Journal of Veterinary Behavior 31 (2019) 43–58

Contents lists available at ScienceDirect

Journal of Veterinary Behavior

journal homepage: www.journalvetbehavior.com

Canine Research

What is the evidence for reliability and validity of behavior evaluations for shelter dogs? A prequel to "No better than flipping a coin"

Gary J. Patronek [a,*], Janis Bradley [b], Elizabeth Arps [b]

[a] Center for Animals and Public Policy, Cummings School of Veterinary Medicine at Tufts University, North Grafton, Massachusetts
[b] The National Canine Research Council, Amenia, New York

ARTICLE INFO

Article history:
Received 2 January 2019
Accepted 1 March 2019
Available online 8 March 2019

Keywords:
animal shelter
dog behavior evaluation
temperament test
validation
dog personality
psychometric
aggression

ABSTRACT

Conversations with stakeholders, as well as remarks in the literature, suggest that there may be confusion about what can be concluded when a canine behavior evaluation has been described as being "validated," "reliable," or "predictive." To assess the evidence, we searched PubMed and ScienceDirect using the terms "canine," "behavior evaluation," "temperament test," and "shelter" to identify articles that assessed the validity or reliability of evaluations based on battery of tests used or intended for screening shelter dogs for behavior labeled aggressive and/or for adoption suitability. Despite 25+ years of publications, including solid studies performed under good to ideal conditions by skilled investigators, findings indicate there is no evidence that any canine behavior evaluation or individual subtest has come close to meeting accepted standards justifying claims that it is validated for routine use in shelters. Furthermore, the mean reported false-positive error rate in study populations was 35.1%, whereas in more typical shelter populations, it was estimated at 63.8%. We propose that the discrepancy between the actual state of the science and what people assume has been accomplished is primarily due to the following: [1] confusion from mixing colloquial with scientific uses of words such as "validated," "predictive," "reliable," and "agreement"; [2] the limitations of correlation and regression as statistical methods for demonstrating agreement or predictive ability; [3] failure to account for the difference between predictive validity of an instrument in populations of dogs in a research exercise versus predictive ability and error rate for individual dogs in real-world settings; [4] conflating statistical significance with clinical significance; and, as a result of 1-4 aforementioned, [5] conferring overall validation status, despite the results of studies being much more circumscribed. Given their published error rates, one explanation may be that behavior evaluations lack basic face validity and/or a clear focus as to what is being measured and its relevance to postadoption outcomes. This argues against use of any behavior evaluation to make important decisions for shelter dogs, especially if the behavior(s) of concern were only observed during provocative testing. These findings indicate an opportunity to acknowledge what has been learned and bring together all stakeholders to consider the real needs of shelter dogs and what the future might look like.

Introduction

Canine behavior evaluations conducted in shelters for the purpose of making rehoming decisions typically comprise a battery of tests (often referred to as a series of subtests) involving different provocations on the part of the tester to attempt to elicit behavioral responses from the dogs. In some cases, the behavior of the dog under normal conditions or without provocation is also scored. The implication is that the test designers believe that the responses expressed during the tests reflect some global, fixed behavioral tendency of the dogs.

In a previous article "No better than flipping a coin: Reconsidering canine behavior evaluations in animal shelters", we (G.J.P. and J.B.) highlighted a fundamental, species-neutral principle of

* Address for reprint requests and correspondence: Gary J. Patronek, Cummings School of Veterinary Medicine at Tufts University, Center for Animals and Public Policy, North Grafton, MA 01536. Tel:(508) 839-7991; Fax:(508) 839-333.
E-mail address: gary.patronek@tufts.edu (G.J. Patronek).

https://doi.org/10.1016/j.jveb.2019.03.001

diagnostic testing that had not received sufficient attention in the canine behavior literature, namely, that using even a validated test in a population with a low prevalence of the behavior of interest would inevitably be associated with an unacceptably high rate of false-positive results (Patronek and Bradley, 2016). Having a low prevalence of behaviors that adopters would find too problematic to live with or that the test designers believed classify the dog as dangerous would apply to most, if not all, populations of shelter dogs that undergo a behavior evaluation. For shelter dogs, this would mean that many could be inaccurately labeled as having behavior problems and/or considered as failing the test, with potentially dire consequences including euthanasia when behavior evaluations are the main determinant of their fate.

Since that study was published, we have received questions along the lines of "I thought it had been shown that a particular behavior evaluation has been validated (or could reliably predict future behavior)?", and we have also read similar assertions in published articles. This was both unexpected and concerning because in our view, scientifically demonstrating overall validity (i.e., having acceptable reliability, validity, and predictive ability conferring suitability for routine use in shelter settings) has not been achieved for any behavior evaluation or subtest published to date.

We propose that much of the disparity between the actual state of the science and what people may assume has been accomplished is primarily due to the following: [1] confusion from mixing colloquial with scientific uses of words such as "validated," "predictive," "reliable," and "agreement"; [2] the limitations of correlation and regression as statistical methods for demonstrating agreement or predictive ability; [3] uncertainty about the difference between predictive validity of an instrument in populations of dogs in a research exercise versus predictive ability and error rate for individual dogs in real-world settings; [4] conflating statistical significance with clinical significance; and, as a result of 1-4 aforementioned, [5] conferring overall validation status despite the results of studies being considerably more circumscribed. The goals of this article are to clarify these five issues with respect to shelter behavior evaluations to assess what has actually been demonstrated by reviewing published studies and to offer some insights on what would be required to make a claim of overall reliability, validity, and utility in the future for any canine behavior evaluation.

Literature search

We searched PubMed and ScienceDirect using the terms "canine," "behavior evaluation," "temperament test," and "shelter" to identify publications that reported results from assessing the validity or reliability of behavior evaluations used for screening dogs for suitability for adoption and/or the presence of behaviors categorized as aggressive by the test designers. We included articles where evaluations were tested on owned dogs if the evaluation had also been used in, or was potentially intended for, shelter dogs. We also reviewed the reference lists of the retrieved articles. We did not include publications whose focus was owner-completed scales used to assess dog behavior, evaluations used to identify suitability for training guide or service dogs, or results from testing purebred dogs for inclusion or exclusion from breeding programs. A final inclusion criterion was that articles had to report some measure of reliability, validity, or predictive ability. We identified a total of 17 articles published from studies conducted in the United States, Italy, Australia, the Netherlands, Sweden, Switzerland, Hungary, and the United Kingdom whose data were consistent enough to summarize. Those results, and our assessment of the degree of support from the findings, are summarized in Figure 1. Specific reasons for our assessments are further explained in Table 1. A glossary of terms is also provided.

Issue 1: Colloquial versus scientific terminology

When discussing measurement of a subjective outcome using a diagnostic instrument, terms such as "validated," "predictive," "reliable," and "agreement" have specific scientific meanings (Streiner et al., 2015; Taylor and Mills, 2006). However, it can be problematic when colloquial meanings are used when considering the results of a scientific article. Slipping into colloquial usage may imply what we defined in the Introduction as "overall validity," when in fact such usage is likely to overstate what has actually been shown. The latter could occur, for example, if only some component of reliability or validity had been explored to some degree, perhaps only reaching the statistical threshold of being correlated, as we explain later (under Issue 4: Statistical versus clinical significance), "somewhat more often than not."

This situation is hardly unique to the literature about canine behavior evaluations. Mokkink et al. (2010) describe the COSMIN study—a Delphi process by which an international group of scientists developed a consensus on taxonomy, terminology, and definitions of measurement properties for health-related, subjective, patient-reported outcomes such as pain or quality of life. They noted that "many different terms for the same measurement property "reliability" are used interchangeably, such as reproducibility, reliability, repeatability, agreement, precision, variability, consistency, and stability", and emphasized how different uses of terminology can lead to confusion about which measurement properties are being assessed (Mokkink et al., 2010, p. 737). Streiner et al. (2015, p. 159) have also noted this problem. Similarly, the US Food and Drug Administration has issued a guidance document for the pharmaceutical industry when seeking a labeling claim for a drug to improve a patient-reported outcome (Department of Health and Human Services, 2009). That document explores many of the same terminology conundrums as the COSMIN study.

The issue of colloquial versus scientific meaning also bears on the question about what shelters think it means to respond affirmatively to questions about whether a validated behavior evaluation would be useful to them, as was performed by Haverbeke et al. (2015). We suspect that the fact that even a validated evaluation would be associated with a particular error rate (i.e., false-positive and false-negative findings) is not fully appreciated when shelters respond in the affirmative to this type of question. The error rate is an essential piece of information for shelters to have before using a test.

Key concepts

In this section, we will only briefly touch on material describing the components of reliability and validity in temperament tests for adult dogs covered in a previous article in this journal (Taylor and Mills, 2006, Appendix A and B for definitions). Instead, we will focus on material that has not been previously addressed in depth and/or new interpretations of previous concepts, particularly the practical meaning and implications of reliability and validity. To illustrate the scope of research needed to satisfy the scientific standard of test validation, we have chosen to focus on attempts to measure reliability and construct validity, each a necessary but not sufficient step for demonstration of overall validity. For a readable, well-established textbook on basic principles that can be purchased online as an e-book, we also recommend "Health Measurement Scales: a Practical Guide to Their Development and Use" (5th edition) (Streiner et al., 2015). We will quote liberally from that text.

Reliability

In our experience, a conversation about the overall utility of a behavior evaluation (implying suitability for use in real-world situations such as animal shelters) often begins with issues

surrounding individual types of validity (i.e., is the instrument measuring what is intended to a sufficient extent). However, as Streiner et al. (2015, p. 9) have indicated, the most important first step is to demonstrate that an instrument is reliable—in other words, that it is at least measuring something in a reproducible fashion (i.e., reliability). Another reason that reliability should be forefront in the discussion about the overall validity or utility of a canine behavioral evaluation is that "reliability places an upper limit on validity" (Streiner et al., 2015, p. 228). This means that if reliability is less than 100%, the extent of any claims about validity must be reduced accordingly.

Inter-rater reliability (i.e., different staff reaching comparable conclusions on the same dog) is the most important type of reliability to establish in the context of shelter behavior evaluations. As Streiner et al. (2015, p. 171) have noted, if inter-rater reliability is acceptable, then by default, intra-rater reliability will be as good or better. Therefore, we did not summarize any results for intra-rater reliability. Although some studies we reviewed reported on inter-rater reliability (Diesel et al., 2008a; Valsecchi et al., 2011), only three studies (Klausz et al., 2014; Mornement et al., 2014; van der Borg et al., 2010) demonstrated acceptable inter-rater reliability (i.e., having mostly strong [$r > 0.70$] and statistically significant [$P < 0.05$] correlations or high kappa values [$\kappa \geq 0.81$]) (Figure 1; Table 1).

A type of reliability that does not bear nearly as much on human clinical situations, but that factors in greatly for canine behavior assessments, might be called "inter-shelter" reliability. Given the heterogeneity of the animal shelter environment, any type of reliability demonstrated at one or even several sites under carefully controlled research conditions cannot be assumed to be the same during actual field use in the extremely diverse settings of animal shelters (Taylor and Mills, 2006). Why is this so important? As an example, in human health care settings, it is even regarded as important to validate something that may seem as trivial as different modes of administration of a scale (e.g., self-administration by paper and pencil or computer entry versus interview by a third party) and not assume that one mode of administration will produce equivalent results to another. For dogs, this would imply that the reliability of a test determined from videos (Diesel et al., 2008a; Klausz et al., 2014; Kroll et al., 2004; Planta and De Meester, 2007), while a useful piece of supportive evidence, cannot be extrapolated to field use (where the typical administration involves observing the dog's behavior from a full vantage point in real time, possibly with a notetaker present in the room). Further studies would have to be conducted under actual shelter field use conditions. No study has demonstrated inter-shelter reliability for any canine behavior evaluation.

With respect to the technical measurement of reliability, the methodological approaches that tend to increase reliability are not practical in a shelter setting. For example, Klausz et al. (2014) rightly note that for humane as well as pragmatic reasons, it would be desirable to reduce the number of subtests included. Unfortunately, that prudent recommendation runs counter to the statistical property whereby brevity (fewer subtests) carries the cost of lower reliability of the evaluation as a whole (Streiner et al., 2015, p. 10). Similarly, the sensible practice of excluding dogs from testing because of known biting history or dangerous behavior, or restricting the study to dogs deemed suitable for rehoming (Hennessy et al., 2001; Mornement et al., 2014; van der Borg et al., 1991), will have the effect of increasing the behavioral homogeneity of the dogs actually tested. From a statistical perspective, the effect of having dogs with less behavioral variety is to make it more difficult to achieve a particular level of reliability. Conversely, if dogs with a history of biting or intensely threatening behavior were included in a study that would have the effect of falsely elevating reliability with respect to the usual population of dogs available for adoption, if the adoption pool did not typically include dogs with such a history (Streiner et al., 2015, pp. 164, 245).

Repeatability (i.e., test-retest reliability) is also often mentioned as an important type of reliability to establish for an evaluation, but both the rationale for conducting such an exercise as well as the logistical issues unique to dogs and shelters merit some additional discussion. We are not the first to note the numerous conceptual problems with attempting to demonstrate elicitation of similar canine behaviors on retesting, including the inability to scientifically determine how soon after intake to perform and when to repeat an evaluation, as well as the uncertain effect of dogs' adaptation (or failure to adapt) to the shelter environment and staff. Although a standard power calculation could be used to determine the number of dogs to be tested to detect a particular difference in scores that would be statistically significant, there is no consensus as to what level of correlation in before and after scores would constitute agreement, and conversely, it is unknown how different the scores on the two tests would need to be in order to be considered clinically significant, therefore indicating non-repeatability. If the results of an evaluation are shown to be somewhat repeatable in dogs that have been in the shelter for a moderate or extended period (e.g., months) of time (Valsecchi et al., 2011) that is not evidence that they would be repeatable in a different population tested and retested shortly after intake. And even if these timing postintake issues were resolved, that still leaves open the question of the learning effect of a dog's repeated experience with the test on their behavior during retesting. Finally, it remains far from established what groups of behaviors in dogs can even be said to have the stability that warrants description as personality traits. Absent this understanding, it's difficult to see what any test-retest protocol can establish, other than a record of specific response in a specific place and time.

Showing that results of one subtest are repeatable also does not mean results of other subtests will be repeatable, and repeatability is not the same as having practical importance. Establishing repeatability of warning and biting behaviors may be of much greater interest and likely more challenging, for example, than establishing repeatability of response to a friendly greeting, reaction to a squeaky toy or of jumping or barking behavior in the kennel, on a leash, or during play. The implication of demonstrating test-retest reliability is also influenced by the distribution of the traits or behaviors in the population studied; if most dogs cluster together on their scores, with little variability among dogs, then demonstrating similar scores at a later time would also be rather expected and say very little about test performance in a sample of dogs with a wider range of scores or for different provocations/behaviors.

Among the studies we reviewed, six attempted to establish (Klausz et al., 2014; Mornement et al., 2014; Netto and Planta, 1997; Poulsen et al., 2010; van der Borg et al., 2010; Valsecchi et al., 2011) but only two (Klausz et al., 2014; Valsecchi et al., 2011) demonstrated test-retest reliability to any extent (Figure 1; Table 1). We describe several here as examples of findings regarding test/retest reliability. For example, one Hungarian study demonstrated test-retest reliability for 5 subtests (Klausz et al., 2014). The test was one developed by the investigator which was scored from video. The first testing was conducted on 73 owned dogs and was repeated after 1 year by the same experimenter and in the same place. However, the retested group was very small, including only 19 of the original group, and the subtests used included response to tug-of-war, where the nature of the activity is to elicit simulated aggression in play. This is an example of concerns noted previously regarding practical significance of demonstrating reproducibility of dogs' responses to relatively benign and likely frequently occurring interactions in a presumably stable environment. Another study conducted in Australia reported test-retest reliability for the Behavioral

Study, date [n of dogs or records] Test name	Dog population	Psychometric–type validation: Measures of reliability: Inter-rater reliability	Inter-shelter reliability	Test-retest reliability	Measures of construct validity: Convergent validity	Discriminative validity	Predictive validity	Type of instrument for comparator	Clinical validation: Predictive ability in study population†: Behavior(s) tested	Prevalence (%)	Sensitivity (%)	Specificity (%)	False positive error (%)	False negative error (%)	Estimated in real world population‡: False positive error (%)	False negative error (%)
van der Borg *et al*, 1991 [n=72] I-D test (Video supplement)	SH	X	X	X	X	X	±	TQ	a	38.9§	82.1	63.6	41.0	15.2	69.9	5.1
									b	22.2§	85.7	77.6	47.8	5.0	57.8	3.4
Netto & Planta, 1997 [n=112 & n=37] I-D test	O	X	NA	±	X	X	X*	Q	c	60.0	33.3	93.3	11.8	51.7	51.4	12.0
									d	60.0	40.0	83.3	21.7	51.9	68.7	12.1
Planta & De Meester, 2007 [n=330 & n=220] SAB test (Video)	O	X	NA	X	X	X	X*	Q	e	28.5	84.0	81.4	35.8	7.2	54.3	3.6
									f	28.5	67.0	94.5	17.1	12.2	30.1	6.2
									g	28.5	71.3	94.5	16.3	10.8	28.8	5.5
									h	16.4	69.4	84.2	53.7	6.6	54.4	6.5
De Meester *et al*, 2008 [n=82, n=64] SAB test (Video)	O OB	X	NA	X	±	X	X	Q/CB	NA	X	X	X	NA	NA	NA	NA
van der Borg *et al*, 2010 [n=479, n=76] SAB test	O	✓	NA	±	?	✓	X*	Q/CB	i	35.7	33.3	80.5	51.3	31.5	75.5	13.6
Svartberg, 2005 [n=697] DMA test	O	X	NA	X	±	X	X	Q/CB	NA	X	X	X	NA	NA	NA	NA
Kroll *et al*, 2004 [n=100] I-D test (Video)	OB	X	NA	X	X	X	X*	Q/VB	j	41.0	75.6	45.8	50.8	27.0	79.0	9.2
									k	42.0	90.5	56.9	39.7	10.8	71.4	3.1
									l	42.0	90.5	63.8	35.6	9.8	67.7	2.8
									m	40.0	75.0	51.7	49.2	24.4	77.2	8.4
									n	54.7	75.0	55.8	32.8	35.1	75.6	7.9
Diesel *et al*, 2008a [n=20] I-D test (Video)	SH	±	X	X	X	X	X	T	NA	X	X	X	NA	NA	NA	NA
Bräm *et al*, 2008 [n=60] 2 Swiss tests intended for "dangerous" breeds and one for general control	O	X	NA	X	±	X	X	T	NA	X	X	X	NA	NA	NA	NA
Poulsen *et al*, 2010 [n=236, n=39] RSPCA Australia test	SH LT A	X	X	±	X	X	±	NA	NA	NA	X	X	X	NA	NA	NA
Valsecchi *et al*, 2011 [n=~32-66] I-D test (Video)	SH LT	±	X	±	X	✓	±	Q	NA	X	X	X	X	X	NA	NA
Bennett *et al*, 2012 [n=67] SAFER™ test Modified Assess-A-Pet™ test	OB	X	NA	X	±	X	X*	Q/CB	o	68.7	60	50	27.5	63.7	81.4	13.2
									p	68.7	73	59	20.4	50.1	74.7	8.0
Marder *et al*, 2013; [n=97] MATCH-UP-II™ test	SH A	X	X	X	X	X	X*	Q	q	28.9	39.3	87.0	45.0	22.1	63.5	11.7
Klausz *et al*, 2014 [n=73, n=19] I-D test (Video)	O	✓	NA	±	X	X	X*	Q	r	NR	48	81	NR	NR	67.5	10.9
									s	NR	56	92	NR	NR	42.9	8.3
									t	NR	24	92	NR	NR	63.6	13.6
									u	NR	8	92	NR	NR	84	16
									v	NR	76	73	NR	NR	65.1	5.9
Mornement *et al*, 2014 [n=48 and n=74] B.A.R.K protocol	SH A	✓	X	±	X	X	±	O/LS	NA	X	X	X	NA	NA	NA	NA
Mornement *et al*, 2015 [n=74] B.A.R.K protocol	SH A	X	X	X	X	X	±	O/LS	NA	X	X	X	NA	NA	NA	NA
Dalla Villa *et al*, 2017 [n=97], SAB test	O	X	NA	X	X	✓	±	Q/CB	NA	X	X	X	NA	NA	NA	NA

Figure 1. Results from, and our ratings of, published studies assessing aspects of validity and reliability of canine behavior evaluations focusing on behaviors categorized as aggressive. †When not provided, values were calculated from data in the article using the following link: http://vassarstats.net/[Clinical Research Calculators: Calculator 1]. ‡Real-world prevalence ~ 16%, our estimate for the maximum proportion of shelter dogs that are not screened out at intake that may express a potentially relationship breaking warning or biting behavior in the adoptive home (Patronek and Bradley, 2016). Calculations were performed using the following link: http://vassarstats.net/[Clinical Research Calculators: Calculator 2]. §Prevalence data were not reported in study and were calculated from the raw data to be consistent with the reported estimates for sensitivity and negative predictive

Assessment for Re-Homing K9's protocol in 48 shelter dogs (Mornement et al., 2014). The two tests were performed 24 hours apart by the same person. The correlations between 12 subtests and five behavioral attributes (anxiety, compliance, fear, friendliness, and activity level) ranged from negligible to high, with the authors deeming the predictive value as poor. It is noteworthy that more than 80% of the dogs in this sample had been previously assessed using another instrument and deemed adoptable, and none had expressed any behaviors of concern to the shelter staff in the minimum of 3 days they had been in residence. This calls into question whether any reliability assessment could be applied to an unscreened, presumably more heterogenous, shelter population. In the final study, mostly (95%) stray dogs in a shelter in Italy were assessed for a wide variety of behaviors using 22 subtests for sociability, playfulness, problem-solving skills, trainability, possessiveness, and reactivity (Valsecchi et al., 2011). Subtests were scored using either 2-, 3-, 4-, 5-, or 10-point scales, and scores on these diverse behaviors were summed. A total of 32 dogs were tested at both 20 days and 60 days after intake. Correlation between tests carried out at 20 days and 60 days was moderate ($r = 0.58$, $P < 0.001$). Postadoption retest results are included under predictive validity in Figure 1. In a study that could not be summarized in Figure 1, Bennett et al. tested 33 US shelter dogs at intake and 3 days later using the Safety Assessment for Evaluating Rehoming (SAFER™) test. The subtest results were unstable between the 2 administrations, and they did not vary in the same direction between dogs.

In summary, it appears that demonstrated reliability of any type is largely absent in published studies of canine behavior evaluations. The take-home message is that if sufficient reliability cannot be established, validity is moot.

Construct validity

Construct validation studies are conducted to establish how strongly an evaluation measures what it claims to be measuring. Many types of comparison studies are included under the broad umbrella of demonstrating construct validity. They include convergent validation, criterion validation, discriminative validation, and predictive validation studies, among others. Even experts may disagree on the precise definitions of these terms, although in some cases, these debates may be largely semantic. The important thing is to understand what is being compared and how the associations between test scores and the comparator, as well as their strength, are being measured.

One initial step could be to try to establish convergent validity—in other words, to demonstrate that the results of an instrument (in this case, a canine behavior evaluation) are correlated with sufficient strength with some other behavior evaluation, instrument, or test purported to measure the same thing, usually with both measures being taken at roughly the same time. For example, convergent validity would be demonstrated if scores on a new behavior evaluation were shown to be correlated with scores on another (ideally older or well-established) instrument or test, with adequate strength (e.g., $r \geq 0.70$-0.80) and statistical significance ($P < 0.05$). The implications of demonstrating convergent validity would also be highly dependent on the degree to which the comparator instrument had been convincingly validated in a similar population. In each of the studies we reviewed, a new behavior evaluation was compared with another unvalidated instrument.

We found four examples that we believed qualified as attempts to measure convergent validity; none was able to establish it convincingly (Figure 1; Table 1) (Bräm et al., 2008; Bennett et al., 2012; De Meester et al., 2008; Svartberg, 2005). For example, in the study by Bräm et al. (2008), three different tests used in Switzerland were compared in owned dogs. Behaviors were scored on each test according to a 5-point scale ranging from "open, friendly, neutral" to "overt aggression/attack/biting without threatening/warning". None of the rating scales were defined behaviorally. Agreement (kappa coefficient, κ) among the 3 tests were found to be $\kappa < 0.135$, which is poor. In another, Bennett et al. (2012) used the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) owner-completed survey to categorize owned dogs in the United States into three categories: [1] no to mild aggression, [2] moderate aggression, and [3] severe aggression. In the entire sample, the prevalence of either moderate or severe aggression as defined by C-BARQ was 46/67 (68.7%) dogs and for severe aggression alone was 24/67 (35.8%) dogs. The C-BARQ classifications were then compared with results of the SAFER™ test and modified Assess-A-Pet™ behavior tests. The incidence of false-positives with the SAFER™ and a modified Assess-A-Pet™ test was 50% and 41%, respectively, meaning that behavior tests identified dogs as being aggressive when they were not scored as aggressive on C-BARQ. Unfortunately, it is impossible to disentangle how much of this was due to the poor performance of the behavior tests versus misclassification of the dogs' behavior status by the C-BARQ scale, or both.

In another study, fear of strangers scores on 6 subtests in the socially acceptable behavior test were compared with the similar rating in C-BARQ and found to be negligible to low (r = 0.21, 0.28, 0.32, 0.36, 0.37, 0.47) (De Meester et al., 2008). In the study by Svartberg (2005), the C-BARQ scale was completed by owners of dogs registered with the Swedish Kennel Club and similarly failed to correlate with a provocative test, the dog mentality assessment administered 1 to 2 years earlier, specifically on the aggression factor. The author concluded that "The aggressiveness trait was not validated by this study, and (the dog mentality assessment) seems to be a poor predictor of aggressive behavior in a dog's everyday life" (Svartberg, 2005, p. 122).

Perhaps it should not be surprising that convergent validity has been poor given the heterogeneity of the various tests and scoring systems. An owner-reported scale such as C-BARQ represents an

value. Abbreviations for behavior evaluations: B.A.R.K., behavioral assessment for rehoming K9s; DMA, dog mentality assessment; I-D, investigator-developed and unique to this study; SAB, socially acceptable behavior; SAFER™, safety evaluation for evaluating rehoming. Abbreviations for dog population: A, dogs adopted from a shelter; O, privately owned dogs; OB, privately owned dogs presenting with a behavior complaint; SH, dogs in a shelter; SHLT, dogs in a shelter in long-term residence. Abbreviations for ratings and data: ✓, criteria established with strong correlation and statistical significance; ±, criteria tested but cautions apply (full explanation for our ratings are presented in Table 1); X, criteria not published; X*, if data for predictive ability were provided, we made no rating for predictive validity unless different data were used; ?, unable to assess; NA, not applicable; NR, not reported. Abbreviations for type of instrument for comparator: O/LS, owner-completed Likert-type scale; Q, questionnaire or history; CB, C-BARQ, canine behavioral assessment & research questionnaire; T, test responses; TQ, owner telephone questionnaire; VB, veterinary behaviorist diagnosis. Abbreviations for behavior tested: a, Aggression toward adults during test; b, Aggression toward dogs during test; c, Biting/attack behavior during test versus bite history for potentially aggressive breeds; d, all aggression observed during test versus bite history for potentially aggressive breeds; e, no biting at all during test accepted versus history of previous bite; f, biting in a maximum of one subtest accepted versus history of previous bite; g, biting in a maximum of one of 8 selected subtests accepted versus history of previous bite; h, lunged, snapped, or bit at least once during testing versus future bite; i, lunged, snapped, or bit at least once during testing versus history of biting a human; j, fearful or aggressive reaction toward doll versus diagnosis of dominance aggression; k, fear or aggressive reaction toward doll versus diagnosis of fear aggression; l, fear or aggressive reaction toward artificial hand versus diagnosis of fear aggression; m, fear or aggressive toward artificial hand versus diagnosis of dominance aggression; n, fear or aggressive reaction toward doll versus history with children; o, scores summed for 7 different SAFER™ subtests with aggression as an emphasis; p, scores summed for 9 different modified Assess-A-Pet™ subtests with aggression as an emphasis; q, food-guarding only; r, aggression during take-away-bone; s, aggression during threatening approach; t, aggression during tug-of-war; u, aggression during rollover; v, aggression during threatening approach or take-away-bone.

Table 1
Justifications for our ratings of strength of evidence for studies summarized in Figure 1

Study	Rating	Basis for rating
van der Borg et al., 1991	±	For predictive validity of problem behaviors, only raw agreement for the proportion of correct positive tests was reported, which ranged from 16.7 to 66.7% (Table 2 in reference).
Netto and Planta, 1997	±	For test-retest, total attack behaviors (snapping, biting, attack behavior) were significantly correlated (Spearman's rho [r_s] = 0.77, $P < 0.0001$). Correlation for snapping was $r_s = 0.52$, $P < 0.0017$ and for biting/attack $r_s = 0.65$, $P < 0.0001$). When agreement was evaluated with the kappa statistic across all 43 subtests, kappa coefficients were ranged from negligible to high. Test-retest was carried out in a small number of dogs (n = 37/112).
Planta and De Meester, 2007	NR	No measures of reliability or validity rated.
De Meester et al., 2008	±	This was a pilot study to assess utility of the SAB test for assessing shyness/confidence. For convergent validity, correlations between the SAB score on 6 subtests and fear of strangers on C-BARQ ranged from negligible to low ($r = 0.21–0.47$); correlations were statistically significant ($P < 0.05$, $P < 0.01$, or $P < 0.001$) for 5/6 subtests.
van der Borg et al., 2010	✓	Using video recordings, inter-rater reliability (kappa) among three raters with 10 years of experience was $\kappa = 0.81$ to 0.87 for threats or attacks.
	±	For test-retest, aggression scores were significantly higher during the first test in the morning compared with the second test in the afternoon ($P < 0.001$).
	?	We were unable to categorize or assess the results presented in Table 5 of this article.
	✓	Dogs labeled as aggressive due to a history of having bitten a person (per owner) had significantly higher mean aggression scores than dogs classified as controls ($P = 0.003$), demonstrating discriminative validity.
Svartberg, 2005	±	For convergent validity, correlations between variable scores on the DMA test and human directed aggression/fear factors from the Swedish version of C-BARQ were negligible ($r = −0.03$ to 0.12).
Kroll et al., 2004	NR	No measures of reliability or validity rated.
Diesel et al., 2008a	±	Weighted kappas for inter-rater reliability were almost all poor to moderate, including when restricted to experienced staff. Kendall's coefficient of concordance was not statistically significant for 4/6 of the aggression characteristics tested, including aggression during handling and aggression when meeting another dog.
Bräm et al., 2008	±	For convergent validity, agreement (kappa) for weighted average of pairwise comparisons for intraspecific behavior ($\kappa = 0.133$, $P = 0.14$) and interspecific behavior toward humans ($\kappa = 0.135$, $P = 0.014$) across the three tests was considered slight. The three tests appeared to agree on dogs that showed open, friendly, neutral behavior, which was most dogs.
Poulsen et al., 2010	±	For test-retest reliability, dogs were assessed in the shelter from 24 hrs to months after intake. The actual shelter assessment was replicated in the home (mean of 80.9 days between tests), with low and non-significant correlation between the two times: $r = 0.29$, $P = 0.08$, as opposed to the assessment being used to predict more general behaviors after adoption. As the second test was performed in the home, these results can also be viewed as an assessment of predictive validity.
Valsecchi et al., 2011	±	For inter-rater agreement, correlations were very high to perfect for simple behaviors (e.g., behavior in kennel, playfulness) but much more variable for more contextually complex behaviors (e.g., reactivity, sociability). Only 19 long-term stay (minimum 1 year in shelter) dogs were tested.
	±	For test-retest reliability in the shelter, there were only moderate correlations for retesting at 20 days and 60 days after intake (n = 32, $r_s = 0.58$, $P < 0.001$).
	✓	For discriminative validity, scores were calculated for 30 long-term dogs categorized into one of three extreme groups based on preassessed behavioral profiles (10 dogs impounded by police and relinquished because of biting history, 10 with a veterinarian-assessed history of fearfulness, and 10 dogs deemed adoptable.) Median scores were significantly different in the expected directions for the three groups.
	±	For predictive validity, only moderate correlations between testing at 20 days after intake versus 4 months after adoption in new home (n = 34, $r_s = 0.44$, $P = 0.009$).
Bennett et al., 2012	±	For convergent validity, correlations between aggression history (no/mild vs. moderate/severe) and evaluations were low and/or not statistically significant (SAFER™, $r = 0.217$, $P = 0.078$; modified Assess-A-Pet™, $r = 0.342$, $P = 0.005$).
Marder et al., 2013	NR	No measures of reliability or validity rated.
Klausz et al., 2014	✓	For inter-rater reliability (kappa), there were high correlations ($\kappa = 1.00$ for aggression, 0.87 for latency of rollover, 0.84 for barking) reported from parallel coding of tests from 15 dogs.
	±	For test-retest reliability, 19 owned pet dogs were evaluated. Retesting was conducted after 1 year; none of the tests (friendly greeting, take-away-bone, threatening approach, tug-of-war, rollover) were different on retesting.
Mornement et al., 2014	✓	For inter-rater reliability, most subtests had correlations from moderate to strong.
	±	For test-retest reliability, correlations between 12 subtests and five behavioral attributes ranged from negligible to high ($r \sim −0.02$ to 1.00).
	±	For predictive validity, correlations between B.A.R.K. scores and owner ratings of the dogs behavior were low to moderate for fear ($r = 0.42$, $P < 0.01$) and friendliness, ($r = 0.49$, $P < 0.01$). Correlations between B.A.R.K. scores and owner ratings for anxiety, compliance, and activity level were not statistically significant.
Mornement et al., 2015	±	For predictive validity, regression analysis showed that there was a statistically significant relationship between B.A.R.K. scores for fear and anxiety and owner ratings on behavioral subscales for fearful/inappropriate toileting, but the magnitude of the correlation was moderate ($r^2 = 0.31$ [calculated $r \sim 0.56$]) with 30.6% of the variance explained. The relationship between owner ratings of problem behavior and aggression scores on B.A.R.K. was not statistically significant and thus not considered predictive.
Dalla Villa et al., 2017	✓	Mean C-BARQ scores for familiar dog aggression, owner-directed aggression, and stranger-directed aggression were significantly higher for dogs categorized as aggressive on the SAB compared with scores for dogs not so categorized, thus showing discriminative validity.
	±	For predictive validity, logistic regression showed that none of the SAB subtests were predictive for C-BARQ scores for owner-directed or familiar dog aggression. Seven of the SAB subtests were predictive for C-BARQ scores for stranger-directed aggression.

✓, criteria established with strong correlation and statistical significance; ±, criteria tested but cautions apply; NR, criteria for rating not reported; SAB, socially acceptable behavior.

impression of cumulative behavioral trends over some time during the period of ownership, whereas a shelter behavioral evaluation reflects behavior at a specific time and place. The effect of using different experimenter-defined thresholds and scoring systems (e.g., dichotomized into present/absent, degrees of escalating fear or aggression) as well as different systems for combining these results into overall scores is unknown.

A closely related type of validation, criterion validation, seems essentially the same as convergent validation, except that the comparator is either an objectively measurable outcome or an

instrument considered a gold standard. We are not aware of a gold standard or objective measure that has been successfully used for a canine behavior evaluation. Blood and fecal cortisol levels have been explored (DePalma et al., 2005; Hennessy et al., 2001), but exactly what differences in cortisol levels would say about different kinds of behavior is unclear. Streiner et al. (2015, p. 11) have noted, however, the following pitfall when considering any measure as having criterion status "Although there are often measures which have, through history or longevity, acquired criterion status, a close review usually suggests that they have less than ideal reliability and validity…". So paying careful attention to what was being promoted as a criterion measure bears careful scrutiny, if such status were to be claimed. Because no external gold (criterion) standard exists, it was not possible for any of the studies to meet this criterion.

There were three studies that demonstrated discriminative validity by showing statistically significant differences in mean or median scores for aggression for dogs categorized as aggressive versus dogs not so categorized (Dalla Villa et al., 2017; van der Borg et al., 2010; Valsecchi et al., 2011). However, an important limitation of this type of comparison is that the clinical importance of a difference in mean scores of a given magnitude from a subjective measure, even if criteria are well defined, cannot be readily assessed. For example, a difference in aggression score between two dogs is not as clinically interpretable as a particular difference in age, body temperature, or weight. So while showing the ability to distinguish between very different/extreme groups is necessary, it constitutes a fairly low level of evidence for overall validity.

Predictive validation refers to the behavior evaluation results being correlated with a past or future outcome measure such as behavior in the home (as opposed to something measured at approximately the same time). Such an exercise, although important, is not the same as showing adequate predictive ability for individual dogs in future homes as demonstrated by the rate of false-positive and false-negative results. We will elaborate more on that critical point under Issue 3: Predictive validity versus predictive ability, and explain why much of what is framed as predictive validation actually reports statistics for predictive ability (e.g., sensitivity, specificity, false positives, and false negatives). Although six studies might be classified as attempting to demonstrate predictive validity independently of predictive ability, none did so convincingly (Figure 1; Table 1) (Dalla Villa et al., 2017; Mornement et al., 2014; Mornement et al., 2015; Poulsen et al., 2010; van der Borg et al., 1991; Valsecchi et al., 2011). Among these, Valsecchi et al. (2011) provide an example. Extending the test-retest results obtained while residing in the shelter, they further compared the predictive validity of tests carried out at 20 days after intake with those carried out in the home four months after adoption. The correlation calculated for 34 dogs was low ($r = 0.44$). This does not strike us as surprising, given the heterogeneity of the behaviors tested and the length of time in between tests, which could be an interval of >1 year from intake testing to testing in the home. In a sixth study described under test-retest, the second test was carried out in the home on 39 dogs approximately 80 days after adoption, so also explored predictive validity (Poulsen et al., 2010). By virtue of being adopted, the dogs tested represented those that initially passed their evaluation in the shelter. One of the findings from that study was that there was no statistically significant correlation between the assessment conducted before adoption in the shelter and after adoption in the home for those tests which involved direct contact between the dogs and the assessor. However, scores for food guarding ($P = 0.02$), reaction to noise and movement ($P = 0.03$), and reaction to toys and play attempt ($P < 0.01$) were significantly correlated before and after.

When evaluating claims about predictive validity, it is also important to consider the practical relevance of what is being assessed and predicted. For example, Christensen et al. (2007) used a modified version of the Assess-A-Pet™ test to determine adoptability of dogs in a US shelter. Sixty-six of the 279 adopters the researchers attempted to contact within 13 months (median of 4 months) of adoption responded to a behavior questionnaire over the phone. The authors reported that the test failed to predict lunging, growling, snapping, and/or biting that occurred in 40.9% of dogs after adoption (the prevalence was 71.2% if they included barking under their "aggression" umbrella). They concluded that behaviors directed at territorial, predatory, and intraspecific conflict provocations had not been adequately tested for, and suggested that the test be modified to minimize the number of dogs expressing these behaviors being adopted from the facility.

The conclusion from the study by Christensen et al. (2007) bears closer examination as some surveys suggest the frequency of the various behaviors they queried as potentially problematic is within the range of normal. For example, in a survey of 3,226 owners recruited from general veterinary practice in Canada, 41% reported that their dog had growled at a household member, including during play, and that 15.6% had bitten a household member (Guy et al., 2001). Christensen et al. (2007), however, asked the dogs' adoptive owners about a much wider variety of contexts. These ranged from an unfamiliar person entering the house or yard or seeing people outside the home, to being in the presence of a squirrel, cat, etc., to seeing another dog while in a car or while being walked on a leash. These situations were in addition to responses to various kinds of provocations on the part of the owner, which included among others, lifting the dog, approaching the dog while on furniture, disturbing while resting, nail-trimming, and verbal or physical punishment. Because it was grouped with snapping, there was no confirmation in this sample of any dog actually biting anyone. Thus, this study's results continue to beg the question of the practical relevance of what exactly is being measured by various subtests in behavior evaluations. As Valsecchi et al. (2011, p. 173) have acknowledged in their study of dogs in four Italian shelters "…it is worth mentioning that [aggressiveness] is a normal component of dogs' behavior: even the most docile and friendly dog could growl to the owner or to an unfamiliar dog under certain conditions."

In summary, although a few studies reported statistically significant findings for various aspects of construct validity, none of the studies demonstrated compelling evidence of construct validity in a more global sense.

Issue 2: Limitations of correlation and regression

Another potential point of confusion is the interchangeable use of correlation and agreement. Statistically speaking, the two are not the same. Indeed, Mukaka (2012, p. 69) indicated "…misuse of correlation is so common among researchers that some statisticians have wished that the method had never been devised at all." The purpose of a correlation is to "…assess the strength and direction of the linear relationships between pairs of variables" (Mukaka, 2012, p. 71). Strength of the relationship is assessed by the magnitude of the correlation coefficient (depending on the form of the data, this could be Pearson's coefficient, Spearman's [rho] coefficient, Kendall's tau, or Cronbach's alpha). A conventional rule of thumb suggests that values between 0-0.30 are considered negligible, 0.30-0.50 low, 0.50-0.70 moderate, 0.70-0.90 high, and 0.90-1.00 very high (Mukaka, 2012, citing Hinkle, 2003), although neither this terminology nor cutoffs are universally acknowledged or used consistently in published articles.

It is scientifically imprecise to frame raw correlation, concordance, or correspondence as agreement. When it is desired to show actual agreement between two continuous variables (e.g., summed

scores), the Bland-Altman method should be used (Giavarina, 2015). For categorical data, such as the presence/absence of a behavior, instead of simply comparing the proportion of all observations in the cells of a 2×2 table that are in concordance versus those that are discordant, the kappa statistic is the correct measure of agreement to use. Kappa will make the important correction for agreement beyond chance, which will inevitably be less—and often substantially less—than raw concordance. For example, Planta and De Meester (2007) reported that the raw agreement between biting behavior among owned dogs during the socially acceptable behavior with their behavior in the home as reported by an owner questionnaire was 81.8%, but when corrected for agreement beyond chance, they indicated that the value for kappa was 0.42 (42.0%), a much less impressive number. For other examples of use of kappa, see the study by Bennett et al. (2015) and Diesel et al. (2008a).

Regression techniques are sometimes used instead of correlation. Although both techniques describe the association between two variables, in regression, one variable is deemed dependent on the other, as opposed to just covarying. The correlation coefficient can be derived by taking the square root of R^2 (the proportion of variability in the dependent variable accounted for by the factors in the model) from a regression model. The main advantage is that if multiple regression is used, the regression coefficients will reflect adjustments for other variables in the model, if such effects exist. The *P*-values are not interpreted any differently.

Issue 3: Predictive validity versus predictive ability

As discussed previously, predictive validity can be established when scores on a behavior evaluation in a population of dogs significantly correlate with a second variable that can reasonably be thought of as dependent on the characteristics being measured, such as future behavior in the home. However, meeting these criteria does in no way imply how accurate the prediction of future behavior will be. Predictive ability, by contrast, reflects the likely accuracy of that evaluation when predicting behavior of individual dogs in the real world, as reflected by the number of errors (i.e., false-positive and false-negative results).

For predictive validity to be met, the bar is lower than that for acceptable predictive ability. Predictive validity can technically be established if the "somewhat more often than not" threshold is met (explained under Issue 4: Statistical versus clinical significance), whereas for a test to have acceptable predictive ability, the number of false positives and number of false negatives must be deemed low enough for the evaluation to have practical value in shelters. This will be determined by the sensitivity and specificity of the test(s), as well as the prevalence of the behavior(s) in the population in question. A key corollary of this principle is that the predictive ability of an evaluation with a particular sensitivity and specificity will be much better in the population in which it was developed (usually a sample of dogs preselected to have a high prevalence of the condition, such as abnormally aggressive behavior demonstrated by previous biting history) than the same evaluation in a more representative population of dogs with a lower prevalence. Finally, it is essential to be mindful that in a low-prevalence situation, it is unsurprising that there would be few false-negative results, so passing the test would provide little new information. By comparison, in the same low-prevalence population, the risk of false-positive errors is of much greater concern because that would label a dog as having a behavior problem, possibly delaying or precluding adoption.

In a previous article, we created hypothetical scenarios from extensively validated instruments for subjective outcomes in human health or behavior but consistent with real-world data from dog behavior evaluations, and demonstrated that even if a validated canine behavior evaluation were available, it would not be useful for making "thumbs up/thumbs down" decisions with respect to adoption due to the high prevalence of false-positive results (Patronek and Bradley, 2016). We also showed that owing to the likely low prevalence of problematic, adoption-preventing behaviors in shelter dogs, little more would be learned from a negative test because most dogs would not exhibit behaviors of concern in the home in response to the triggers in question.

Here, we examine attempts to establish the specific predictive ability of individual instruments as recounted in the literature. As Figure 1 shows, reported and/or calculated false-positive rates for various individual dog behaviors in the study populations of mostly owned dogs ranged from 11.8 to 53.7% (mean, 35.1%); false-positive rates we estimated for real-world shelter populations using the stated values for sensitivity and specificity ranged from 28.8% to 84% (mean, 63.8%). The mean false-negative rate in study populations was 25.6%, whereas the estimated mean false-negative rate for typical shelter populations was estimated as 8.5%. A level of false-positive error sometimes exceeding 70% goes far beyond the simple random chance performance of "no better than flipping a coin", and suggests that the tests were measuring a different construct entirely or inducing a behavioral response that was an artifact of the test situation. This degree of error may be of lesser importance when selecting dogs for a breeding program for example, where maximizing negative predictive value is the most critical metric and consequences of a false positive are minimal (i.e., not being bred). By comparison, avoiding false-positive error has a much greater level of importance when tests are used to determine the fate of dogs in shelters.

It is noteworthy that with the exception of the first article published on this topic (a study of shelter dogs in the Netherlands by van der Borg et al., 1991), and one article that evaluated just the food-guarding subtest administered to a group of dogs in a US shelter (Marder et al., 2013), all of the studies reporting predictive ability of behavior tests have involved owned, rather than shelter dogs. These include studies of pet dogs in the United States (Bennett et al., 2012; Kroll et al., 2004), in Hungary (Klausz et al., 2014), and in the Netherlands (Netto and Planta, 1997; Planta and De Meester, 2007; van der Borg et al., 2010). From a pragmatic perspective in the early stages of research, using owned dogs is understandable, but it should be recognized that if carried out in the home, that environment poses its own challenges with respect to standardization.

Issue 4: Statistical versus clinical significance

We believe that conflating statistical versus clinical significance is another source of potential confusion when interpreting or discussing results of canine behavioral evaluations (Wasserstein et al., 2019). Although correlation can give an indication of the likely strength of the relationship between scores on a dog behavior evaluation and another variable categorizing that dog, statistical significance testing is needed to ascertain the probability that any correlation is indeed a real effect in the larger population from which the sample was drawn. However, a demonstration of statistical significance says nothing about whether the correlation has practical importance in the real world, only the probability that the relationship in the sample is real in the larger population. Along those lines, small highly significant values such as $P < 0.001$ only mean that you are more certain that the relationship is real than you are with a *P*-value of say $P = 0.04$. A critical point here is that statistical significance is heavily influenced by sample size, with increasing the sample size in a study being akin to using a magnifying glass to detect a defect such as a scratch in a piece of antique pottery. Consequently, with a sufficiently large sample size, even weak correlations can become statistically significant. This is one

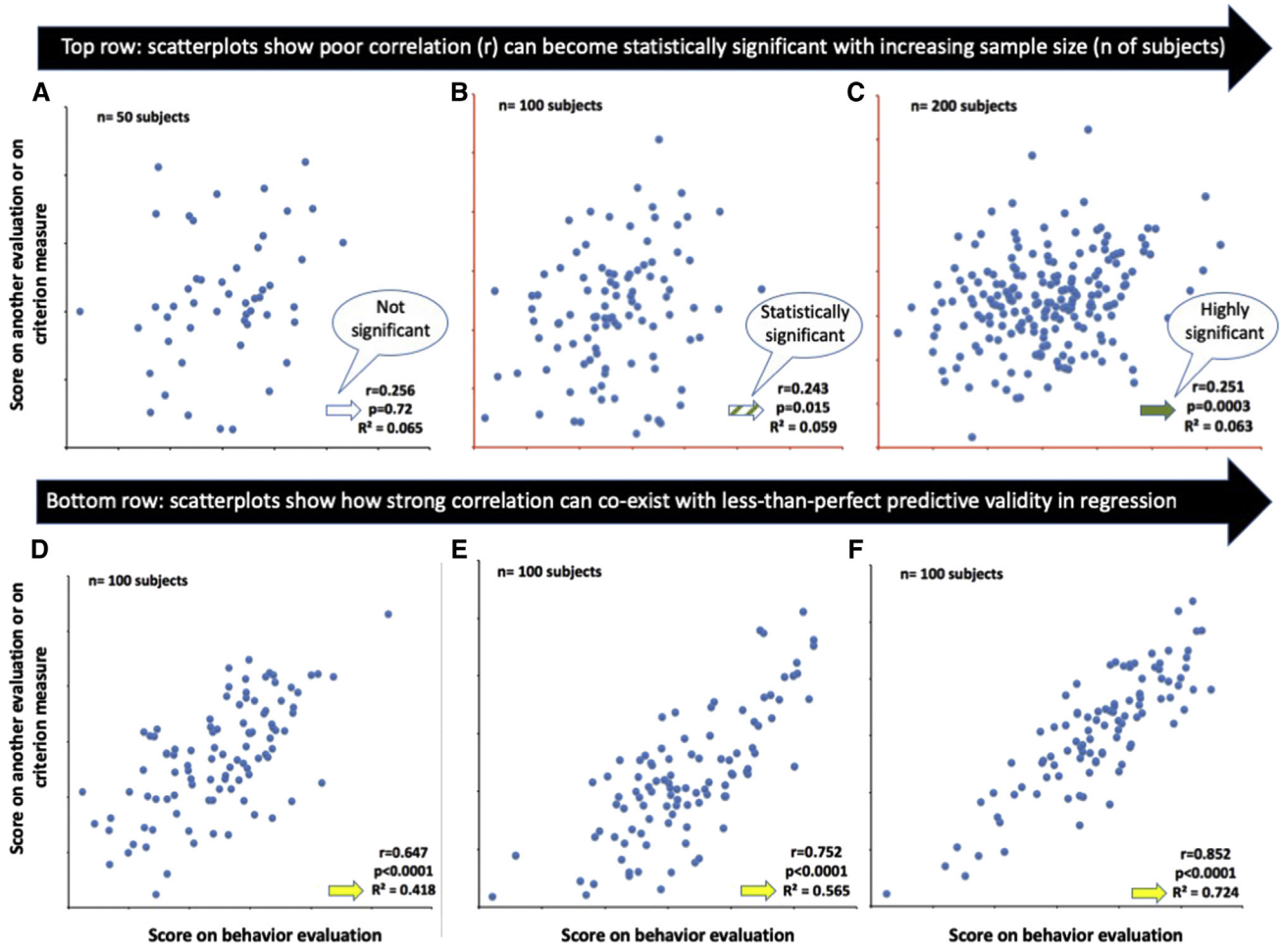

Figure 2. Example of a possible relationship between sample size, statistical significance, and agreement between scores on a behavior evaluation. The figure shows data simulated in an Excel spreadsheet, which was programmed to deliver random samples of a given size (number of subjects) with specified mean, standard deviation, and correlation. (a-c) negligible correlation of $r = 0.243$-0.256 ($\sim$25%) has minimal practical significance ("somewhat more often than not" threshold), but *P*-values show how it becomes "statistically significant" (hashed arrow) and even "highly significant" (green arrow) when sample size increases, even though the nature of the relationship between score on the behavior evaluation and comparator measure does not change. (d-e) progressively increasing strong correlations of $\sim$65%-85% (yellow arrows) are statistically and practically significant, and help establish different types of "construct validity" for the evaluation: this would be "convergent validity" if the comparator measure was score on another behavior instrument, and "predictive validity" if the comparator measure were a criterion (gold standard) instrument or a clinical outcome measured in the future after the behavior evaluation was performed. However, as the scatter of the points in these panels suggests, scores on the behavior evaluation still explain only from 41.8% to 72.4% of the variance in the comparator measure (R^2). This is one reason why the correlation for predictive validity is not the same as the predictive value of positive or negative test (predictive ability).

reason why statistically significant predictive validity can coexist with poor predictive ability (i.e., high false positives and false negatives) in the real world. So in the most conservative situation, all that statistical significance implies is that two variables are likely to be related somewhat more often than not. The possibility of reaching only the "somewhat more often than not" threshold for correlations that are of low to moderate strength is important to keep in mind to avoid overinterpreting the importance of statistical significance when assessing claims of reliability or validity.

We illustrate these principles in scatterplots of two variables (for purposes of discussion, scores on a canine behavior evaluation and some other instrument or criterion measure) (Figure 2). In this simulation, we have deliberately omitted trend lines or lines of best fit through the data because that visual representation tends to draw viewers to the line and the points closest to it; our goal with this simulation was to encourage viewers to focus on the pattern of data points instead, when thinking about what correlation reveals about the underlying data and relationship between two variables. Scatterplots C and D would be examples of the "somewhat more often than not" threshold. As Figure 2 demonstrates, relationships that visually appear considerably less than perfect can be statistically significant, thus potentially leading to claims of being validated.

Issue 5: Establishing overall validity

Perhaps the most important general principle to appreciate is that overall validity can only be claimed after multiple rigorous studies, using precisely the same instrument and conducted in different populations exactly the same way, have demonstrated acceptable strength of association and statistical significance. As Williams et al. (2006) have emphasized, "Validation is not an event whereby an instrument is deemed valid once and for all—rather, it is a periodic and systematic examination of a measure that helps to build a body of evidence in support of an instrument's validity and utility." Streiner et al. (2015, p. 12 and p. 230) also noted that "…the burden of evidence in testing construct validity arises not from a single powerful experiment, but from a series of converging experiments" and that "…we can never say, 'This scale is valid', because it is not the scale that is being validated. The most that we can conclude regarding the results of any one particular study is,

'We have shown the scale to be valid with this group of people and in this context.'" Streiner et al. (2015, p. 12) have also emphasized that "…reliability and validity are not fixed, immutable properties of scale, that once established, pertain to the scale in all situations. Rather, they are a function of the scale, the group to which it is administered, and the circumstances under which it was given".

The environmental aspect is especially relevant to establishing validity of canine behavior evaluations, which may perform very differently in home and shelter environments, and within sub-environments in the shelter (e.g., meeting rooms, outdoor play areas, offices, hallways, and kennels), where the circumstance may not seem at all equivalent when experienced through the eyes, ears, noses, and previous experiences of dogs. Another issue is the influence of attachment (or lack thereof) to persons conducting or present during the tests. Kis et al. (2014) showed that in a sample of 50 Hungarian pet dogs, more warning behaviors categorized as aggression were recorded when their owners were present than when they were absent. Nevertheless, the frequency of aggressive responses was described as being rare. Using different tests, De Meester et al. (2008) also demonstrated that the presence or absence of the owner during a test influenced the results. Barrera et al. (2010) reported that shelter dogs showed a higher frequency of proximity and ears down behavior when they were exposed to an unfamiliar experimenter who acted passively compared with the person being more active. This was not observed among owned dogs. However, attachment figure presence effects on test behavior may not be confined to owned dogs, as Gácsi et al. (2001) showed that it was possible to detect gains in attachment that affected behavior after as little as 3-10 minute exposures of a dog living in a shelter to an unfamiliar person. Although it's not known whether these behaviors would be long-lasting and thus conform to the definition of attachment between mothers and offspring, and more recently, between owners and pet dogs often studied with regard to this concept, the results on versions of the Ainsworth test are strikingly similar to those used to evaluate attachment in long-term relationships.

The staff performing the evaluations are also an important potential source of variability and error in measurement, particularly given the complexity of canine behavior evaluations. For example, in their survey of 11 shelters in Australia, Mornement et al. (2010) describe how most did not have any documentation or step-by-step instructions on how to administer whatever behavior evaluation was being used, nor was there any standardization between or within shelters as to how soon after intake the evaluations were conducted. Evidence of training and auditing of that training to ensure consistency was limited at best. Given the limited resources and other constraints faced by many shelters just to stay abreast of daily intake and animal care requirements, we believe this is not an uncommon situation in the United States as well. As noted by Mornement et al. (2010), minimizing measurement error in daily field use must be an ongoing process that can only be ensured through quality of initial training and auditing of training to be sure original standards continue to be adhered to. Factoring in staff turnover in shelters, this is not a trivial concern.

A major point in the validation process for instruments used in human health care is that it cannot be assumed that an instrument developed in one population (e.g., children or adolescents) would be applicable to a different population (e.g., adults) who may have clinically meaningful differences in demographic characteristics or risk profiles. As additional examples, an instrument designed to measure fatigue or pain in a population of patients with cancer might not be suitable to measure those same constructs in a population of patients with a different disease, say rheumatoid arthritis. Indeed, Streiner et al. (2015, p. 12) have emphasized that "Potential users of a scale must assure themselves that reliability and validity have been determined in groups similar to the ones with which they want to use it." The parallel situation with dogs would include consideration of applicability in immature dogs versus adults. Other repercussions would be that an instrument developed in dogs living as pets in homes, service dogs, or police canines, for example, could not be just assumed to be useful in a population of dogs living in shelters, without further and substantial validation. With respect to dogs living in shelters, the fact that they have been relinquished or lost and are in an unfamiliar environment introduces critical environmental differences compared with studies in populations of dogs living in homes as pets. Indeed, as we mentioned earlier, it is possible that shelters may be sufficiently different from one another that an evaluation would need to be individually validated in each one. In fact, "shelter" is a term that covers an enormous variety of environments and simply means any place where animals that are not currently owned by private individuals are kept.

Face validity—the elephant in the room

In contrast to the statistical methods used to establish construct validity, face validity is a common sense determination that an instrument appears to measure what is intended. As a subjective assessment of reasonableness, it does not involve any statistical calculations. Thus, it was not possible to include face validity in the characteristics of studies reviewed in Figure 1. In our opinion, this critical issue is often glossed over during discussion of behavior evaluations, and there does appear to be an underlying assumption that canine behavior evaluations (and therefore, the subtests within those evaluations) have face validity.

Regardless of the target expressed behavior settled on, perhaps the most important question to consider when thinking about face validity would be whether it is reasonable to assume that any behavior on a test battery (or individual subtest) administered at a single time in a stressful, unfamiliar shelter environment surrounded by strangers can generally be considered a reasonable surrogate for that dog's behavior in a future home, where she/he would presumably be attached, settled, and frequently in the company of familiar people, and potentially have very different concepts of territory and different reactions to stimuli. Embedded in this concept is an assumption that the dog's behavior in the home would be invariant after adoption and that the time of evaluation after rehoming to evaluate the success of the behavior evaluation in the shelter would not be important. That assumption does not strike us as plausible. This concern is often expressed or alluded to when various instruments are discussed in publications, but remains largely unresolved. In our opinion, when this issue is raised, the discussion quickly moves on to ways to make a better widget, so to speak, as opposed to critically thinking about whether a widget is needed in the first place.

Given the absence of information, we have not included face validity as a category in Figure 1, but that omission does not indicate a lack of importance of the topic when assessing a canine behavior evaluation. Indeed, it may well be the most fundamental.

To begin with, face validity requires a clear statement about precisely "what" is being measured. We have not found a study where this "what" is clearly stated, although 2 main issues are mentioned to demonstrate the need for the tests: dog bite risks and adoption return risk. We discuss both these rationales in more detail in the section Dogs and risk, but some specific description of agonistic behavior as the target of face validity in behavior evaluations is needed here.

Behavior evaluations are used presumably to ensure good matches and identify and mitigate risks. In human health care practice, when developing an instrument used in assessing a

subjective outcome such as quality-of-life or degree of disability or pain, it is expected that experts as well as patients with the condition of interest are typically involved (Department of Health and Human Services, 2009). This practice suggests that if facilitating safe and lasting adoptions is a goal of behavior evaluation, pet owners and potential dog adopters who know what they find desirable and what they find unacceptable in a relationship with a companion dog would be important to include at an early stage of the development of a canine behavior evaluation for shelter dogs. This type of exercise has, to our knowledge, never been carried out. In our opinion, the result of such an exercise would be to emphasize the wide range of experience, motivations, and hopes that potential adopters bring to the process of choosing a companion pet.

Authors sometimes begin with descriptions of dog bite injuries as a public health issue, from which one can infer that the instrument is an attempt to identify dogs with some degree of elevated likelihood to inflict an injurious bite (Dalla Villa et al., 2017; Kroll et al., 2004; Netto & Planta, 1997; van der Borg et al., 2010). The tests then attempt to elicit biting or warning behavior by introducing various stimuli as described previously. However, the behaviors defined as problematic—and even those deemed disqualifying for adoption—vary from test to test, as do the stimuli used to elicit them. An important question is at what level of provocation could biting and warning behaviors be expected from any dog, including those that we deem "normal"? There is also no consensus regarding at what point the intensity, frequency, duration, or context of the behaviors included cross a threshold from the species norm into unacceptable behavior for animals living as companions to humans. Because most dogs will exhibit warning behaviors at some point in their lives, it is not just an academic question to ask "how much is too much?" Even among the validation efforts for various evaluations that followed up with adopters, only one (Marder et al., 2013) reported whether the dogs had actually bitten anyone after adoption (among 89 dogs that were reported as approached or touched while eating, 1 dog was reported as biting "sometimes" over delicious food items, and 2 dogs reported as biting "rarely"; there were no reports that any of the bites had injured anyone). Another lumped warning behaviors along with biting into a single category (Christensen et al., 2007), with the implication that any expression of any postadoption warning behavior, even without any harm, was an indicator of a failure of the behavior evaluation.

This question of "what is being measured" goes hand-in-hand with the closely related concept of content validity, which addresses whether all relevant facets of whatever has been defined under face validity (item generation in statistical parlance) have been included in the assessment process. Although mathematically complex statistical techniques such as factor analysis are sometimes used to confirm associations among different items in a scale, thus helping to assert that they are measuring roughly the same thing, ultimately, establishing content validity for battery tests used in canines is mostly a subjective determination. Consideration of the dimensionality of an evaluation comprising a series of subtests is crucial here, particularly when subtests are lumped together to create an overall score. If a battery of tests is not measuring one identifiable, underlying construct such as danger from a bite, but instead many different things, as suggested by the range of everyday situations, play, and provocative stimuli, as well as behaviors they are intended to assess, then the rationale for combining scores from all subtests into an overall score may not only reduce face validity but beg the question as to what exactly is being measured in the first place. These differences often also preclude making meaningful comparisons across studies.

To illustrate the extent of this challenge, the types of provocations or interactions mentioned in various reports of behavior evaluations include but are not limited to the following: routine stroking, petting, handling, or brushing; hugging; squeezing toes; examining teeth; touching the muzzle; putting on a collar; approaching the dog in friendly manner; approaching the dog in the kennel; staring into the dog's eyes; staring at the dog while trying to appear imposing such as by extending arms or making a striking motion; knocking on a car window and attempting to open the door; introducing a live or fake dog; presenting a doll or a silhouette of a cat; chasing a dog around with a doll; appearing dressed in an unusual manner or costume such as large hats, buckets on heads, or draped in sheets; confronting with a human dummy figure; approaching with a toy lawnmower; making sudden motions such as a flapping blanket or opening an umbrella in the dog's face; offering a tug-of-war toy; moving suddenly or running to initiate play; making/causing loud noises such as bells or gunshots; issuing verbal or hand commands; walking the dog while on leash, ignoring the dog; leaving a dog alone in a room or car; removing food, treats, or possessions with a fake hand or stick (Bennett et al., 2012; Christensen et al., 2007; Mornement et al., 2014; Svartberg, 2005; van der Borg et al., 2010; Valsecchi et al., 2011; van der Borg et al., 1991).

Consider aggression, which has been described as a collection of distance-increasing behaviors (Beaver, 2009). This is by no means the only definition being used, and it should be noted that the term had already become so murky by the 1960s that no less a luminary in animal behavior science as John Paul Scott considered it scientifically useless (Scott, 1966). According to Voith and Borchelt (1996) "despite thousands of research reports investigating aggression, no uniformly acceptable definition has ever been agreed upon," and that "a simple, precise, unitary definition of aggression has been impossible to formulate." In a recent compilation of articles on dog bites, Mills (2017) points out that "a fundamental principle of 'behaviorism' and the science of animal behavior is that behaviors are objectively describable, but this does not appear to be the case when it comes to 'aggression'…." (Mills, 2017, p.15). Lockwood (2016) points out that classifications also vary widely, with some based on the function of or motive for the behaviors, some on context, and others on the recipients of the behavior. Nevertheless, using the definition of distance-increasing behaviors as an example, this implies that any behavior a dog uses to increase distance could fall under the general umbrella of the term. This makes the scenarios test designers might legitimately invent, along with the behaviors they are attempting to elicit, nearly infinite. Furthermore, given this diversity, making any inferences across studies strikes us as problematic. Most of these provocations, the justifications behind them, and their subsequent interpretations have not been empirically evaluated. However, data have shed doubt on the ability of model devices such as a human-like doll (Barnard et al., 2012) and a fake dog to assess responses to real children and dogs (Shebelansky et al., 2015). Both of these are used in some behavior evaluations.

Therefore, to summarize, one explanation for the generally poor predictive ability of the behavior evaluations studied so far may be that the tests, or at least certain subtests, lack basic face validity. For shelter dogs, it may well be that some evaluations are detecting little more than fear, uncertainty, and lack of adaptation to the shelter environment when they are conducted, as opposed to some stable, generalizable behavior that will carry over into a different situation. However, performance of tests on owned dogs has not been any better.

Our findings in perspective

The most important conclusion that can be drawn from the findings summarized in Figure 1 is that, based on accepted criteria

for measuring subjective outcomes, no canine behavior evaluation or individual subtest has sufficient evidence to be deemed reliable and valid for routine use in shelters. Given the considerable heterogeneity of the studies with respect to dogs, tests, type of stimuli, behaviors elicited or observed, and interpretations of those behaviors, it would also not be appropriate to daisy chain the sporadic statistically significant results from inconclusive studies to argue that the collective body of work is somehow greater than the sum of the parts, thereby disregarding the bulk of the evidence. The sensitivity and specificity of current behavior evaluations, if applied to the general population of shelter dogs that are not screened out immediately due to history of a bite or behavior at intake and that have a low prevalence of potentially problematic behaviors (we used $\leq$16% based on published shelter data), indicates that a positive finding (i.e., failing the test) would be misleading in that an unacceptably high proportion of positive tests would be in error (false positives). This is particularly germane if the only problematic behavior exhibited by the dogs in the shelter was during provocative testing. These latter results from published dog behavior evaluations are in line with those from previous hypothetical scenarios (Patronek and Bradley, 2016).

In human clinical use, it is also well-established that once validated in a specific population, an instrument cannot be modified (at least without permission of the developers) as even minor changes to a scale, series of questions, or other rating system may alter performance and negate any claims to validity. Indeed, it is not uncommon for developers of scales to make this requirement an explicit condition for using a particular instrument, which may be copyrighted. There is evidence to suggest this core principle is regularly ignored in the practical application of canine behavior evaluations in shelters, with shelters modifying items, deleting items, and/or adding items of their own, as well as creating their own evaluations from scratch (D'Arpino et al., 2012; Mornement et al., 2010; Mohan-Gibbons et al., 2018). This type of diversion from accepted practice is a cause of great concern, and underscores the pitfalls of what happens when procedures are implemented in an environment where there is little to no regulatory or professional oversight to prevent such misuse, however, well-intentioned. We concede this is currently a moot point, as it is not possible to invalidate something that had never been validated in the first place. But it does highlight a concern for the future.

Some of our skepticism about the feasibility of validating such instruments for day-to-day application in shelters arises from the history of canine behavior evaluations. The practice of formal evaluations of canine behavior in Europe largely predates that in the United States and reflects purposes that are absent in the United States. One of these is an effort to guide breeding decisions among purebred dogs in an attempt to select animals that may be more likely than the general population to produce progeny that will express very specific working task aptitudes. In some cases, this means specifically targeting the reduction of the expression of behaviors categorized as manifestations of fear and aggression below what is presumed to be the level typical for the breed (van der Borg et al., 2017). With some breeds in some countries, this has been taken to the length of testing breeding animals as a requirement for registering their offspring as pedigreed individuals (van der Borg et al., 2010; van der Borg et al., 2017). That endeavor has its own daunting limitations and challenges, but they are dramatically different from those of attempting to predict the behavior in an adoptive home of a dog currently living in a shelter, or to make a life-and-death decision for an individual dog.

Current evidence also confirms what was first noted over a decade ago (Taylor and Mills, 2006) and repeated more recently (Mornement et al., 2014): that canine behavior evaluations have leapfrogged from various stages of development and/or initial research studies to widespread clinical use, with profound consequences for shelters and dogs. The implications of using poorly validated instruments in human health care settings have been eloquently stated by Streiner et al. (2015, p. 14), and we believe they are equally applicable for dogs: "A scale that places a 'normal' person in the abnormal range (a false-positive result) or misses someone who actually does have some disorder (false-negative result) is not too much of an issue when the scale is being used solely within the context of research. It will lead to errors in the findings, but that is not an issue for the individual completing the scale. It is a major consideration, though, if the results of the scale are used to make a decision regarding that person—a diagnosis or admission into a program. In these circumstances, the issue of erroneous results is a major one, and should be one of the top criteria determining whether or not a scale should be used." This concern would also apply to use of owner-completed scales as an assessment tool for dogs being relinquished to shelters, such as described by Duffy et al. (2014).

Dogs and risk

As mentioned earlier, the issue of what risk of injury to people a given dog may represent figures large in the discussion of the function of behavior evaluations in shelters. While preventing serious dog bite injuries is a legitimate concern, we also believe that underlying the persistence in the design and use of behavior evaluations is a premise that most shelter dogs present an elevated danger to people, and that this must be ruled out before a dog is offered to the public for adoption. Although dangerous dogs certainly can be encountered in shelters, dogs with a history of biting or deemed too dangerous to handle at or shortly after intake are often identified and removed from the adoption pool without needing to resort to a formal behavior evaluation (see Mornement et al., 2014; Hennessy et al., 2001; van der Borg et al., 1991). These are a minority of dogs. Of course, this initial screening is not foolproof as some owners may fail to disclose problematic behavior at surrender. But to our knowledge, there is no evidence that truly dangerous dogs routinely enter the shelter adoption pool and end up exposing the community to risk.

Catastrophizing about dog bite incidence is certainly common in the human medical literature (Arluke et al., 2017), which may offer some explanation for the hypervigilance expressed about the need for behavior evaluations in the articles we examined. However, if one looks instead at the proportion of the ~86 million dogs believed to reside in human households in the United States who inflict even a medically attended bite (recognizing that many people seek medical attention after a minor incident for reasons other than injury severity), that statistic frames the risk from dogs quite differently. In the United States, it can be concluded "that >90% of dogs did not bite anyone in a given year and that <1.5% of dogs inflicted a bite for which medical assessment was sought, regardless of whether treatment was actually necessary. The proportion actually requiring medical treatment or hospitalization would be much lower" (Patronek & Bradley, 2016, p. 71). There is no compelling evidence to suggest that this latter group is so disproportionately represented in shelters that resource-intensive testing is needed for the general population of dogs that are not already screened out at intake.

We speculate that some of this belief about inherent danger of shelter dogs may also derive from claims from, or misreading of, relinquishment studies. Indeed, as mentioned previously, it seemed almost *de rigueur* that an article describing a behavior evaluation begins with a claim about behavior being the most common or very common reason for relinquishment. Nearly half (8/17) of the studies we examined cited secondary sources to support this rationale (Bennett et al., 2012; Bräm et al., 2008; Dalla Villa et al., 2017; Diesel et al., 2008a; Kroll et al., 2004; Planta and De Meester, 2007; Poulsen

et al., 2010; van der Borg et al., 1991). But closer inspection of the cited studies often indicates that relinquishment is often complicated by owner-associated factors such as moving, lack of time, etc, that make it difficult or impractical for a particular owner to accommodate the needs of a given dog (Haverbeke et al., 2015; Hemy et al., 2017; Marston et al., 2004; Salman et al., 1998). Furthermore, most of the behavioral reasons cited in these studies are normal behaviors reflecting a lack of training, not threatening behaviors or biting. There is also a potential for confusion when citing data from risk factor studies if relative versus absolute risk become conflated. For example, if a behavior is reported more frequently in dogs relinquished to shelters than in samples of owned dogs, then it may well be a strong risk factor for relinquishment. However, if that behavior is uncommon, then even a doubling of the risk may have little practical importance because it still remains an uncommon occurrence. It has also been argued that relinquishing owners may under-report behavior problems out of concern for jeopardizing the rehoming prospects of their dog and that this possibility should be factored into interpretation of these statistics (Segurson et al., 2005). However, recent high ($\geq\sim$77%) average placement rates from thousands of US shelters and rescue organizations reporting to standardized databases support the contention that the prevalence of adoption-preventing behaviors is generally low (see live release rates: https://shelteranimalscount.org/data/Explore-the-Data; https://bestfriends.org/2025-goal). These data are likely skewed in that they over-represent organizations that may have incorporated an operating ethic geared toward saving every placeable animal or that do not admit certain dogs, but the large number of organizations and inclusion of open-admission shelters is supportive evidence.

The issue of under-reporting also must be assessed in light of whether owners consider some behaviors detected by screening questionnaires such as C-BARQ as actual problems affecting their ability to live with a dog. Research indicates that pet owners make a clear distinction between behaviors that are considered undesirable and those that are considered problematic (Pirrone et al., 2015), and that owners' responses do not necessarily match the assumptions of rehoming professionals (Marder et al., 2013; Mohan-Gibbons et al., 2012; van der Borg et al., 1991). As a result, it is possible that test designers and users in shelters are overestimating the relevance of certain behaviors to the clientele they are hoping to serve. For the only behavior where this has been empirically studied in depth (food guarding), there is research demonstrating that behavior a shelter deems problematic may, in fact, not be of much concern to adopters and/or be easily manageable in the home (Marder et al., 2013; Mohan-Gibbons et al., 2012). Interestingly, this lack of concern about food guarding among owners of adopted dogs was noticed in the first publication of a behavioral assessment in shelter dogs (van der Borg et al., 1991).

One survey of 784 clients of a veterinary teaching hospital (excluding dogs with chronic illness or dogs attending the behavior services) reported that behaviors such as stranger-directed aggression (21.6%) or fear (12.3%) or dog-directed aggression (23.5%) were not uncommon among their dogs (Segurson et al., 2005). The data in the study by Christensen et al. (2007) also confirm what practical experience shows—that many family pets bark or growl when someone comes to the door (a normal and sometimes desired behavior), when they see an unfamiliar dog or cat walking down the street, or a squirrel running across a phone wire and still remain successfully in their homes.

Are there practical uses of unvalidated tests?

With respect to continued use of unvalidated canine behavioral tests in shelters, some animal scientists (Mornement et al., 2014) and one national humane organization (ASPCA, 2018) have explicitly called out the consequences of using those instruments as objective evidence to make life-and-death decisions for dogs. We share these concerns, and extend them to making any important decision. It seems that in shelter situations, their utility would be limited to being part of a series of acknowledged subjective assessments that might help, in conjunction with history and observations of other behavior during routine care, inform placement decisions or possibly identify the need for mitigation strategies to help dogs better cope while in the shelter. In those circumstances, it is important for shelters to carefully consider whether the time and resources required for conducting these standardized screening tests on all dogs deemed potentially placeable after intake is worthwhile, given that time and resources could be spent in other efforts to assess behavior and safety through normal daily interaction, provide enrichment, and help rehome those dogs that may stand to benefit from more individualized attention. Nevertheless, even such use is not without peril, as the simple act of using an evaluation may convey an unconscious belief that the tests do have objective predictive meaning. There is also a slippery slope in that behaviors observed (typically by people who are not veterinary behaviorists) in the evaluation may have a tendency to turn into labels, labels to morph into quasi-diagnoses, and these quasi-diagnoses to become behavioral pathology that justifies the need for more intensive screening and/or rehabilitation that only the shelter can provide.

Another goal of using these tests has been said to be to aid in making better matches between dogs and new owners (Mornement et al., 2015; van der Borg et al., 1991; Valssechi et al., 2011) and thereby presumably preventing returns after adoption. Although this may seem intuitively compelling, as discussed previously, this goal may well rest on erroneous assumptions about risk factors for relinquishment. And indeed, we are not aware of any empirical data suggesting that it does, in fact, make a difference. There are shelters that do not use formal behavior evaluations before placing dogs for adoption, and it might be possible to empirically assess this concern by comparing adoption-return rates between those shelters and shelters that do use formal evaluations, as had already been performed on a small scale with regard to testing dogs for food-guarding behavior (Mohan-Gibbons et al., 2018).

Along these same lines are concerns over preventing the so-called "failed adoption" which, for reasons that are not entirely clear, have been ingrained in shelter culture as a terrible thing to befall a dog and a black mark on the efforts of the shelter as well. Although we completely agree that prudent steps to minimize returns are desirable, we are not aware of any data justifying this degree of self-flagellation. Indeed, if anything should be surprising, it is that most adoptions succeed so well, given the number of variables in play and the limited amount of time available to devote to any individual adoption. Typical return rates published from studies in the United States and United Kingdom seem to fluctuate around 14% (Bollen and Horowitz, 2008; Diesel et al., 2008b; Posage et al., 1998) although one Australian study of more than 4,400 dogs reported returns of only 7.22% (Marston et al., 2004) and a study from the United Kingdom indicated 36 returns/556 completed surveys (Wells and Hepper, 2000). It should also be noted, when considering these statistics, that the duration of time after adoption for which a surrendered dog would be considered returned could vary substantially.

There is no reason to expect that any matchmaking effort would be 100% successful. Although behavior is one important reason for return, it is not the only reason. Dogs do not know that the adoption has "failed," and the notion that shelters are averse to the "failed" adoption because dogs will be traumatized by such an experience is contradicted by the common practice of sending dogs to temporary foster homes for training, medical care, or

stress relief, after which it is expected the dog will be returned to the shelter for later adoption. A recent study has contributed some interesting empirical data on this issue (Patronek and Crowe, 2018, Supplementary Table 1). That study found that dogs in an open-admission shelter returned from adoption were nearly all readopted and had a statistically significant odds of length of stay $\leq$1 week, compared with first-time owner-surrendered dogs (odds ratio, 1.44 [95% confidence interval 1.22; 1.70], $P < 0.001$). This suggests that a temporary adoption experience may have functioned akin to a foster experience, giving the dog an opportunity to get away from the stress of the shelter and the adopters an opportunity to report back much more accurate information about the dog's behavior and needs when the dog was returned. This of course does not mean that some dogs may not find the experience of being sent to a foster home or being returned from adoption mentally stressful, but it does suggest that the effects of those experiences did not have an adverse effect on length of stay in shelter before a future adoption.

Of course, there are justifiable reasons for using some type of systematic process to identify lack of coping by a dog in the shelter environment. Identifying poor welfare due to stress in the kennel could then lead to mitigation strategies, such as placing the dog into an alternative environment (e.g., office or foster home). However, the goal of assessing "well being" in the shelter is sufficiently different from predicting future problematic behavior in a home that different types of evaluations would be required. Attempting to provoke previously unobserved unwanted behavior seems less appropriate than simply identifying and then reacting appropriately to actual stress-induced behavior observed on a day-to-day basis while in residence.

Conclusion

In the past, when the deficiencies of behavior evaluations have been identified, most authors have called for further research. Calls for further research are understandable when deficiencies in existing work are newly recognized, and they are justifiable when there is some reasonable expectation that successful efforts will be forthcoming. The latter would imply a consensus at minimum over which evaluation(s) and/or subtests and behaviors are the most meaningful; how the work will be funded and replicated sufficiently to support a claim of overall reliability, validity, and utility; how confirmation of that validation will occur and be supported at individual shelters; how continual quality control over time will be achieved; and what error rate false-positive results would be acceptable to shelters; and most importantly, how the problem of poor predictive ability of a positive test (false-positive errors) in the face of low prevalence will be overcome. We see the latter as insurmountable based on the likely low prevalence of adoption-preventing behaviors in shelter dogs, as the method of calculating false-positive and false-negative errors is a fundamental, species-invariant principle of evaluating diagnostic test performance.

We would argue that when looking at the cumulative results of 25+ years of publication in this field, including solid studies performed under good to ideal conditions by skilled investigators, that calls for additional research, at least for assessing aggression-related behaviors in shelter dogs, is akin to waiting for Godot. This delay only serves to dodge the need for engaging in a much-needed conversation starting with first principles, including, as Mornement et al. (2010 p. 316) noted, considering the meaning of behavior observed under "…highly artificial conditions and during a limited time." Ethically, we would argue, given the lack of scientific evidence for validity, reliability, and predictability of canine behavior evaluations for individual dogs in a field setting, a moratorium on any uses of these evaluations as the sole determinant of a dog's fate is warranted, particularly when problematic behavior on the evaluation is the only cause for concern for a dog that has otherwise acted normally in the shelter. Careful observation of a dog's daily behavior in the shelter during routine interactions is a more natural way to gauge a dog's needs.

Although the findings from this review may seem surprising, the human medical literature is replete with reports of dozens of dearly held medical practices eventually discarded once they were refuted by actual data showing they were of low-value or even harmful (Elshaug et al., 2012; Prasad et al., 2013). In fairness although, achieving this has not typically been an instantaneous process, often taking many years after the deficiencies were first noted for users to abandon them. An important discussion is also now underway in the field of forensic investigation, where government reports have noted that many common forensic techniques such as forensic dentistry, blood spatter analysis, and other types of pattern matching used to convict (and in the most extreme cases argue for the death penalty) in human crimes have never been adequately scientifically validated (National Research Council, 2009; President's Council of Advisors on Science and Technology, 2016). Concerns have also been raised about problems plaguing genetic testing for pets, which according to the authors, has proceeded despite lack of necessary validation or evidence of ability to predict health outcomes (Moses et al., 2018). Therefore, it should not be surprising that a similar situation might exist in shelter practice as well. Indeed, the American Society for the Prevention of Cruelty to Animals has now recommended that "unless aggressive behavior during an assessment is egregious, shelters should consider it valid only if corroborated in another environment" (ASPCA, 2018). However we would argue that even if used in such a fashion, it must still be recognized that the clinical importance of the behavior(s) remains subjective and should not be interpreted as a scientifically validated indicator of future behavior.

What could the future hold?

We see this as an opportunity to acknowledge what has been learned from past efforts and bring together researchers, behaviorists, shelter staff, and even pet owners and potential adopters to consider what the future might look like. Going forward, if studies of canine behavior evaluations appear in the literature, we suggest that journals require, and authors use, a standardized reporting metric where authors must be specific about what particular aspects of a behavior evaluation are being evaluated and/or claimed to be validated. This would ensure greater rigor and more accurate reporting of such studies. Guidelines exist regarding best practices for conducting and reporting validation studies (Kottner et al., 2011). A very comprehensive model for the predictive ability component is the Standards for Reporting of Diagnostic Accuracy Studies adopted by leading human clinical journals and some psychology journals (Bossuyt et al., 2015). We realize certain aspects of these may be impractical or not relevant for dogs, but at minimum, when clinimetric statistics (sensitivity, specificity, and predictive ability) are available and/or calculable, they should be included in addition to psychometric measures. No test should ever be deemed validated without explicit, published error rates. We also recommend that journals discourage colloquial use of terms that may have specific scientific meanings when discussing or reporting results from scientific studies of behavior evaluations. If terms are used in the scientific sense, the evidence should be robust enough to support the scientific meaning. Finally, to avoid introducing further confusion in the literature, statements in the Introduction or Discussion section of an article claiming or implying that other instruments have been "validated" or are "reliable"

should not be allowed unless there is compelling evidence to show that such a state of affairs does in fact exist.

Acknowledgments

The authors thank Donald Cleary of the National Canine Research Council for help with proofreading an earlier version of this article and for providing helpful comments. There was no funding source beyond the relationships with the National Canine Research Council noted in the conflict of interest statement.

Authors' contributions: The idea for this article was conceived by G.J.P. and J.B.; all authors contributed to the identification and selection of articles, writing and review, and approved this submission.

Ethical considerations

Not required.

Conflict of interest

G.J.P. is a paid consultant to Maddie's Fund and the National Canine Research Council, a subsidiary of Animal Farm Foundation. J.B. and E.A. are employees of the National Canine Research Council.

References

Arluke, A., Cleary, D., Patronek, G., Bradley, J., 2017. Defaming rover: error-based latent rhetoric in the medical literature on dog bites. J. Appl. Anim. Welf. Sci. 25, 1–13. Available at. https://www.tandfonline.com/doi/full/10.1080/10888705.2017.1387550. Accessed February 28, 2019.

ASPCA, 2018. Position statement on shelter dog behavior assessments. Available at. https://www.aspca.org/about-us/aspca-policy-and-position-statements/position-statement-shelter-dog-behavior-assessments. Accessed February 28, 2019.

Barnard, S., Siracusa, C., Reisner, I., Valsecchi, P., Serpell, J.A., 2012. Validity of model devices used to assess canine temperament in behavioral tests. Appl. Anim. Behav. Sci. 138, 79–87.

Barrera, G., Jakovcevic, A., Elgier, A.M., Mustaca, A., Bentosela, M., 2010. Responses of shelter and pet dogs to an unknown human. J. Vet. Behav.: Clin. Appl. Res. 5, 339–344.

Beaver, B.V., 2009. Canine Behavior: Insights and Answers, 2nd ed. Elsevier Health Sciences, St. Louis, MI.

Bennett, S.L., Litster, A., Wenig, H.-Y., Walker, S.L., Luescher, A.U., 2012. Investigating behavior assessment instruments to predict aggression in dogs. Appl. Anim. Behav. Sci. 141, 139–148.

Bennett, S.L., Weng, H.-Y., Walker, S.L., Placer, M., Litster, A., 2015. Comparison of SAFER behavior assessment results in shelter dogs at intake and after a 3-day acclimation period. J. Appl. Anim. Welf. Sci. 18, 153–168.

Bollen, K.S., Horowitz, J., 2008. Behavioral evaluation and demographic information in the assessment of aggressiveness in shelter dogs. Appl. Anim. Behav. Sci. 112, 120–135.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C., Kressel, H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., Cohen, J.F., STARD Group, 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin. Chem. 61, 1446–1452. Available at. http://clinchem.aaccjnls.org/content/61/12/1446.long. Accessed February 28, 2019.

Bräm, M., Doherr, M., Mills, D., Lehmann, D., Steiger, A., 2008. Evaluating aggressive behavior in dogs: a comparison of 3 tests. J. Vet. Behav.: Clin. Appl. Res. 3, 152–160.

Christensen, E., Scarlett, J., Campagna, M., Houpt, K.A., 2007. Aggressive behavior in adopted dogs that passed a temperament test. Appl. Anim. Behav. Sci. 106, 85–95.

D'Arpino, S., Dowling-Guyer, S., Shabelansky, A., Marder, A.R., Patronek, G.J., 2012. The use and perception of canine behavioral assessments in sheltering organizations. In: Proceedings of the American College of Veterinary Behaviorists/American Veterinary Society of Animal Behavior Veterinary Behavior Symposium, San Diego, CA, pp. 27–30.

Dalla Villa, P., Barnard, S., Di Nardo, A., Iannetti, L., Vulpiani, M.P., Trentini, R., Serpell, J.A., Siracusa, C., 2017. Validation of the socially acceptable behaviour (SAB) test in a central-Italy pet dog population. Vet. Ital. 53, 61–70. Available at. http://www.izs.it/vet_italiana/2017/53_1/61.htm. Accessed February 28, 2019.

Department of Health and Human Services, 2009. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. Available at. https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf. Accessed February 28, 2019.

De Palma, C., Viggiano, E., Barillari, E., Palme, R., Dufour, A.B., Fantini, C., Natoli, E., 2005. Evaluating the temperament in shelter dogs. Behaviorism 142, 1307–1328.

De Meester, R.H., De Bacquer, D., Peremans, K., Vermiere, S., Planta, D.J., Coopman, F., Audenaert, K., 2008. A preliminary study on the use of the socially acceptable behavior test as a test for shyness/confidence in the temperament of dogs. J. Vet. Behav. Clin. Appl. Res 3, 161–170.

Diesel, G., Brodbelt, D., Pfeiffer, D.U., 2008a. Reliability of assessment of dogs' behavioral responses by staff working at a welfare charity in the UK. Appl. Anim. Behav. Sci. 115, 171–181.

Diesel, G., Pfeiffer, D.U., Brodbelt, D., 2008b. Factors affecting the success of rehoming dogs in the UK during 2005. Prev. Vet. Med. 84, 228–241.

Duffy, D.L., Kruger, K.A., Serpell, J.A., 2014. Evaluation of a behavioral assessment tool for dogs relinquished to shelters. Prev. Vet. Med. 117, 601–609.

Elshaug, A.G., Watt, A.M., Mundy, L., Willis, C.D., 2012. Over 150 potentially low-value health care practices: an Australian study. Med. J. Aust. 197, 556–560. Available at. https://www.mja.com.au/journal/2012/197/10/over-150-potentially-low-value-health-care-practices-australian-study. Accessed February 28, 2019.

Gácsi, M., Topál, J., Miklósi, Á., Dóka, A., Csányi, V., 2001. Attachment behavior of adult dogs (*Canis familiaris*) living at rescue centers: forming new bonds. J. Comp. Psychol. 115, 423–431.

Giavarina, D., 2015. Understanding Bland Altman analysis. Biochem. Med. (Zagreb) 25, 141–151. Available at. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470095/. Accessed February 28, 2019.

Guy, N.C., Luescher, U.A., Dohoo, S.E., Spangler, E., Miller, J.B., Dohoo, I.R., Bate, L.A., 2001. Demographic and aggressive characteristics of dogs in a general veterinary caseload. Appl. Anim. Behav. Sci. 74, 15–28.

Haverbeke, A., Pluijmakers, J., Diederich, C., 2015. Behavioral evaluations of shelter dogs: literature review, perspectives, and follow-up within the European member states's legislation with emphasis on the Belgian situation. J. Vet. Behav.: Clin. Appl. Res. 10, 5–11.

Hemy, M., Rand, J., Morton, J., Paterson, M., 2017. Characteristics and outcomes of dogs admitted into Queensland RSPCA shelters. Animals (Basel) 7, 67. Available at. http://www.mdpi.com/2076-2615/7/9/67. Accessed February 28, 2019.

Hennessy, M.B., Voith, V.L., Mazzei, S.J., Buttram, J., Miller, D.D., Linden, F., 2001. Behaviour and cortisol levels of dogs in a public shelter, and an exploration of the ability of these measures to predict problem behavior after adoption. Appl. Anim. Behav. Sci. 73, 217–233.

Hinkle, D.E., Wiersma, W., Jurs, S.G., 2003. Applied Statistics for the Behavioral Sciences, 5th ed. Houghton Mifflin, Boston.

Kis, A., Klausz, B., Persa, E., Miklósi, Á., Gácsi, M., 2014. Timing and presence of an attachment person affect sensitivity of aggression tests in shelter dogs. Vet. Rec. 174, 196.

Klausz, B., Kis, A., Persa, E., Miklósi, Á., Gácsi, M., 2014. A quick assessment tool for human-directed aggression in pet dogs. Aggress. Behav. 40, 178–188.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M., Streiner, D.L., 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J. Clin. Epidemiol. 64, 96–106.

Kroll, T.L., Houpt, K.A., Erb, H.N., 2004. The use of novel stimuli as indicators of aggressive behavior in dogs. J. Am. Anim. Hosp. Assoc. 40, 13–19.

Lockwood, R., 2016. Ethology, ecology and epidemiology of canine aggression. In: Serpell, J. (Ed.), The Domestic Dog: Its Evolution, Behavior, and Interactions with People, 2nd ed. Cambridge University Press, Cambridge, pp. 160–181.

Marder, A.R., Shabelansky, A., Patronek, G.J., Dowling-Guyer, S., Segurson D'Arpino, S., 2013. Food-related aggression in shelter dogs: a comparison of behavior identified by a behavior evaluation in the shelter and owner reports after adoption. Appl. Anim. Behav. Sci. 148, 150–156.

Marston, L.C., Bennett, P.C., Coleman, G.J., 2004. What happens to shelter dogs? An analysis of data for 1 year from three Australian shelters. J. Appl. Anim. Welf. Sci. 7, 27–47.

Mills, D.S., 2017. Dog bites and aggressive behavior - key underpinning principles for their scientific study. In: Mills, D.S., Westgarth, C. (Eds.), Dog Bites: A Multidisciplinary Perspective. 5M Publishing Ltd., Sheffield, pp. 11–24.

Mohan-Gibbons, H., Weiss, E., Slater, M., 2012. Preliminary investigation of food guarding behavior in shelter dogs in the United States. Animals (Basel) 2, 331–346. Available at. http://www.mdpi.com/2076-2615/2/3/331. Accessed February 28, 2019.

Mohan-Gibbons, H., Dolan, E.D., Reid, P., Slater, M.R., Mulligan, H., Weiss, E., 2018. The impact of excluding food guarding from a standardized behavioral canine assessment in animal shelters. Animals (Basel) 8, 27. Available at. http://www.mdpi.com/2076-2615/8/2/27. Accessed February 28, 2019.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63, 737–745.

Mornement, K.M., Coleman, G.J., Toukhsati, S., Bennett, P.C., 2010. A review of behavioral assessment protocols used by Australian animal shelters to determine the adoption suitability of dogs. J. Appl. Anim. Welf. Sci. 13, 314–329.

Mornement, K.M., Toukhsati, S., Coleman, G.J., Bennett, P., 2014. Development of the behavioural assessment for re-homing K9's (B.A.R.K.) protocol. Appl. Anim. Behav. Sci. 151, 75–83.

Mornement, K.M., Toukhsati, S., Coleman, G.J., Bennett, P., 2015. Evaluation of the predictive validity of the behavioural assessment for re-homing K9's (B.A.R.K.) protocol and owner satisfaction with adopted dogs. Appl. Anim. Behav. Sci. 151, 75–83.

Moses, L., Niemi, S., Karlsson, E., 2018. Pet genomics medicine runs wild. Nature 559 (7715), 470–472. Available at. https://www.nature.com/articles/d41586-018-05771-0. Accessed February 28, 2019.
Mukaka, M.M., 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. 24, 69–71. Available at. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/. Accessed February 28, 2019.
National Research Council, 2009. Committee on identifying the needs of the forensic science community. In: Strengthening Forensic Science in the United States: A Path Forward. National Academies Press, Washington, DC. Available at. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf. Accessed February 28, 2019.
Netto, W.J., Planta, D.J.U., 1997. Behavioural testing for aggression in the domestic dog. Appl. Anim. Behav. Sci. 52, 243–263.
Patronek, G.J., Bradley, J., 2016. No better than flipping a coin: Reconsidering canine behavior evaluations in animal shelters. J. Vet. Behav.: Clin. Appl. Res. 15, 66–77. Available at. http://www.journalvetbehavior.com/article/S1558-7878(16)30069-7/fulltext. Accessed February 28, 2019.
Patronek, G.J., Crowe, A., 2018. Factors associated with high live release for dogs at a large, open-admission municipal shelter. Animals (Basel) 8, 45. Available at. http://www.mdpi.com/2076-2615/8/4/45. Accessed February 28, 2019.
Pirrone, F., Pierantoni, L., Mazzola, S.M., Vigo, D., Albertini, M., 2015. Owner and animal factors predict the incidence of, and owner reaction towards, problem behaviors in companion dogs. J. Vet. Behav.: Clin. Appl. Res. 10, 295–301.
Planta, J.U.D., De Meester, R.H.W.M., 2007. Validity of the Socially Acceptable Behavior (SAB) test as a measure of aggression in dogs towards non-familiar humans. Vlaams Diergen Tijds 76, 359–368.
Posage, J.M., Bartlett, P.C., Thomas, D.K., 1998. Determining factors for successful adoption of dogs from an animal shelter. J. Am. Vet. Med. Assoc. 213, 478–482.
Poulsen, A.H., Lisle, A.T., Phillips, C.J.C., 2010. An evaluation of a behavior assessment to determine the suitability of shelter dogs for rehoming. Vet. Med. Int. 2010, 523781. Available at. https://www.hindawi.com/journals/vmi/2010/523781/. Accessed February 28, 2019.
Prasad, V., Vandross, A., Toomey, C., Cheung, M., Rho, J., Quinn, S., Chacko, S.J., Borkar, D., Gall, V., Selvaraj, S., Ho, N., Cifu, A., 2013. A decade of reversal: an analysis of 146 contradicted medical practices. Mayo. Clin. Proc. 88, 790–798. Available at. http://www.mayoclinicproceedings.org/article/S0025-6196(13)00405-9/fulltext. Accessed February 28, 2019.
President's Council of Advisors on Science and Technology, 2016. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Available at. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. Accessed February 28, 2019.
Salman, M.D., New, J.G., Scarlett, J.M., Kass, P.H., Ruch-Gallie, R., Hetts, S., 1998. Human and animal factors related to relinquishment of dogs and cats in 12 selected animal shelters in the United States. J. Appl. Anim. Welf. Sci. 1, 207–226.
Scott, J.P., 1966. Agonistic behavior of mice and rats: a review. Am. Zool. 6, 683–701.
Segurson, S.A., Serpell, J.A., Hart, B.L., 2005. Evaluation of a behavioral assessment questionnaire for use in the characterization of behavioral problems of dogs relinquished to animal shelters. J. Am. Vet. Med. Assoc. 227, 1755–1761.
Shabelansky, A., Dowling-Guyer, S., Quist, H., D'Arpino, S.S., McCobb, E., 2015. Consistency of shelter dogs' behavior toward a fake versus real stimulus dog during a behavior evaluation. Appl. Anim. Behav. Sci. 163, 158–166.
Streiner, D.L., Norman, G.R., Cairney, J., 2015. Health Measurement Scales: A Practical Guide to Their Development and Use, 5th ed. Oxford University Press, New York, NY.
Svartberg, K., 2005. A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Appl. Anim. Behav. Sci. 91, 103–128.
Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav: Clin. Appl. Res. 1, 94–108.
Valsecchi, P., Barnard, S., Stefanini, C., Normando, S., 2011. Temperament test for rehomed dogs validated through direct behavioral observation in shelter and home environment. J. Vet. Behav.: Clin. Appl. Res. 6, 161–177.
van der Borg, J.A.M., Netto, W.J., Planta, D.J.U., 1991. Behavioural testing dogs in animal shelters to predict problem behaviour. Appl. Anim. Behav. Sci. 32, 237–251.
van der Borg, J.A.M., Beerda, B., Ooms, M., Silveira de Souza, A., van Hagen, M., Kemp, B., 2010. Evaluation of behaviour testing for human directed aggression in dogs. Appl. Anim. Behav. Sci. 128, 78–90.
van der Borg, J.A.M., Graat, E.A.M., Beerda, B., 2017. Behavioural testing based breeding policy reduces the prevalence of fear and aggression related behaviour in Rottweilers. Appl. Anim. Behav. Sci. 195, 80–86.
Voith, V.L., Borchelt, P.L., 1996. Aggression in dogs and cats. In: Voith, V.L., Borchelt, P.L. (Eds.), Readings in Companion Animal Behavior. Veterinary Learning Systems, Trenton, NJ, pp. 217–229.
Wasserstein, R.L., Schirm, N.A., Lazar, N.A., 2019. Moving to a World Beyond "$p < 0.05$". Am. Stat 73 (sup1), 1.
Wells, D.L., Hepper, P.G., 2000. Prevalence of behavior problems reported by owners of dogs purchased from an animal rescue shelter. Appl. Anim. Behav. Sci. 69, 55–65.
Williams, V.S.L., Smith, M.Y., Fehnel, S.E., 2006. The validity and utility of the BPI interference measures for evaluating the impact of osteoarthritic pain. J. Pain Symptom Manage. 31, 48–57.

Glossary of terms (in the context of canine behavior evaluations)

Absolute risk: The probability of a behavior such as a bite occurring among the population of dogs being studied.
Battery tests: A series of situations or stimuli, some of which may be provocative in nature, intended to elicit particular behavior(s) in dogs if those behaviors are present.
Clinical significance: A subjective assessment of how important the result is for dogs in the real world.
Construct validity: How well an evaluation actually measures the trait or behavior it claims to be measuring. This can include subcategories such as criterion, convergent, discriminative, and predictive validity.
Content validity: How well an evaluation assesses all relevant facets of a behavior or trait.
Convergent validity: How well the results of the evaluation correlate with another measure presumably assessing the same behavior.
Correlation: How strongly two measures tend to change (increase or decrease) together.
Criterion validity: Correlates with another measure of the same behavior, preferably one that is a gold standard.
Discriminative validity: How well the results of the evaluation discriminates between dogs that have a behavior or trait and dogs that do not.
Face validity: A common-sense assessment of whether an evaluation or subtest appears to be assessing the behavior or trait of interest. Note: What is being measured must be clearly indicated.
False-positive errors: The proportion of dogs identified by the evaluation as having a behavior or trait when they do not.
False-negative errors: The proportion of dogs identified by the evaluation as not having the behavior or trait when they do have it.
Incidence: The number of new cases of a condition or behavior.
Intra-rater reliability: Level of agreement when the same rater assesses the same dog at different times.
Inter-rater reliability: Level of agreement when different raters assess the same dog at the same time.
Overall validity: A characteristic indicating that a behavior evaluation has been scientifically established to be suitable for routine use across shelters to make predictive decisions about the likely future behavior of a dog after adoption which are sufficiently accurate to ensure public safety and dog welfare.
Predictive ability: Demonstrating that the rate of false-negative and false-positive results on a behavior evaluation in both research and real-life settings is acceptable.
Predictive validity: Demonstrating that a result of a behavior evaluation is of acceptable strength and statistically significantly associated with other objective or validated measures of behavior.
Prevalence: The proportion of dogs that have a specific behavior or trait at a particular point in time. How common is the behavior or characteristic? Note: The positive and negative predictive value of any diagnostic test is heavily dependent on the prevalence of the behavior in the population tested.
Predictive value of positive tests: The proportion of dogs that test positive on an evaluation that actually have the problematic behavior.
Predictive value of a negative test: The proportion of dogs that test negative on an evaluation that are actually free of the problematic behavior.
Relative risk: The probability of a behavior such as a bite occurring in a population of dogs with a certain characteristic relative to the probability of a bite occurring in a population of dogs without that characteristic.
Reliability: Reproducibility of the evaluation as established by inter-rater, intra-rater, test-retest reliability, and inter-shelter reliability.
Sensitivity: The ability of the test to correctly identify dogs that have the problematic behavior; it is calculated as the proportion of dogs with the behavior that test positive.
Specificity: The ability of the test to correctly identify dogs that do not have the problematic behavior; it is calculated as the proportion of dogs without the behavior that test negative.
Statistical significance: A mathematical assessment of the probability that the result reflects what is truly happening in a population of dogs versus the result being due to having used an atypical sample of dogs that are not representative of the dog population.
Test-retest reliability: The ability of the test to obtain the same result when repeated on the same dog.