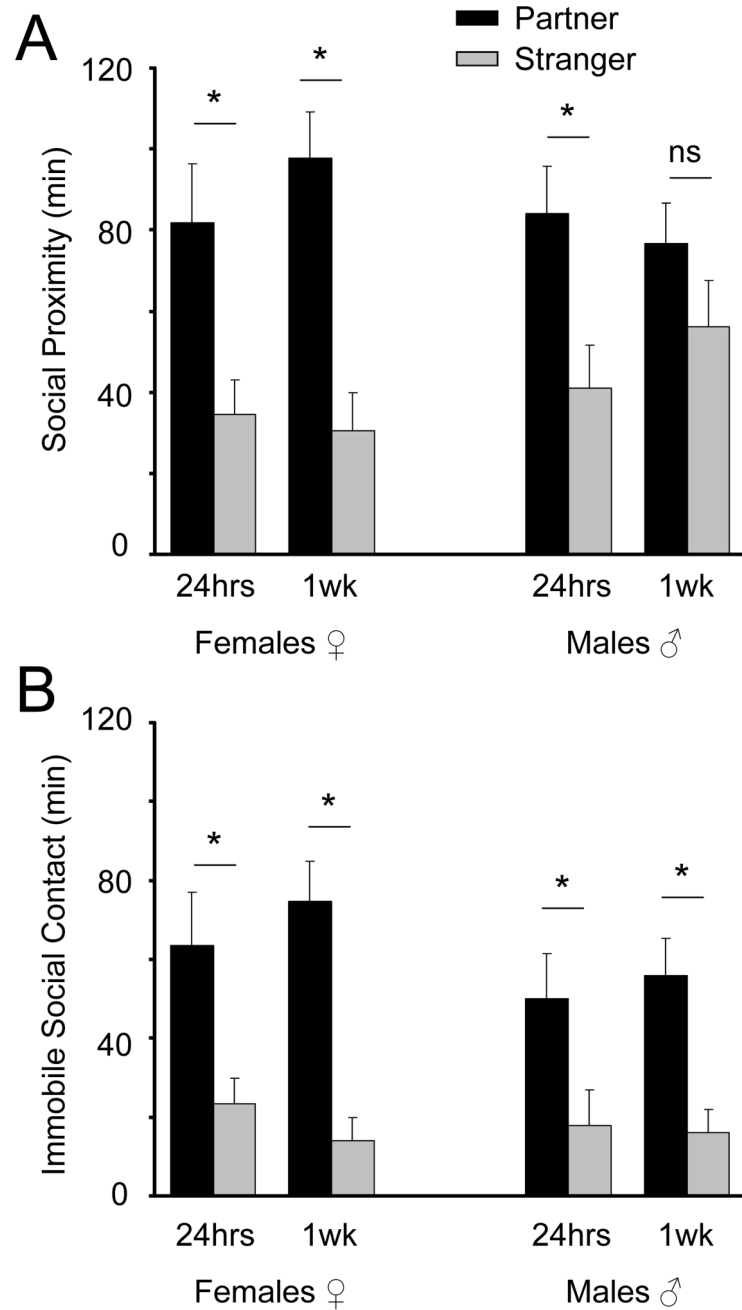
Figure 4.
Males and females underwent two separate 3-hour PPTs, once after 24-hours of cohabitation, and once after a full week of cohabitation. All tests were processed *post hoc* and assessed for "social proximity" time (A) and "immobile social contact" time (B). The "immobile social contact" metric detected a significant partner preference in all four data sets; the "social proximity" failed to detect a partner preference in males after 1 week of cohabitation. Bar-graphs represent mean + SEM; (*) represents a statistically significant preference for the partner over the stranger, P < 0.05. "ns" = not significant.
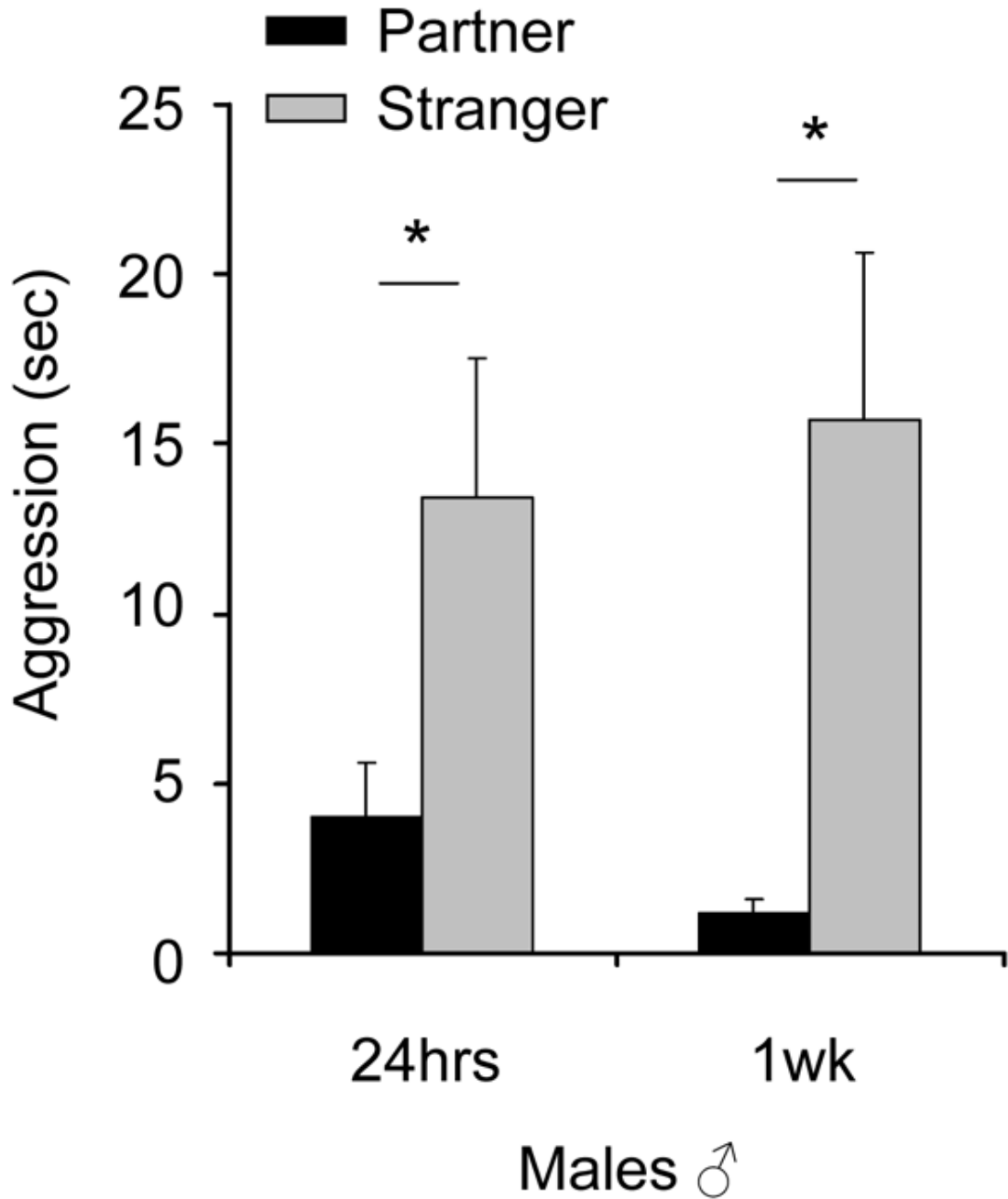
Figure 5.
Males were assessed for aggression in twenty-four 15-minute video segments, half during the post-24 hour cohabitation PPT test and half 1-week PPT. Males spent more time in aggressive scuttles with stranger females than with their female partners. Bar-graphs represent mean + SEM. (*) P < 0.05.

Table 1

Manual and automated scoring methods were used to analyze twenty-four 15-minute video segments for social behavior, center zone time, and center zone entries. Each scoring method was correlated to each of the others and a Pearson's R-value was calculated. **Bolded** R-values met an arbitrarily good correlation cutoff of 0.95. All social contact data were produced by SocialScan 2.0. The values in parentheses denote the "immobility" value used by SocialScan 2.0 to distinguish "immobile social contact" from mere "social contact." All correlations produced significant linear models (P < 0.001).

Social time			Pearson's R
Real-time manual huddling	vs	16X manual huddling	**0.989**
		Social proximity (EthoVision 3.0)	0.897
		Social proximity (SocialScan 2.0)	0.904
		Social contact	**0.972**
		Immobile (0.05) social contact	**0.987**
		Immobile (0.04) social contact	**0.992**
		Immobile (0.03) social contact	**0.991**
		Immobile (0.02) social contact	**0.960**
16x manual huddling	vs	Social proximity	0.879
		Immobile (0.04) social contact	**0.980**
Center zone time			**Pearson's R**
Real-time manual	vs	16x manual	**0.996**
		EthoVision 3.0	**0.998**
		SocialScan 2.0	**0.999**
16x manual	vs	EthoVision 3.0	**0.993**
		SocialScan 2.0	**0.980**
Center zone entries			**Pearson's R**
Real-time manual	vs	16x manual	0.925
		EthoVision 3.0	0.814
		SocialScan 2.0	**0.986**
16x manual	vs	EthoVision 3.0	0.815
		SocialScan 2.0	0.950