What makes a good test of canine behaviour: Lessons from the behavioural sciences

Presenter: Pauleen Bennett, Director, Anthrozoology Research Group Animal Welfare Science Centre, Monash University

Email: pauleen.bennett@med.monash.edu.au

Abstract

A core responsibility for local government in Australia is to increase community safety through the removal or control of dogs likely to represent a threat to the health of humans or other animals, while simultaneously facilitating responsible ownership of the vast majority of dogs, who represent no such threat but confer on the community a great number of benefits. Dangerous dogs are not currently identifiable by breed, type, size or any other measurable physical or physiological characteristic. Assessors therefore typically rely on observations of canine behaviour to make inferences about the temperament or personality of a particular dog, and to predict how that dog is likely to react in various situations. This is made difficult by the fact that dogs which are dangerous in one situation may be entirely safe and well-behaved in another. Similar problems have challenged behavioural scientists, studying animal and human behaviour in numerous different settings, for decades. The result is that there are well-established principles that must be adopted when developing, evaluating, administering and interpreting behavioural tests. The objective in this presentation is to clarify exactly what it is we should be trying to measure in potentially dangerous dogs, and then to provide a simple overview of three guiding principles that guide how behavioural tests can be used and interpreted – reliability, validity and feasibility. At the completion of the presentation, those attending will have a better idea of how to evaluate existing tests of canine behaviour.

Why we need to identify dangerous dogs

There are two very good reasons to identify dangerous dogs. The first is that some dogs are dangerous. The second is that most are not. That some dogs are dangerous is self-evident. Newspaper stories of toddlers being mauled or killed by dogs are horrific. Governments throughout the world have reacted by introducing regulatory measures designed to protect community members. The effectiveness of these measures depends on being able to identify those dogs at risk of causing harm.

That most dogs are not dangerous is also self-evident, although newspaper stories tend to obscure this fact. Bradley (2005) compared dog bite statistics with other forms of injury in the USA. She concluded that fatal dog bites are 'fantastically rare', with approximately 16 reported each year. This compares with 82 fatalities caused annually by lightening strikes and 68 caused by forklifts. These figures are remarkable given that over 77 million dogs reside in the USA (American Pet Products Association 2009). If forklifts and electrical storms were equally prevalent, fatality rates would skyrocket! Even for toddlers, acknowledged to be at greatest risk of dog bite injury, the risks are tiny. The average number of children (<10 years) killed by dogs each year in the USA is 10. This compares with 11 attributed to balloons, 15 to playground accidents, 22 to buckets and 826 to family members and friends (Bradley 2005).

Recent statistics for Victoria, Australia, reflect a similar picture, with an average of one death per year for the period from 2005-2007 (Cassell & Ashby 2009). Of some concern, however, is the fact that hospitalisation for dog-bite injury may be increasing in frequency; according to Cassell and Ashby, 8.3 Victorians per 100,000 were admitted to hospital for dog bite injury in 1999. The corresponding figure in 2007 was 9.5 / 100,000. Despite these figures most Australian dog owners are very satisfied with their dog's behaviour. Few report experiencing behavioural problems (Bennett & Rohlf 2007). The need to protect most dogs from the threat posed by those who would label all dogs as dangerous, by identifying those few which do constitute a significant risk, is substantial.

Why it is difficult to identify dangerous dogs

Having established that identifying dangerous dogs is an important undertaking, it is appropriate to consider how this might be accomplished. In some respects this is a standard risk management problem. We can therefore employ an actuarial approach – clearly define the risk, identify risk factors, rate target dogs on each factor and compare the score they obtain against some kind of agreed upon standard.

The problem with this approach is that the concept of dangerous is poorly defined. In addition, relevant risk factors have not been identified with any degree of certainty. Addressing these two issues is an essential first step. As an analogy, imagine that you are asked by your employer to investigate whether a local baboon is malcron. You know very little about baboons and have never heard the word 'malcron', so the first thing you need to do is find out what it means. If malcron is something concrete, like weight, your task will be relatively easy. Everyone you consult will give a similar definition and agree on how malcron should be measured. All you have to do is access an appropriate tool, measure the trait in your target baboon, and report back to your employer.

On the other hand, what if malcron turns out to be more complex? What if you consult ten experts and they all give different definitions? What if malcron is not something that can be measured directly, but has to be inferred on the basis of something you can measure? What if there is widespread disagreement about exactly what you should be measuring and also about whether those measurements are accurate? And what if malcron is something that comes and goes – such that a baboon might show malcron on one occasion, but not on another? Worse still, what if your own behaviour effects what you are trying to measure?

In this situation you might initially be tempted to give up. But what if you believe passionately that identifying malcron is important? And what if, when you discuss this situation with your psychologist wife, she reminds you that this is just the type of situation she's been complaining about for years. In her work, she is required to measure the intelligence of human clients. Intelligence, like the hypothetical malcron, is not something that is easily defined. In fact, despite decades of research and debate, experts continue to define it in very different ways (reviewed in Sternberg 1994). One popular way of conceptualizing human intelligence is that it exists as a single, general ability. Another recognizes two distinct types of intelligence. Still another proposes multiple intelligences, adding to the basic dichotomy a number of other dimensions, such as musical intelligence and interpersonal intelligence.

The definition of intelligence adopted has profound implications. To make things even more complicated, psychologists have not yet found a way to measure it directly. Instead they infer intelligence from their clients' behaviour during highly structured assessments which require each person to undertake a range of different tasks (Groth-Marnat 2003). The most popular test of intelligence comprises 13 subtests. The same person, therefore, may be rated as highly intelligent in some areas but below average in others. Generally, test scores are aggregated to infer how intelligent the client is relative to other people. They are also used to predict how a person will fare at school and in different employment contexts.

This works well enough most of the time, largely because psychologists have access to decades of data against which to compare test scores and make predictions. It is possible, however, for serious mistakes to be made. A person may perform badly because he or she does not understand the questions, is deaf, or is not motivated to perform well. Many skills are not examined at all. A child perceived to be of low intelligence in a standard examination may excel in music classes, exhibiting skills overlooked by existing tests. Clearly, then, the results obtained partially reflect the skills of the test administrator – his or her competency with the test, ability to motivate the client and sensitivity to other factors that may affect test scores. To counter these issues every psychologist in Australia is required to use exactly the same format to assess intelligence. It takes years of training in appropriately accredited courses to learn how to administer, score and interpret an intelligence test correctly.

How is this relevant to the issue of identifying dangerous dogs? The point is that dangerousness is similar to intelligence in that there is no universal agreement about how the term should be defined. There is also no agreement about how it should be measured. While it is relatively easy to ban dogs of certain breeds on the basis that they 'might' be more aggressive and, hence, more dangerous, than other dogs, existing evidence suggests that this approach is both conceptually flawed (Delise 2007) and ineffective (Rosado et al 2007). Conceptually, it makes more sense to identify risk factors associated with dog-bite injuries, so that an evidence-based approach can be developed. Many people assume that these may include the size of a dog and its breed or breed-type, but additional factors might include, among other things, a dog's tendency to react in a potentially dangerous way when it is frightened or frustrated, the environment in which a dog is kept, its socialisation and training history, the management skills and motivation of a dog's owner and, in some cases, the behaviour of a potential victim. Again, however, research unequivocally linking these factors with dog bites is lacking.

In the absence of this critical information, the most sensible approach to protecting the community from dangerous dogs is to assess the behaviour of individual dogs in order to predict how they might react in future situations. Doing so, however, requires consistency in how the term is defined. It also requires consistency in the test used to measure dangerousness and in how this test is administered, scored and interpreted.

Developing a definition of dangerous

Defining what is meant by the term dangerous is a difficult undertaking. While serious dog bite injuries are 'fantastically rare', less serious injuries are common (Bradley 2008). All dogs possess the weaponry required to cause serious harm to humans. Most people run foul of this at some point. Donaldson (2008) reports that anybody living to the age of sixty is likely to be bitten at least once. Fortunately, nearly all dog bites fall into the lowest injury-severity rating possible – quick recovery and no lasting impairment. To put this in context, practically all of us, especially children, sustain similar 'band-aid' level injuries on a weekly basis from falling off bikes and playground equipment, bumping into furniture and arguing with friends.

The huge difference between the very high rate of minor incidences and the very low rate of major incidences means that we cannot simply label as dangerous any dog with the potential to cause harm to a human or, indeed, any dog that has caused such harm. As an analogy, try to imagine what would happen if we identified drivers likely to cause fatal car accidents on the basis of whether they ever bumped another car while reverse parking. Cancelling the licence of everyone involved in a minor bingle would, in principle, reduce the rate of fatal accidents, but of course this strategy would lead to unacceptable social consequences. Similarly, we could reduce the number of serious dog bites by identifying as dangerous every dog that ever nips or bumps a human. Or we could eliminate those dogs that, in some kind of assessment session, can be incited to nip or rush or bark, or jump up on the assessor. The resulting absence of dogs from our community would undoubtedly please some people but would not be acceptable to most. Our task in defining dangerousness, therefore, is to develop a definition that encompasses those animals which present a level of risk higher than what we, as a community, deem to be acceptable, while simultaneously excluding those that fall below this threshold. This is no easy task and is yet to be accomplished.

Assessing the quality of behavioural tests

Once we are clearer about what it is we are trying to measure in potentially dangerous dogs, the pressing task becomes to develop a means of measuring this property. This requires a well developed assessment protocol, able to be administered and scored objectively and interpreted in a meaningful way. A wide variety of protocols have been used to assess dog behaviour. These vary from those developed informally for personal use through to those developed formally, often by scientists working with dog behaviour experts. Informal procedures can be very effective. Personal opinions about individual dogs, however, are not sufficient to justify decisions about their management, particularly in a legal sense. More formal instruments are required.

The strengths and weaknesses of existing assessment protocols is reviewed elsewhere (Mornement et al 2009), as are attempts to characterise and measure the stability of dog personality traits (Ley & Bennett 2009) and to develop a standardised test of dog behaviour (King et al 2009). To evaluate the merits of these attempts it is necessary to consider what features are desirable in a test of canine behaviour. Here the field of human intelligence testing can again provide powerful insights. Although intelligence is poorly defined and impossible to measure precisely, psychologists are able to measure it well enough to make reasonably valid predictions about individual humans (Groth-Marnat 2003). They do this by using tests that are reliable, valid and feasible, three characteristics desirable in any test of behaviour (Martin & Bateson 2007).

Reliability

For any test to be of value it must be reliable. This means that any given person or animal taking the test should receive the same score if it is administered on more than one occasion. If a dog receives a different score on a test of dangerous each time it is administered, we would say either 1) that the property of dangerousness is inherently unstable and therefore cannot be reliably measured using any test, 2) that the test is unreliable and of little use in detecting dangerousness, or 3) that the person using the test has introduced some kind of error bias. Any one of these explanations could be correct.

Decades of research have established that all biological organisms have relatively stable predispositions to behave in certain ways. Behaviour, however, does tend to be less reliable than many other properties. This is particularly evident in relation to dangerousness. Forensic psychologists, charged with the task of assessing the potential for human criminals to re-offend, acknowledge this and restrict their predictions to very specific circumstances (Cohen 1997). Unless we can assume that dangerousness in dogs is much more stable than it is in humans, measuring a dog's responses in one situation in order to predict its behaviour in another is largely a waste of time.

Even if we assume that dangerousness is a relatively stable property, measuring it reliably is likely to be difficult. The fact that it cannot be measured directly but must be inferred from the behaviour of the dog reduces the objectivity of the test, introducing a possible source of error. It is not difficult to imagine that two people observing the same dog undergoing the same test might differ in the scores they award or in how these scores are interpreted, or that the same person might award different scores to two dogs who exhibit identical responses. Imagine a scenario where every time you step on your bathroom scales they give you a markedly different result. This would be unacceptable – the scales would be dismissed as being unreliable.

Examiners are often surprised by how much their own behaviour influences test scores. Consider a study conducted in 1974 in which the experimenters observed naïve participants as they interviewed job applicants (Word et al 1974). Although the questions were identical, subtle differences were observed in the mannerisms of the interviewers depending on the race of the applicant. When participants were interviewing African-American applicants they sat further away, terminated the interview earlier and made more speech errors than when they were interviewing Caucasian participants. When the researchers copied these mannerisms as they interviewed Caucasian applicants, the applicants were rated by independent observers as having much less composure and as being much less adequate for the job. This is referred to as a self-fulfilling prophecy, where the examiner brings about the very outcome she expects by unconsciously altering her own behaviour.

Reliability, then, is an important property that can be compromised at various levels. Statistically, there are various ways in which the reliability of a behavioural test can be established. An important principle, however, is that extensive training is usually required to ensure that any test is administered, scored and interpreted as consistently as possible. We cannot help but bring to any assessment our own prejudices. These influence how we act and how we perceive the person or animal we are testing, potentially biasing any assessment we conduct and leaving our conclusions open to challenge in the legal system.

Validity

Validity is a more complex concept than reliability (Aiken 2003). Suppose that your bathroom scales give you the same reading every time you stand on them, but underestimate your weight by several kilograms. The scales are perfectly reliable, but wrong nonetheless. Similarly, imagine that you assess a dog on a test of dangerousness on several occasions and conclude on each occasion that it is a very 'safe' dog. If that dog subsequently inflicts serious harm, your test is not a valid indicator of that behaviour. Nor is it a valid test if it results in the mislabelling of dogs that are not dangerous. A poorly designed test may not be sufficient to elicit the behaviour being tested for, or it might elicit behaviour that would never be exhibited under 'normal' circumstances. To have any confidence in a behavioural test it is necessary to establish its validity; the capacity of the test to measure accurately what we think it is measuring.

Validity is not something that can be measured directly and it is difficult to establish (Aiken 2003). Tests of intelligence, for example, are deemed to be valid only when they generate scores equivalent to those obtained on similar tests (concurrent validity) or on a specific 'gold-standard' task (criterion validity), or when scores on different types of tests either converge (convergent validity) or diverge (divergent validity) as predicted. Intelligence tests are also expected to predict how a person will perform on tasks assumed to require different levels of intelligence, either in the laboratory (predictive validity) or in 'real-life' (ecological validity). Importantly, validity is not necessarily a static trait, such that a test is equally valid in all situations. A test of intelligence that is valid for children may not be valid for adults and a test that is valid in the USA may not be valid in Australia. After many years of using intelligence tests psychologists agree that current tests are valid in certain cultural groups, when administered in certain ways and in terms of predicting some outcomes, such as school performance. They are less valid when used in other circumstances (Groth-Marnat 2003). Knowing the limitations of a test is important, allowing test results to be interpreted in a defendable manner.

The type of validity of most interest in the current context is predictive validity – the ability of the test to accurately predict which dogs will inflict serious harm. Establishing the predictive validity of a test for dangerousness is particularly challenging. Ideally, it requires testing a sample of dogs prospectively so that their scores can later be compared across groups with different outcome. Dogs who obtain high scores on a test of dangerousness should later be found to have inflicted more harm than those who obtain low scores. Conducting such an experiment would clearly be unethical since it would intentionally expose humans to dogs considered to be at high risk of harming them. It would also be impractical. Because dogs inflict serious harm at such a low rate, many thousands of dogs would need to be tested to obtain even a small sample of dangerous dogs. An alternative is to carefully and comprehensively evaluate existing cases of dog bite injury as they arise. The accumulation of data about potential risk factors would greatly assist development of a valid assessment.

Feasibility

A final factor in determining the usefulness of a behavioural test is how feasible it is (Martin & Bateson 2007). The most reliable and valid way to determine if a dog is likely to harm a running, jumping, screaming toddler is simply to expose it to one on multiple occasions, making sure to vary the conditions of the test until the dog's behaviour in all possible circumstances has been established. This is clearly not feasible for a host of practical and moral reasons. What is needed are feasible ways of testing a dog's potential for inflicting harm; methods which can be administered safely and economically but which are effective.

Conclusions

The objective in this paper was to clarify what it is we should be trying to measure in potentially dangerous dogs, and then to provide a simple overview of three principles that guide how behavioural tests can be used and interpreted. As should be evident by now, there are many good reasons to identify dangerous dogs. There is also little doubt that a behavioural approach, rather than one based on the size or breed of the dog, is the preferred strategy. There is much work to be done before such an approach can be used successfully, however, since the concept of dangerousness remains poorly defined and significant measurement issues abound. At present, we have very limited knowledge about risk factors for serious dog bites. We also have very little information about the properties of available measurement instruments. If these are not reliable, valid and feasible, using them is impossible to defend.

References

Aiken, L.R. (2003). Psychological testing and assessment (11th edn). New York, USA: Allyn & Bacon.

American Pet Products Association (2009). 2009/2010 National Pet Owner's Survey. Summary statistics retrieved 4th Sept 2009, from www.americanpetproducts.org/press_industrytrends.asp

Bennett, P.C., & Rohlf, V.I. (2007). Owner-companion dog interactions: relationships between demographic variables, potentially problematic behaviours, training engagement and shared activities. *Applied Animal Behaviour Science*, 102, 65–84.

Bradley, J. (2005). Dogs bite, but slippers and balloons are more dangerous. California, USA: James and Kenneth.

Cassell, E., & Ashby, K. (2009). Unintentional dog bite injury in Victoria: 2005–7. *Hazard*, 69, 1–24. Victoria, Australia: Monash Accident Research Centre

Cohen, D.A. (1997). Notes on the clinical assessment of dangerousness in offender populations. Psychiatry On-Line (retrieved 4th sept 2009, from priory.com/psych/assessin.htm

Delise, K. (2007). The Pit Bull placebo: the media, myths and politics of canine aggression. New Jersey, USA: Anubis Publishing.

Donaldson, J. (2008). Oh behave! Dogs from Pavlov to Premack to Pinker. Washington, USA: Dogwise.

Groth-Marnat, G. (2003). Handbook of psychological assessment (4th edn). New Jersey, USA: John Wiley and Sons.

King, T., Marston, L.C., & Bennett, P.C. (2009). Development of a valid and reliable test of amicability in dogs. Australian Institute of Animal Management Annual Conference, October 14–16, Geelong, Victoria, Australia.

Ley, J., & Bennett, P.C. (2009). Assessment of personality in dogs. Australian Institute of Animal Management Annual Conference, October 14–16, Geelong, Victoria, Australia.

Martin, P., & Bateson, P. (2007). Measuring behaviour (3rd edn). New York, USA: Cambridge University Press.

Mornement, K., Toukhsati, S.R., Coleman, G.J., & Bennett, P.C. (2009). Reliability, validity and feasibility of existing tests of canine behaviour. Australian Institute of Animal Management Annual Conference, October 14–16, Geelong, Victoria, Australia

Rosado, B., García-Belenguer, S., León, M., & Palacia, J. (2007). Spanish dangerous animals act: effect on the epidemiology of dog bites. Journal of Veterinary Behavior, 2, 166–174.

Sternberg, R.J. (1994). Encyclopedia of intelligence. New York, USA: Macmillan.

Word, C.O., Zanna, M.P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. Journal of Experimental Social Psychology, 10, 109–120.

About the author

Pauleen Bennett is currently the Director, Anthrozoology Research Group, Animal Welfare Science Centre, Monash University. Pauleen trained as a Behavioural Neuroscientist and Clinical Neuropsychologist but was always interested in animals. After completing her PhD she began breeding and training dogs and became more and more interested in the complex roles that companion animals play in our lives. She was fascinated that some people are more devastated by the loss of a pet than by the loss of a family member, and that animals manage to find themselves embedded in every aspect of human community. This prompted Pauleen to change career paths, and she now teaches courses about animals in society and animal welfare in both Australia and the USA, as well as conducting research into various aspects of companion animal care and management. She hopes to have a real impact on the way that dogs and cats and other species are cared for in our community and ensure that companion animals continue to enrich the lives of community members. When she's not at work she's at home on her farm, surrounded by dogs, angora goats and alpacas. She also occasionally assists her partner Ron, who is responsible for caring for dogs and cats admitted to the local pound.