CHAPTER THREE

DECONSTRUCTING SOCIABILITY, AN AUTISM-RELEVANT PHENOTYPE, IN MOUSE MODELS

Andrew H. Fairless[1], Rhia Y. Shah[1], Ashley J. Guthrie[1], Hongzhe Li[2], Edward S. Brodkin[1]

In press at *The Anatomical Record*

[1]Center for Neurobiology and Behavior, Department of Psychiatry, University of Pennsylvania School of Medicine, Translational Research Laboratory, 125 South 31st Street, Room 2220, Philadelphia, PA 19104-3403 USA

[2]Statistical Genetics and Genomics Laboratory, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 215 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021 USA

Correspondence should be addressed to Edward S. Brodkin, tele (215)-746-0118; fax: (215)-573-2041; email ebrodkin@mail.med.upenn.edu

Manuscript correspondence should be addressed to Edward S. Brodkin

Grant sponsor(s): NIMH; Burroughs Wellcome Fund
Grant number(s): R01MH080718; 5-T32-MH017168-26

Key words: sociability; behavior; mouse; model; autism; endophenotype

Abstract

Reduced sociability is a core feature of autism spectrum disorders (ASD) and is highly disabling, poorly understood, and treatment refractory. To elucidate the biological basis of reduced sociability, multiple laboratories are developing ASD-relevant mouse models, in which sociability is commonly assessed using the Social Choice Test. However, various measurements included in that test sometimes support different conclusions. Specifically, measurements of time the "test" mouse spends near a confined "stimulus" mouse (chamber scores) sometimes support different conclusions from measurements of time the test mouse sniffs the cylinder containing the stimulus mouse (cylinder scores). This raises the question of which type of measurements are best for assessing sociability. We assessed the test-retest reliability and ecological validity of chamber and cylinder scores. Compared with chamber scores, cylinder scores showed higher correlations between test and retest measurements, and cylinder scores showed higher correlations with time spent in social interaction in a more naturalistic phase of the test. This suggests that cylinder scores are more reliable and valid measures of sociability in mouse models. Cylinder scores are reported less commonly than chamber scores, perhaps because little work has been done to establish automated software systems for measuring the former. In this study, we found that a particular automated software system performed at least as well as human raters at measuring cylinder scores. Our data indicate that cylinder scores are more reliable and valid than chamber scores, and that the former can be measured very accurately using an automated video analysis system in ASD-relevant models.

INTRODUCTION

Reduced sociability (reduced tendency to seek social interaction) is a core feature of autism spectrum disorders (ASD) that is highly disabling and for which effective treatments are sorely lacking (Hill and Frith, 2003). The biological basis of social impairments in ASD remains largely unexplained. A vital research priority is to develop mouse models relevant to ASD, because the experimental control that model systems afford will be indispensable for unraveling the complex biological basis of sociability deficits. An important part of these efforts to develop model systems will be refining methods for measuring sociability in mice.

Social affiliative behaviors have been studied in mice using several experimental paradigms. Mice may be observed freely interacting in a novel environment (Social Interaction Test) (de Angelis and File, 1979, File and Seth, 2003) or in their home cages (Lijam et al., 1997, Mondragon et al., 1987, Terranova et al., 1994) to allow quantification of naturalistic behaviors, including passive social behaviors, such as huddling together, as well as more active social behaviors, such as sniffing and allogrooming. An advantage of these assays is their ecological validity: the behaviors of these freely interacting mice resemble those of feral mice in their natural environments. A disadvantage is the complexity of these social interactions: because either or both mice can easily initiate, maintain, or modulate an interaction, disentangling the contributions that each mouse makes to the interaction can be difficult.

By contrast, more controlled social affiliation assays appear to be less ecologically valid but simplify the social interaction, making it more feasible to measure the tendency of a specific mouse to approach or avoid another mouse. These controlled

assays include the Partition Test (Kudryavtseva, 2003, Moretti et al., 2005, Spencer et al., 2005) and the Social Choice (or Social Approach) Test (Brodkin et al., 2004, Moy et al., 2004, Nadler et al., 2004, Sankoorikal et al., 2006). The Social Choice Test is conducted in a three-chambered apparatus, or box, with a transparent, air-permeable cylinder in each of the two end chambers. After a period in which the "test" mouse is habituated to the apparatus, a "stimulus" or "target" mouse is then confined inside one cylinder, so that the stimulus mouse cannot approach the test mouse to initiate or maintain a social interaction. With only enough space to turn around, the stimulus mouse is always close to or against the cylinder wall, so that it can be easily sniffed through the holes in the cylinder wall and otherwise investigated by the test mouse, which is free to move throughout the box. This high degree of control limits affiliative behaviors, especially by the stimulus mouse, but ensures that any active social interaction can occur only if the test mouse initiates and maintains that interaction. Thus, the social choice assay is well suited for isolating and quantitatively measuring the sociability of the test mouse.

One may analyze sociability in the Social Choice Test using any of several measures. "Social chamber time" can be defined as the amount of time that the test mouse spends in the end chamber that contains the stimulus mouse, and "social cylinder time" can be defined as the amount of time that the test mouse sniffs and otherwise investigates the cylinder that contains the stimulus mouse. One may calculate "chamber/cylinder preference" scores and "chamber/cylinder preference change" scores (see Methods and Materials), which theoretically may improve control of the analysis (Sankoorikal et al., 2006). Thus, one may potentially assess sociability by six different

but related scores: social chamber time, social cylinder time, chamber preference, cylinder preference, chamber preference change, and cylinder preference change.

This multiplicity of scores raises the question of whether any of the scores are more valid than the others. If any score always yields the same conclusion as the other scores, then they are redundant. On the other hand, if the scores sometimes disagree, then on which resulting conclusion should one rely for further research?

The scores usually support the same conclusions (Crawley et al., 2007, Nadler et al., 2004, Ryan et al., 2008, Sankoorikal et al., 2006, Yang et al., 2007b), but they sometimes disagree with each other, and this has created some ambiguity in published studies. Moy and colleagues (2007) tested 10 inbred strains of mice for sociability and, in an analysis akin to a preference score, identified three strains for which chamber and cylinder scores disagreed on whether the strains should be considered sociable. Similar disagreements occurred in one cohort of vasopressin receptor 1B (Avpr1b) null mutants and heterozygotes tested during the circadian light phase (Yang et al., 2007a, Fig. 3I,J) and in one cohort of Fragile X mental retardation 1 mutants ($Fmr1^{-/y}$) on a FVB/129 genetic background (Moy et al., 2009, Figs. 2B, 3A). Fairless et al. (2008) hypothesized a positive correlation within the BALB/cJ inbred mouse strain between sociability and the size of the corpus callosum. The chamber preference change score showed no such correlation, yet the cylinder preference change score did. Such discrepancies may be resolved by determining which, if any, of the sociability scores are more valid than the others.

There are many criteria by which to assess the validity (henceforth called “general validity”) of a measurement. One criterion is test-retest reliability. Assuming that a

behavior is temporally stable and is not substantially changed by the testing procedure, measurements of that behavior in the same mice at different times should yield a positive correlation. This approach has been used to study rodent behaviors in the elevated plus maze (Andreatini and Bacellar, 2000, Lister, 1987, Rodgers et al., 1997), forced swim test (Drugan et al., 1989, Hilakivi and Lister, 1990), free-exploratory paradigm (Teixeira-Silva et al., 2009), open field (Henderson, 2005) and other experimental paradigms. A second criterion by which to generally validate an experimental measurement is its ecological validity, or how closely the measurement relates to behaviors in naturalistic situations. One way to assess ecological validity is to measure the correlation between behavior in one test with behavior in a more naturalistic or ecologically relevant test. Other criteria of general validity of animal models of human disease, such as etiological/construct validity and predictive validity (Crawley, 2004) are beyond the scope of the present study.

To assess the test-retest reliability of the six sociability scores, we re-analyzed data from a previously published experiment (Sankoorikal et al., 2006) in which the same mice underwent the Social Choice Test once per day on two consecutive days. At the conclusion of the Social Choice Test on the second day, the cylinders were removed so that the test and stimulus mice could both move about and interact freely. This phase of the test closely resembled the Social Interaction Test and was more naturalistic than the Social Choice Test. To assess the ecological validity of the six sociability scores, we correlated those scores with the amount of time that the test mouse investigated the stimulus mouse during this “Free Social Interaction” period. We hypothesized that the

chamber preference change score would show the highest test-retest reliability and that the cylinder preference change score would show the highest ecological validity.

Our data analysis in the present study indicates that cylinder scores are more reliable and ecologically valid measures than chamber scores for measuring sociability in the Social Choice Test. Automated tools that locate a mouse are well established and allow chamber scores in the Social Choice Test to be easily obtained (Nadler et al., 2004, Page et al., 2009). However, fewer automated methods exist for recording sniffing and other active behaviors on which the cylinder scores are based. Here we evaluate software that can accurately measure cylinder scores. We hypothesized that an automated video analysis system could measure cylinder scores with accuracy comparable to that of a human rater.

Finally, some have suggested that a measurement of the mouse's proximity to the social cylinder – that is, a measurement of the amount of time that the test mouse spends in an area that is near the social cylinder but smaller than the entire social chamber– would be a valid alternative measurement of sociability (alternative to chamber and cylinder scores) in the Social Choice Test (Page et al., 2009). We hypothesized that this alternative measurement would correlate highly with social cylinder time and thus provide an adequate substitute for directly measuring cylinder scores.

MATERIALS AND METHODS

Experiment 1: General validity of six sociability scores

The general validity of the six sociability scores was investigated by analyzing archived data from a previously published experiment, and a comprehensive description

of that experiment's methods can be found in the prior report (Sankoorikal et al., 2006). Briefly, mice were obtained from The Jackson Laboratory (Bar Harbor, ME) and housed at the University of Pennsylvania on a 12-hour light-dark cycle with the light cycle occurring from 7:00 a.m. – 7:00 p.m. Food and water were available to the mice ad libitum. Test mice were male and female C57BL/6J and BALB/cJ mice that were tested for sociability at 4 or 9 weeks-of-age, which was 5 – 7 days following their arrival. Test mice were housed 2 females or 2 males to a cage. Stimulus mice were 4-week-old DBA/2J mice that were housed 4 females or 4 males to a cage. The 4-week-old, prepubescent test mice and the 9-week-old adult test mice were separate cohorts; the 9-week-old mice had not been previously tested for sociability. All animals were treated according to the National Institutes of Health Guide for the Care and Use of Laboratory Animals, and all procedures were approved by the University of Pennsylvania Institutional Animal Care and Use Committee.

All testing occurred in a dimly lit (<7 lux) , sound attenuated behavioral testing room between noon and 5 p.m., during the light phase of the light-dark cycle. Mice were brought in their home cages from the colony room to the behavioral testing room and were allowed to habituate to the behavioral testing room for ~30 minutes prior to the start of the social choice test. The test was conducted in a three-chambered box with a cylinder in each of the two end chambers (Fig. 1A,B). The test mouse was initially placed into the middle chamber and allowed to explore all 3 chambers during a 5-min "Habituation" period (Phase 1). After Phase 1, a DBA/2J stimulus mouse was placed into one of the two cylinders. The stimulus mouse was prepubescent and of the same sex as the test mouse to minimize any sexual or aggressive motivations of the test mouse.

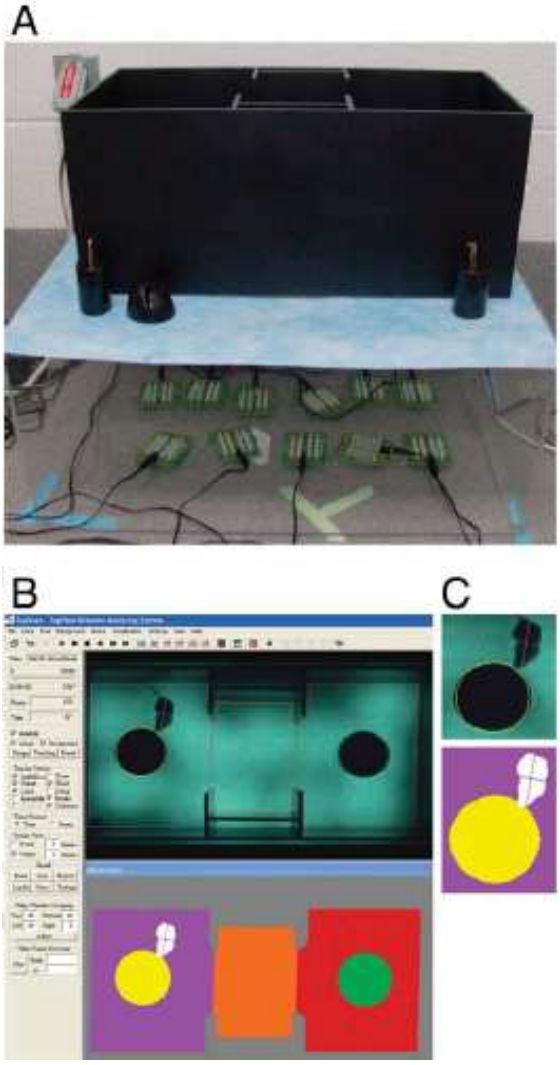

Fig. 1. **(A)** A side-view of the three-chambered box used for sociability testing. For automated scoring, the box sits on a blue mat that covers a transparent Plexiglas stand. Below are 11 Infrared (LED) Lighting Kits that illuminate the floor of the box. In front of the box are two 0.5-kg weights that are placed on top of the cylinders, and the paper weight that serves as a novel object. **(B)** A screenshot from TopScan (Clever Sys., Inc.) software. The upper portion shows the video image, in which an albino (BALB/cJ) mouse sniffs the left cylinder and is silhouetted against the infrared-lit floor. The lower portion shows the software's location of the mouse, both cylinders, and each chamber. **(C)** A close-up of the mouse sniffing the cylinder. The yellow circle outline in the top panel shows where the software operator has drawn the cylinder's outline. The yellow circle in the bottom panel indicates the cylinder's location, including a 4-mm extension of the radius.

The walls of the transparent cylinders contained holes that allowed the mice to sniff each other. During the next 5-min period ("Social Choice," Phase 2), the test mouse could approach and affiliate with the stimulus mouse. During Phases 1 and 2, the experimenters recorded the amounts of time that the test mouse spent in the end chamber that contained the stimulus mouse (the "social chamber") and the chamber that did not (the "nonsocial chamber") and the amounts of time that the test mouse sniffed and

otherwise investigated the cylinder that contained the stimulus mouse (the "social cylinder") and the cylinder that did not (the "nonsocial cylinder").

Mice were tested on two consecutive days. On Day 1, mice were tested in Phases 1 (Habituation) and 2 (Social Choice). On Day 2, mice underwent Phases 1, 2, and 3 (Free Social Interaction). The test mouse was paired with a different DBA/2J stimulus mouse on Day 2 from the one it was paired with on Day 1. During the Free Social Interaction period, which occurred immediately following Phase 2 on Day 2, the cylinders were removed for a 5-min period during which the test and stimulus mice could both move about the three-chambered box and interact freely (Phase 3). The experimenters recorded how much time the mice were in direct contact, which included time that the two mice spent sniffing and/or allogrooming each other.

Of the 161 mice originally tested (Sankoorikal et al., 2006), 10 were excluded from all analyses here because their data from Phase 1 or Phase 2 of the test were incomplete. Another three mice were excluded because they showed aggression (vigorously lunging at and biting the stimulus mouse) during the Free Social Interaction (Phase 3), suggesting that their earlier motivations for social approach and investigation during Phase 2 may have been aggressive. Data from the remaining 148 mice were analyzed for test-retest reliability. These mice were 4-week-old C57BL/6J females (n = 18), 9-week-old C57BL/6J females (n = 17), 4-week-old C57BL/6J males (n = 20), 9-week-old C57BL/6J males (n = 18), 4-week-old BALB/cJ females (n = 20), 9-week-old BALB/cJ females (n = 19), 4-week-old BALB/cJ males (n = 18), and 9-week-old BALB/cJ males (n = 18).

Sociability was assessed by six scores: social chamber time, social cylinder time, chamber preference, cylinder preference, chamber preference change, and cylinder preference change. Social chamber time was the total number of seconds that the mouse spent in the social chamber in Phase 2 (Social Choice). Social cylinder time was the total number of seconds that the mouse spent sniffing, scratching, gnawing, or otherwise investigating the social cylinder (with the head pointed at the cylinder) in Phase 2.

Chamber and cylinder preference scores were each calculated by subtracting the nonsocial chamber/cylinder time from the social chamber/cylinder time during Phase 2 (Social Choice). Thus, a positive preference score indicated that a mouse spent more time in the social chamber than in the nonsocial chamber or more time sniffing the social cylinder than sniffing the nonsocial cylinder. In other words, the mouse "preferred" the social chamber or social cylinder. Likewise, a negative preference score indicated a preference for the nonsocial chamber/cylinder, and a preference score of zero indicated no preference. Preference scores were designed to exclude a mouse's tendency to explore any nonsocial stimulus (chamber or cylinder) and leave only the social investigatory component.

Both preference change scores were calculated by subtracting the chamber/cylinder preference scores during Phase 1 (Habituation) from the chamber/cylinder preference scores during Phase 2 (Social Choice). Thus, a mouse with a positive preference change score preferred the social chamber/cylinder over the nonsocial chamber/cylinder more so during Phase 2 (Social Choice) than in Phase 1 (Habituation). Likewise, a negative preference change score indicated a change in preference towards the nonsocial chamber/cylinder from Phase 1 to Phase 2, and a

preference change score of zero indicated no change in preference between the phases. While the two end chambers were virtually identical, individual mice usually show some variation in their preference scores during Phase 1 (Habituation). The preference change scores may correct for this individual variation.

Test-retest reliability was assessed on the six sociability scores. Because the mice represented eight different experimental groups, the data were analyzed in a general linear model with strain (C57BL/6J vs. BALB/cJ), sex (female vs. male), and age (4 weeks vs. 9 weeks) as factors: $y = \alpha + \beta_1$ (strain) + β_2 (sex) + β_3 (age) where α, β_1, β_2, and β_3 are the regression coefficients. The residuals from the model were calculated for all scores, and the residuals of each of the six scores from Day 1 were correlated with the residuals of the corresponding score on Day 2 using the Pearson product-moment correlation coefficient (Pearson's *r*). These correlations were not independent and were compared using methods described in Steiger (1980) where the covariances among the Fisher-transformed values of these correlations are taken into account.

To simplify the analysis and control the familywise error rate, the six scores were initially compared in two sets of group comparisons. In the first set, the three chamber scores were compared to the three cylinder scores. This comparison answered the question: are chamber scores more or less reliable than cylinder scores? In the second set, three groups were compared: the social chamber/cylinder times, the chamber/cylinder preference scores, and the chamber/cylinder preference change scores. This comparison answered the question: are any of these three types of scores (i.e., "time", "preference", and "preference change") more or less reliable than the others? If

either set of comparisons showed a statistically significant difference, appropriate pairwise comparisons were performed (Steiger, 1980).

To address the question of ecological validity, data from 146 mice were used to compare the six sociability scores of Phase 2 (Social Choice) to Phase 3 (Free Social Interaction). Phase 3 data were missing for two other mice, a 4-week-old C57BL/6J male and a 9-week-old BALB/cJ male. Pearson's *r* was calculated on the residuals of a general linear model, as described above, to correlate each of the six sociability scores of Phase 2 to the amount of time the test mouse spent sniffing and otherwise investigating the stimulus mouse during Phase 3 on Day 2. The six sociability scores from Day 1 and from Day 2 were used, yielding a total of 12 correlations with Phase 3. The six correlations of Phase 2 on Day 1 with Phase 3 on Day 2 were analyzed separately from the six correlations of Phase 2 on Day 2 with Phase 3 on Day 2. For each analysis, two sets of group comparisons and any appropriate pairwise comparisons were performed as described above.

Experiment 2: Reliability of manual and automated methods of scoring sociability

Manual and automated methods of measuring sociability were compared using a cohort of mice that was not included in any previously published reports. These test mice were male C57BL/6J (n = 4) and BALB/cJ (n = 16) mice that were bred at the University of Pennsylvania. If litters exceeded six pups, then at 3 – 5 days of age they were culled to six pups so that as many males as possible were retained. Upon weaning at 23 – 25 days of age, two or three males were housed together in a cage, and only males were used for behavioral testing. All male test mice were tested at 31 days of age. Stimulus mice

were adult male A/J mice ordered from The Jackson Laboratory (Bar Harbor, ME) and housed 3 – 5 to a cage. All the A/J stimulus mice had been castrated before puberty. We used castrated males as stimulus mice in order to minimize the extent to which stimulus mice elicited sexual or aggressive motivations from test mice. The A/J mice were habituated to use as stimulus mice by repeated use as stimulus mice prior to the sociability testing described here. All mice were housed in a 12-hour light-dark cycle with the light cycle occurring from 7:00 a.m. – 7:00 p.m., and food and water were available ad libitum. Testing was carried out between 9:00 a.m. and 3:30 p.m. All animals were treated according to the National Institutes of Health Guide for the Care and Use of Laboratory Animals, and all procedures were approved by the University of Pennsylvania Institutional Animal Care and Use Committee.

Each mouse was tested for sociability on a single day. A test mouse was initially placed into the 3-chambered box and allowed to explore for 10 min (Phase 1). After Phase 1, an A/J stimulus mouse was placed into one of the two cylinders, and a novel object was simultaneously placed into the other cylinder. The novel object was a paper weight that was roughly the size of a mouse and it served as a novel, nonsocial stimulus. The test mouse was then able to sniff the cylinder with the stimulus mouse or the cylinder with the novel object for 10 min (Phase 2). Up to three mice were tested simultaneously in three separate 3-chambered boxes in very dim lighting. The intensity of light of the visible spectrum at the floor of the 3-chambered boxes measured 1 – 2 lux. The test sessions were recorded by video cameras overhead for subsequent manual and automated analysis.

The automated tracking of the mice by computer software depends on a high contrast between the mouse and the mat that serves as the "floor" of the 3-chambered box. Because the mat is light blue, the black C57BL/6J mice provide a high contrast and are easily tracked. The white BALB/cJ mice, however, provide a low contrast and are not easily tracked by the software. To resolve this difficulty, infrared light was shone upwards from below the mats/floors of the 3-chambered boxes, and video cameras that could detect infrared light were used. Thus, in the video the mat/floor appeared as a brightly lit area, while black and white mice appeared as silhouettes against it (Fig. 1B,C). This lighting arrangement provided adequate contrast for the software to track all the mice.

The precise arrangement of the apparatus was as described here: a clean blue mat was placed on top of a transparent sheet of Plexiglas that served as a table raised 68.5 cm above the room's floor (Fig. 1A). A 3-chambered box was placed on the light blue mat (Fisherbrand Absorbent Underpads, 20 x 24 in. (51 x 91cm), Fisher Scientific, Pittsburgh, PA) so that the mat served as a temporary "floor" of the 3-chambered box. Infrared lighting sources were placed beneath the Plexiglas table at the level of the room's floor. These infrared sources were either 10 – 12 Infrared (LED) Lighting Kits (Ramsey Electronics, LLC, Victor, NY) or a single infrared lighting panel (Clever Sys., Inc., Reston, VA). A video camera (Sony DCR-SR85 Handycam Camcorder, Sony Corporation, Tokyo, Japan) was mounted so that the lens of the camera was 167 cm above the mat/floor of the 3-chambered box. The camera's infrared-sensing "Nightshot" feature was enabled during testing. To prevent the stimulus mouse from moving the

social cylinder, 0.5-kg or 1-kg balance weights (Troemner Brass Gram Weights, Fisher Scientific, Pittsburgh, PA) were placed on top of both cylinders during Phase 2.

To assess the accuracy of manual scoring of sociability, videos were analyzed by three raters using The Observer XT Video-Pro 5.0 (Noldus Information Technology, Wageningen, The Netherlands) software. For each mouse, 5 min of video with the stimulus mouse present (Phase 1) or absent (Phase 2) was analyzed. Each rater produced two analyses of cylinder sniffing. First, each rater scored the video at the actual rate (1x) that the events occurred. For this "realtime" analysis, raters were not allowed to pause the video, re-watch any portion of it, or correct any perceived mistakes. This realtime analysis thus simulated a live scoring. For the second, "correctable" analysis, raters were allowed to score the videos at any rate that was comfortable to them. In practice, this rate was often 0.5x while the mice were sniffing and 1x or 2x during long time periods between sniffs. Raters were also allowed to pause and re-watch the video and to correct any perceived mistakes. Each rater scored all the mice for the realtime analysis before scoring any of the mice for the correctable analysis.

After producing the realtime and correctable analyses, Rater 1 (A.H.F.), the most experienced rater, scored the videos with a high degree of precision to determine the exact video frame when each sniff of a cylinder began and ended. A sniff was recorded only when the mouse's nose was very near and oriented towards a cylinder. This analysis produced a precise benchmark to which all other scoring methods were compared.

Automated analyses were performed by TopScan version 2.00 (Clever Sys., Inc., Reston, VA) software using its default settings (Fig. 1B). The TopScan software can track not only the position of the mouse but also the position of the mouse's head and

hind and the orientation of the mouse's nose. Therefore, the software can quantify the time that the mouse spends specifically sniffing directly at each cylinder. To prepare for the analysis, the software user must select or create a "background" image of the testing area that excludes the image of the mouse, calibrate the software's measurements of distances in the video so that metric units can be accurately used, draw an outline of the testing area and cylinders, and select the actions of the mouse that the software should record. While TopScan allows adjustment of multiple settings that can affect its measurement of cylinder sniffing, the primary setting that exerts a large effect is the size of the manually-drawn outline of the cylinder.

The outline of the cylinder can be approximated by using TopScan's circle drawing tools. Because the camera's view of a cylinder is not directly over that cylinder, the circular top of the cylinder may appear as an ellipse in the video. The circle drawn in TopScan is thus placed so that the area of the ellipse along its major axis that falls outside the circle approximately equals the area of the circle that falls outside the ellipse along its minor axis (Fig. 1C).

After the circle is placed, its radius can be lengthened beyond its original value. This expands the circle to varying sizes and allows sniffing of the cylinder to be recorded to varying degrees of accuracy. To determine what circle size allowed the most accurate measurements of cylinder sniffing, the radius of the circle was expanded from 0 mm – 18 mm beyond the original radius. The software was instructed to record the mouse's sniffing of the circle. To determine whether the mouse's proximity to the cylinder provided a reasonable approximation of the mouse's cylinder sniffing, the radius of the

circle was expanded from 10 mm – 70 mm beyond the original radius. The software was instructed to record the mouse's presence within the circle.

Video segments of 20 mice were analyzed. Each segment was 5 min in length and was sampled randomly from either Phase 1 or Phase 2. The amount of time that each mouse sniffed each cylinder was counted as a single data point.

An intraclass correlation coefficient (ICC) was used to compare the reliability of each manual or automated analysis to Rater 1's manual benchmark analysis. Specifically, the ICC(A,1) was used to account for the absolute agreement between each pair of analyses (McGraw and Wong, 1996). If and only if each mouse's sociability score of a given analysis equaled the corresponding score in the benchmark analysis, did the ICC(A,1) = 1. If the two scores for any mouse were not equal, then the ICC(A,1) < 1. Thus, unlike Pearson's *r* (or the ICC(C,1)), the ICC(A,1) is sensitive to any disagreement between scoring methods.

All analyses for both experiments were run on the statistical software R (R Development Core Team, 2007) with the software packages R Commander (Fox, 2007) and irr (Gamer, 2007). All *p* values less than 0.05 were considered statistically significant.

RESULTS

Experiment 1: General validity of six sociability scores

Following adjustment for strain, sex, and age, test-retest reliability of the six sociability scores was assessed by correlating scores obtained on Day 1 of testing with those obtained on Day 2 of testing (Fig. 2). Cylinder scores were more reliable than

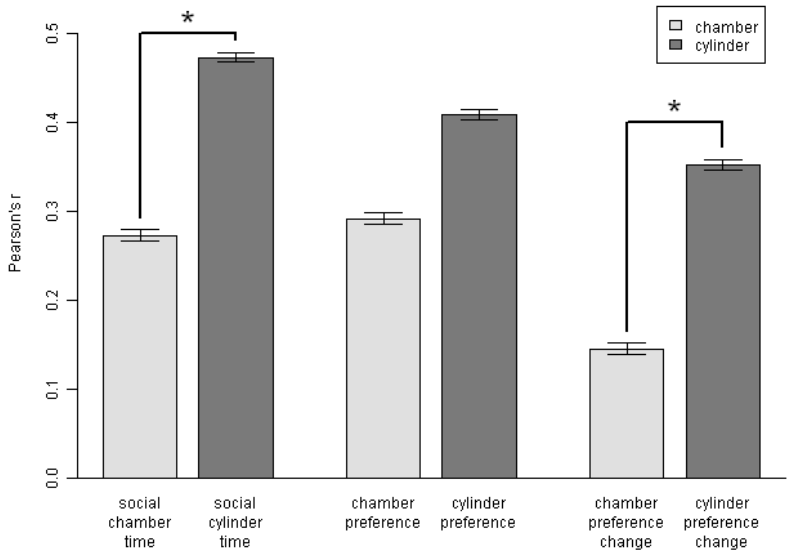

Fig. 2. Reliability of the six sociability scores. Each sociability score from Day 1 of testing was correlated with the corresponding score from Day 2. $N = 148$. Pearson's $r \pm$ SE. The three cylinder scores showed higher correlations than the three chamber scores, $P < 0.05$. $^*P < 0.05$ for pairwise comparisons, chamber vs. cylinder score.

chamber scores, $\chi^2(3, n = 148) = 9.55$, $p = 0.023$. Pairwise comparisons showed that social cylinder time was more reliable than social chamber time, $Z_2^*(145) = -2.39$, $p = 0.017$, and cylinder preference change was more reliable than chamber preference change, $Z_2^*(145) = -2.23$, $p = 0.025$. No difference between cylinder preference and chamber preference could be confirmed, $Z_2^*(145) = -1.35$, $p = 0.17$. There were no differences in reliability among social chamber/cylinder times vs. preference scores vs. preference change scores, $\chi^2(3, n = 148) = 3.52$, $p = 0.32$.

Although cylinder scores were more reliable than chamber scores overall, this effect might have applied to only some of the experimental groups tested, rather than

TABLE 1. Reliability (Pearson's *r*) of sociability scores for each experimental subgroup

Strain	Sex	Age	*n*	Social chamber time	Social cylinder time
C57BL/6J	Female	4 weeks	18	0.15	0.40
C57BL/6J	Female	9 weeks	17	0.46	0.59
C57BL/6J	Male	4 weeks	20	0.30	0.73
C57BL/6J	Male	9 weeks	18	0.38	0.60
BALB/cJ	Female	4 weeks	20	0.08	-0.05
BALB/cJ	Female	9 weeks	19	0.27	0.30
BALB/cJ	Male	4 weeks	18	0.63	0.82
BALB/cJ	Male	9 weeks	18	0.41	0.55

being broadly applicable to a variety of mice. To investigate this possibility, we inspected the correlations for each of the eight experimental groups, as defined by strain, sex, and age. This inspection focused on social cylinder time compared to social chamber time, because social cylinder time showed the highest reliability of all the scores (though nonsignificantly, compared to other cylinder scores). Social cylinder time was more reliable than social chamber time for seven out of eight experimental groups (Table 1). Mice of the remaining group, the 4-week-old BALB/cJ females, did not behave reliably from Day 1 to Day 2, as shown by their near-zero correlations for both social chamber and cylinder times. Social cylinder time was only marginally more reliable than social chamber time for 9-week-old BALB/cJ females. Overall, social cylinder time was more reliable than social chamber time for all groups, with the exception of BALB/cJ females, for which the two scores performed about equally.

To address the question of ecological validity, the six sociability scores were next assessed for their correlations to the Phase 3 (Free Social Interaction) score. Following adjustment for strain, sex, and age, the six sociability scores obtained during Phase 2 (Social Choice) on Day 1 were correlated with the amount of time the test mice spent sniffing the stimulus mice during Phase 3 on Day 2 (Fig. 3A). Cylinder scores did not

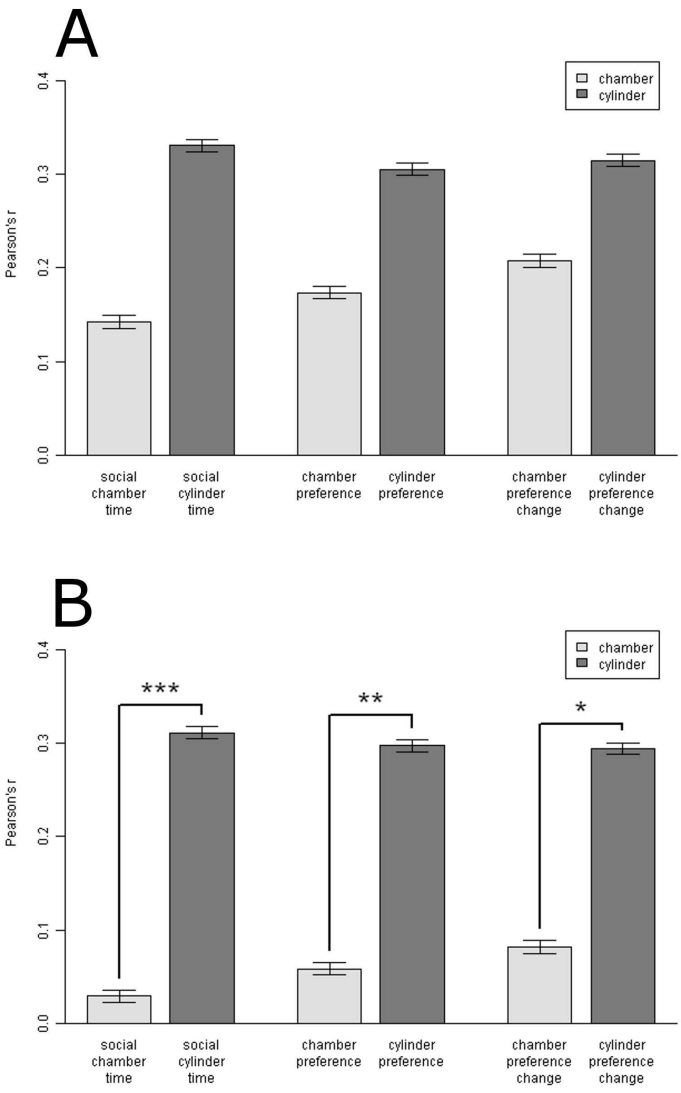

Fig. 3. Ecological validity of the six sociability scores. **(A)** Each sociability score from Day 1 of testing was correlated with the Phase 3 (Free Social Interaction) score from Day 2. There were no significant differences among the scores. N = 146. Pearson's r ± SE. **(B)** Each sociability score from Day 2 of testing was correlated with the Phase 3 (Free Social Interaction) score from Day 2. N = 146. Pearson's r ± SE. The three cylinder scores showed higher correlations than the three chamber scores, $P < 0.01$. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, for pairwise comparisons, chamber vs. cylinder score.

show significantly higher correlations with the Phase 3 score than did chamber scores, $\chi^2(3, n = 146) = 5.55$, $p = 0.135$. There were also no differences in ecological validity among social chamber/cylinder times vs. preference scores vs. preference change scores, $\chi^2(3, n = 146) = 0.39$, $p = 0.94$. An inspection of the social cylinder time and social chamber time by experimental groups indicated that cylinder scores showed higher correlations than chamber scores for seven out of eight experimental groups (Table 2).

TABLE 2. Ecological validity (Pearson's *r*) of sociability scores for each experimental subgroup

Strain	Sex	Age	*n*	Social chamber time, Day 1	Social cylinder time, Day 1	Social chamber time, Day 2	Social cylinder time, Day 2
C57BL/6J	Female	4 weeks	18	0.17	0.47	0.02	0.39
C57BL/6J	Female	9 weeks	17	-0.02	-0.21	-0.04	0.03
C57BL/6J	Male	4 weeks	19	0.38	0.47	-0.05	0.49
C57BL/6J	Male	9 weeks	18	-0.07	0.10	-0.02	0.38
BALB/cJ	Female	4 weeks	20	-0.09	0.06	-0.08	0.26
BALB/cJ	Female	9 weeks	19	0.02	0.28	0.01	0.16
BALB/cJ	Male	4 weeks	18	0.28	0.55	0.42	0.69
BALB/cJ	Male	9 weeks	17	0.13	0.45	-0.08	0.38

For the remaining group, the 9-week-old C57BL/6J females, the social cylinder time showed a modest negative correlation.

Following adjustment for strain, sex, and age, the six sociability scores obtained during Phase 2 (Social Choice) on Day 2 were then correlated with the amount of time the test mice spent sniffing the stimulus mice during Phase 3 on Day 2 (Fig. 3B). Cylinder scores showed higher correlations with the Phase 3 score than chamber scores, $\chi^2(3, n = 146) = 14.86$, $p = 0.0019$. Cylinder scores outperformed chamber scores for all three pairs of comparisons: social cylinder time vs. social chamber time, $Z_2^*(143) = -3.37$, $p = 0.0007$; cylinder preference vs. chamber preference, $Z_2^*(143) = -2.87$, $p = 0.004$; and cylinder preference change vs. chamber preference change, $Z_2^*(143) = -2.49$, $p = 0.013$. No differences were present among social chamber/cylinder times vs. preference scores vs. preference change scores, $\chi^2(3, n = 146) = 0.24$, $p = 0.97$. The Phase 3 correlations with social cylinder time exceeded the Phase 3 correlations with social chamber time for all eight experimental groups (Table 2), though the 9-week-old

C57BL/6J females showed only a marginal difference and virtually no correlation overall.

Experiment 2: Reliability of manual and automated methods of scoring sociability

Manual and automated methods of scoring cylinder sniffing were compared to a precise, manual benchmark analysis to assess their scoring accuracies (Fig. 4). In

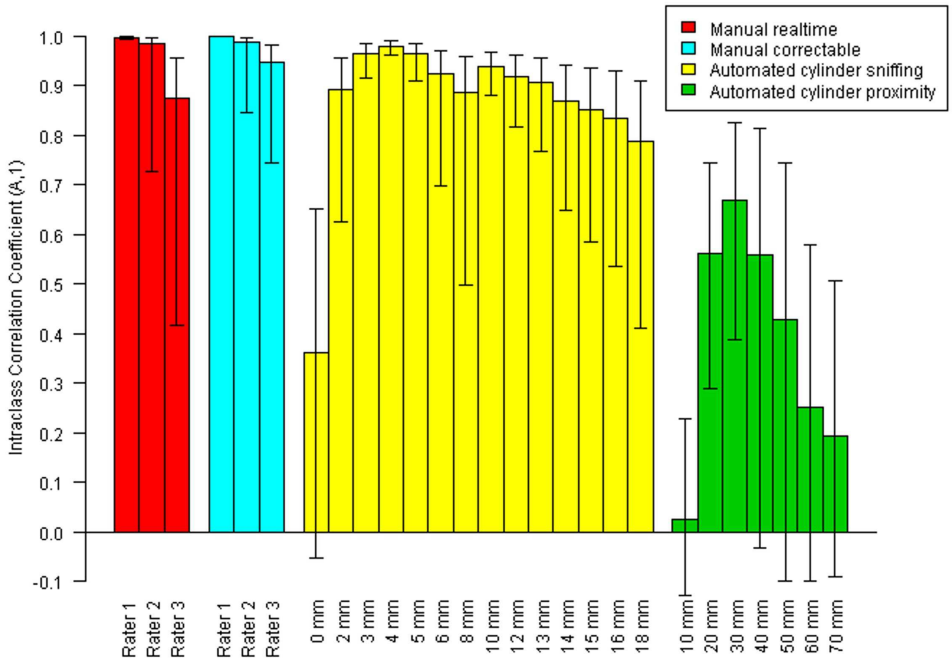

Fig. 4. Agreement of manual and automated methods of measuring social cylinder time with the benchmark analysis. $N = 40$. ICC(A,1) ± 95% confidence interval. For automated cylinder sniffing, 0 – 18 mm denotes the length by which the software operator extended the cylinder's radius to accommodate the mouse's sniffing of the cylinder. For automated cylinder proximity, 10 – 70 mm denotes the length by which the software operator extended the cylinder's radius to accommodate the mouse's location within the resulting circular area surrounding the cylinder (see Materials and Methods).

addition to the ICC, the lower 95% confidence interval (LCI) values of the ICC are examined to account for a very conservative estimate (a likely underestimate) of the correlation among the analyses. The realtime analyses by the three human raters produced ICC values ranging from 0.87 (Rater 3) to 0.997 (Rater 1), and LCI values ranging from 0.42 (Rater 3) to 0.994 (Rater 1). The correctable analyses showed higher ICC and LCI values, which indicated scoring more consistent with the benchmark analysis than the realtime analyses. Specifically, the correctable ICC values ranged from 0.95 (Rater 3) to 0.999 (Rater 1), and the LCI values ranged from 0.75 (Rater 3) to 0.998 (Rater 1).

The automated scoring of cylinder sniffing by the TopScan software yielded its highest ICC and LCI values when the radius of the circle that demarcated the cylinder was extended by 4 mm beyond the original radius (Fig. 4). Under this condition, the software produced an ICC of 0.980 and a LCI of 0.96. These values were only slightly below those of Rater 1 for both the realtime and correctable analyses. The software's ICC exceeded that of Rater 3 for both realtime and correctable analyses, and was slightly exceeded by those of Rater 2 for both analyses. The software's LCI value was higher than the LCI values for both Raters 2 and 3 for both analyses. Thus, the results from the TopScan software correlated with the benchmark analysis as well as or better than the human raters did.

The performance of the human raters and the TopScan software without reference to the benchmark analysis are also reported (Table 3). The three human raters achieved inter-rater reliabilities, as measured by the ICC, ranging from 0.87 to 0.99 for the realtime analyses, and ranging from 0.95 to 0.99 for the correctable analyses. The correlations of

TABLE 3. Inter-rater reliabilities (ICC(A,1)) among three human raters and TopScan software

		Realtime analysis	Correctable analysis
Rater 1	Rater 2	0.99	0.99
Rater 1	Rater 3	0.87	0.94
Rater 2	Rater 3	0.92	0.98
Rater 1	TopScan (4mm)	0.98	0.98
Rater 2	TopScan (4mm)	0.97	0.98
Rater 3	TopScan (4mm)	0.89	0.95

the TopScan software's results with the human raters ranged from 0.89 to 0.98 for the realtime analyses and from 0.95 to 0.98 for the correctable analyses. These results again show that the TopScan software performed comparably to the human raters in measuring sociability.

The TopScan software was also used to assess whether a mouse's proximity to the social cylinder -- without scoring the mouse's sniffing of the social cylinder – provided an adequate estimate of the mouse's sociability, as measured by the benchmark analysis of social cylinder time. The software attained its highest correlation to the benchmark analysis when it scored the mice while the mice were within 30 mm of the social cylinder (Fig. 4). This ICC value (0.67) was substantially below the lowest ICC of a human rater (0.87) and below the software's ICC when it measured sniffing of the social cylinder (0.98). Thus, measuring the test mouse's proximity to the social cylinder is an inadequate approximation of the mouse's sniffing of the social cylinder.

Scatterplots of several of the manual and automated analyses as compared to the benchmark analysis are reported (Fig. 5).

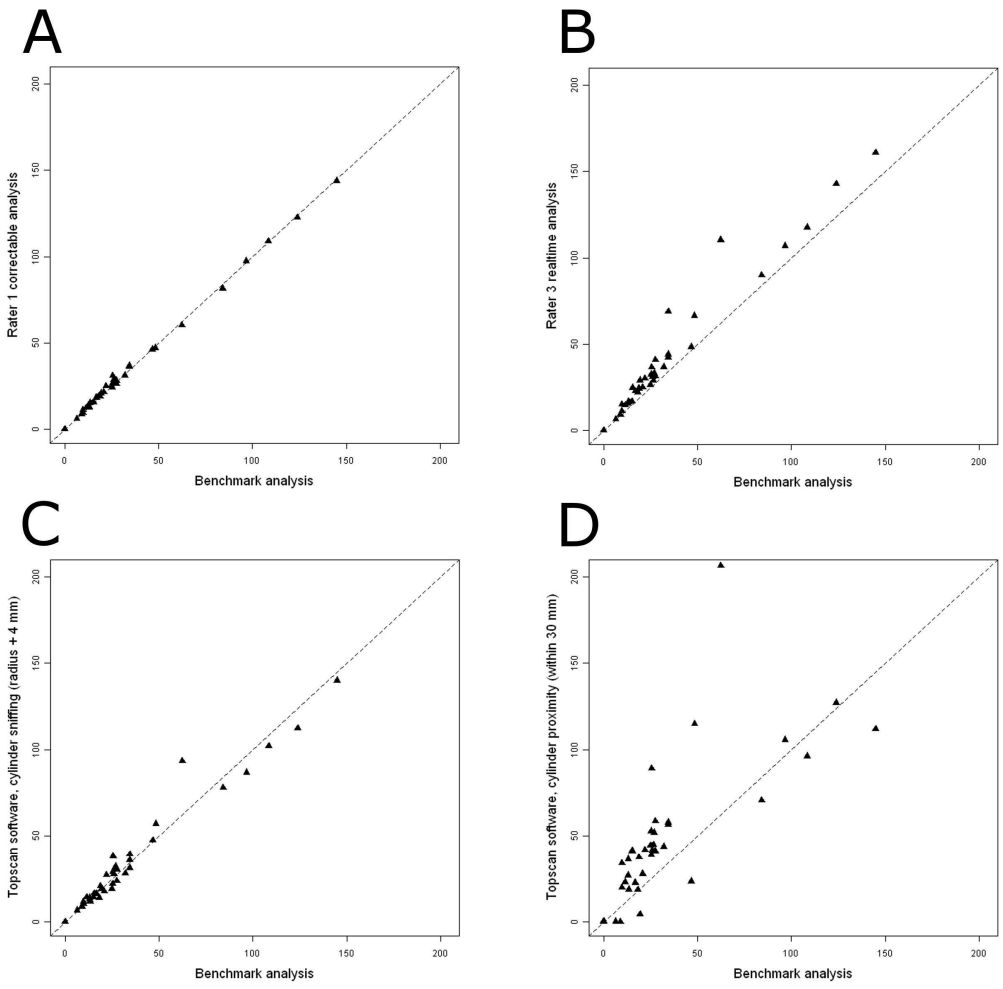

Fig. 5. Scatterplots of manual and automated methods of measuring social and nonsocial cylinder time with the benchmark analysis. $N = 40$. The dotted line indicates $y = x$, which represents perfect agreement between the two analyses. **(A)** Rater 1 correctable analysis (the manual analysis with the highest correlation to the benchmark analysis) vs. the benchmark analysis. **(B)** Rater 3 realtime analysis (the manual analysis with the lowest correlation to the benchmark analysis) vs. the benchmark analysis. **(C)** TopScan software, cylinder sniffing with the radius of the cylinder extended by 4 mm vs. the benchmark analysis. **(D)** TopScan software, cylinder proximity within 30 mm vs. the benchmark analysis of social cylinder time.

DISCUSSION

Contrary to our hypothesis, cylinder scores – not the chamber preference change score – showed the highest test-retest reliability. Additionally, cylinder scores obtained on Day 2 were more ecologically valid than Day 2 chamber scores. Although the ecological validity of Day 1 cylinder scores appeared to be higher than the ecological validity for Day 1 chamber scores, this difference did not reach statistical significance. Yet even in this case, all the cylinder scores attained higher correlations than did any of the chamber scores. Thus overall, cylinder scores achieved higher test-retest reliability and ecological validity than did chamber scores.

Our sample size did not provide sufficient statistical power to detect the relatively small differences among social chamber/cylinder times, preference scores, and preference change scores. Yet notably, the social cylinder time showed the highest test-retest reliability of all the scores. Furthermore, the social cylinder time – and not the hypothesized cylinder preference change score – showed the highest ecological validity, though only slightly higher than the other cylinder scores. Thus, the three cylinder scores were more generally valid than the three chamber scores, and the social cylinder time may be the most generally valid of the three cylinder scores; however additional study with larger sample sizes would be required to determine whether the latter point is true.

We had hypothesized that preference change scores would be the most generally valid scores because they theoretically control for a mouse's tendency to explore a nonsocial stimulus (as do preference scores) and for its individual preference for a chamber or cylinder. It was therefore surprising that the cylinder preference and cylinder preference change scores were not more generally valid than social cylinder time in any

case. The information about nonsocial stimulus investigation and prior chamber/cylinder preference may be weakly or not related to sociability, so by including it, these scores may introduce nearly random "noise" to the "signal" of social cylinder time. The lower test-retest reliability of these complex scores, compared to social cylinder time, may support this notion, but a similar pattern does not appear for ecological validity. Regardless, the experimenter should be cautioned against assuming that a more complex score necessarily improves the general validity of a behavioral analysis, and it may sometimes decrease general validity.

The superior general validity of cylinder scores over chamber scores suggests that sociability should be measured primarily by including only the active behaviors that are most directly related to social investigation. The predominant active social behavior is sniffing the cylinder. When its nose is in contact with the cylinder, the test mouse can likely perceive both volatile and nonvolatile odorants from the stimulus mouse (Brennan and Kendrick, 2006, Luo et al., 2003, Sanchez-Andrade and Kendrick, 2009). Other active social behaviors include scratching, gnawing, climbing on, and rearing against the cylinder. Chamber scores may include other, passive social behaviors that can occur with some distance between the mice, such as when the test mouse chooses to be near another mouse, watches that mouse, or smells volatile odorants that have diffused some distance from that mouse. But these scores also include behaviors that are not clearly social, such as sniffing the chamber walls, walking through the chamber but not towards the social cylinder, and remaining still next to the chamber wall. Likewise, low locomotor activity may substantially affect a chamber score. Excluding these behaviors by accounting for

only active behaviors directed toward the social cylinder yields a more generally valid measurement of sociability.

Because this study analyzed a heterogeneous group of mice, some conclusions may not apply evenly across all subgroups. With no more than 20 mice in each subgroup, statistical power was not sufficient to test robustly for correlational differences among the subgroups, and the estimates of the magnitude of the correlations for each subgroup are imprecise. However, some general patterns are noteworthy.

Social cylinder time showed higher test-retest reliability and ecological validity than social chamber time for nearly all subgroups. This was not true for the test-retest reliability of the 4-week-old BALB/cJ females, which behaved inconsistently across test sessions. Thus, for no experimental group did social chamber time indicate greater reliability than social cylinder time. For ecological validity, the 9-week-old C57BL/6J females were an exception, where the social chamber time correlation exceeded that of social cylinder time. However, the chamber time correlation was near zero, while the cylinder time correlation was surprisingly negative, though not of high magnitude. Thus even in this case, the social cylinder time may show a relationship that the chamber time does not show. In sum, there is no evidence that the chamber scores are more generally valid than the cylinder scores for any subgroup.

The correlations presented here are based on the behaviors of individual mice. While assessing reliability on an individual level is a common approach (Andreatini and Bacellar, 2000, Drugan et al., 1989, Henderson, 2005, Hilakivi and Lister, 1990, Lister, 1987, Teixeira-Silva et al., 2009), studies of anxiety-related behaviors suggest that examining behaviors on a group level can yield different results (Ramos, 2008). In some

cases, group-level analyses were able to detect behavioral correlations that were not present at an individual level. Thus, the possibility remains that a group-level analysis could detect higher general validity of chamber scores than has been found here. In developing the Social Choice Test, Moy et al. (2004) presented evidence that chamber scores are generally reliable at a group level: adult C57BL/6J and DBA/2 mice showed largely similar chamber scores between a test and re-test 11 – 12 days later. However, cylinder scores were not reported in this experiment, so it is unclear whether the chamber scores' reliability equals that of the cylinder scores at a group level. Given the large difference in general validity between chamber and cylinder scores found here, it is unlikely that a group-level comparison of chamber and cylinder scores would undermine our recommendation to primarily use cylinder scores to evaluate sociability.

This study was limited by the use of archival data (Sankoorikal et al., 2006) that were not originally designed to answer questions on the general validity of the sociability scores. One limitation was potential test order effects: the interactions of the mice during Phase 2 (Social Choice) might have affected their subsequent interactions in Phase 3 (Free Social Interaction). No mice were tested in Phase 3 before Phase 2 to identify any test order effects. Furthermore, a test mouse was exposed to the same stimulus mouse for Phase 2 on Day 2 and for Phase 3 (also on Day 2), which may have attenuated their interaction during Phase 3 due to a habituation effect. However, any attenuation of social interaction that affected the mice fairly uniformly would not have greatly affected the correlations between Phase 2 and Phase 3, which were based on Pearson's *r*. Notably, attenuation of social interaction (habituation) seems even less likely between Phase 2 on Day 1 and Phase 2 on Day 2, because each test mouse was tested with

different stimulus mice on Day 1 and Day 2 and because of the day-long interval between tests. Additionally, any test order effects might have been minimal: the effects of prior testing experience depend on the specific paradigms used and do not necessarily affect results substantially (Henderson, 2005, McIlwain et al., 2001).

The Social Choice Test is a highly controlled assay for social affiliation, and this high level of control entails curtailing some naturalistic aspects of social interactions between mice. Confining the stimulus mouse to a cylinder in Phase 2 allows one to isolate, to some degree, the sociability of the test mouse. But it also alters the quality or nature of the social interaction, because the confinement of the stimulus mouse limits its ability to initiate, maintain, and terminate a social interaction and to respond to social cues from the test mouse. Social behaviors of the test mouse may also be affected by being in a novel environment, which can induce exploratory and anxiety-related behaviors, and by the inability to fully contact the stimulus mouse due to the presence of a partial barrier between them (cylinder wall with holes in it). However, it is worth noting that this controlled social interaction shows some similarity to a more naturalistic interaction, as shown by the positive correlations between the social measures of Phase 2 (Social Choice) and Phase 3 (Free Social Interaction) (Fig. 3). Moreover, we have chosen to regularly include a Free Social Interaction phase in the Social Choice Test in all of our studies (Brodkin et al., 2004, Sankoorikal et al., 2006, Fairless et al., 2008), in order to include both a more controlled and a more naturalistic way of observing social interactions in the context of the Social Choice Test.

Phase 3 (Free Social Interaction) is more naturalistic than the Phase 2 (Social Choice), during which the stimulus mouse is confined to a cylinder, because both mice

can move freely in Phase 3. Nevertheless, Phase 3 still differs substantially from a social situation between feral mice in their natural environment. Among many other artificial factors, the mice in the Free Social Interaction (Phase 3) are laboratory-bred; are restricted to a novel, artificial environment; and interact in the presence of a human. Strategies that reduce or eliminate such factors to attain more naturalism – such as observing mice in home cage environments or semi-natural burrow habitats – can be related to the Social Choice Test to further investigate its ecological validity. Importantly, mice of the inbred strain BTBR *T+ tf*/J show lower social behaviors than C57BL/6J mice in both the Social Choice Test and semi-natural burrow habitats (McFarlane et al., 2008, Pobbe et al., 2010). Unlike the present study, these results are based on group-level analyses, but they do support the notion that results from the Social Choice Test can be relevant to more naturalistic social situations.

Since the archival data were collected, the procedure for the Social Choice Test has been altered. In the earlier experiment (Sankoorikal et al., 2006, this study, Experiment 1) the stimulus mouse was placed into one cylinder while the other cylinder remained empty at the start of Phase 2. When the stimulus mouse was introduced, it was both a social stimulus and a novel stimulus. To control for novelty as a possible confound, subsequent experiments have included a novel object that is introduced into the other (nonsocial) cylinder at the same time that the stimulus mouse is introduced into the social cylinder at the start of Phase 2 (Fairless et al., 2008; this study, Experiment 2). Given this change, it is possible that the results concerning test-retest reliability and ecological validity from Experiment 1 would not apply well to subsequent experiments. We consider this unlikely because the procedure change (presence of the novel object in

the nonsocial cylinder) has not substantially changed behaviors of test mice: the test mice generally sniff the nonsocial cylinder little compared with the social cylinder using either procedure, and experimental results in C57BL/6J and BALB/cJ mice have been very similar before and after the procedural change (e.g., juvenile BALB/cJ mice consistently have shown lower sociability than juvenile C57BL/6J mice, both before and after the procedure change; Sankoorikal et al., 2006, Fairless et al., 2008).

Tools that can automate the measurement of chamber scores are well established (Nadler et al., 2004, Page et al., 2009) and widespread, and this may account for the prevalence of using only chamber scores to assess sociability in the Social Choice Test. Given the cylinder scores' superior general validity indicated in our study, exclusive use of chamber scores may produce a higher rate of undetected false positives and false negatives in the Social Choice Test. To facilitate the use of cylinder scores, we have validated the software TopScan for automated measurement of cylinder sniffing in the Social Choice Test. At the settings that we specified, TopScan performs as well as or better than human raters at this task, as we had hypothesized.

Some have suggested that a mouse's proximity to a cylinder provides an adequate measure of sociability (Page et al., 2009). Contrary to this hypothesis, our data show that this approach provides a measurement of sniffing inferior to that of directly measuring sniffing of the cylinder, either by manual or automated methods. We have observed that test mice often walk beside or along the cylinder wall, but orient their heads towards the cylinder for only brief, intermittent periods to sniff. This behavior may account for much of the discrepancy between the "cylinder proximity" measurements and our recommended "cylinder sniffing" approach. In summary, use of the cylinder proximity

approach may risk a higher rate of false positives and false negatives in assessing sociability in the Social Choice Test; our results support the use of direct measurements of cylinder sniffing.

The higher general validity of cylinder scores compared to chamber scores suggests that active investigation of a conspecific is the predominant component of sociability in the Social Choice Test. Sociability, the tendency to approach and affiliate with an unfamiliar conspecific, is a relatively simple social behavior, but it is important in many species as a prelude to more complex behaviors, such as the formation of social bonds. Research into the biological factors that influence sociability in mouse models of ASD may eventually yield insight into the social impairments of ASD, and optimal measurement of sociability is essential to obtaining clear results in this endeavor.

Acknowledgements

This work was supported by National Institutes of Health Grants R01MH080718 (E.S.B.) and 5-T32-MH017168-26 (training grant supporting A.H.F.) and a Burroughs Wellcome Fund Career Award in the Biomedical Sciences (E.S.B.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.