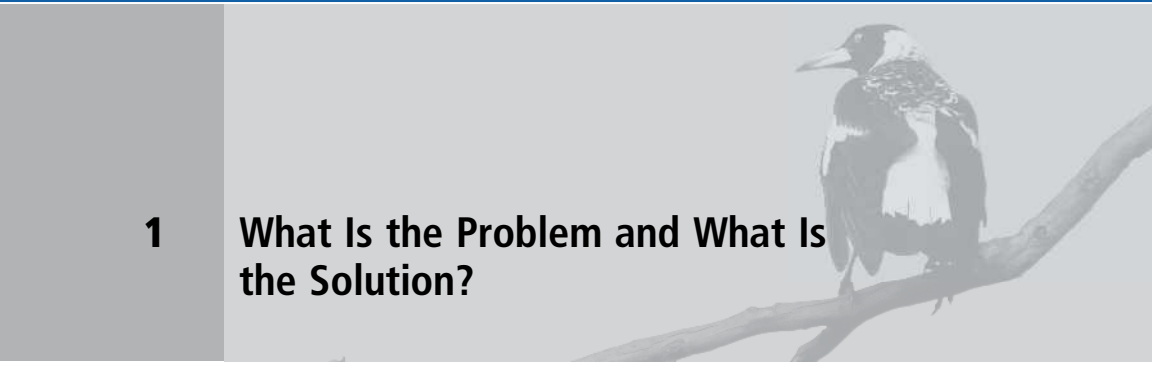

1 What Is the Problem and What Is the Solution?

> Count what is countable, measure what is measurable and what is not measurable, make measurable.
>
> Attributed to Galileo Galilei (Finkelstein, 1982, p. 25)

From its inception in the 1500s, a key element in the success of the scientific revolution has been measurement. Rendering the intangibles of nature into numerical values has allowed for precise comparisons. The coupling of accurate measurements with the statistical methods that were developed in the late nineteenth and early twentieth centuries led to the ability to test hypotheses, objectively, about the causal processes that underlie observable phenomena. In this regard, the scientific study of the behaviour of animals is no different to any other branch of modern science as perusal of any journals that involve studying animal behaviour attests (e.g., *Animal Behaviour*, *Behavioral Neuroscience*, *Behaviour*, *Behavioural Brain Research*, *Ethology*, and *Journal of Comparative Psychology*).

However, perusal of papers on topics related to animal behaviour that were published in the 1940s, 1950s, and 1960s show that, even though they became increasingly quantitative, they were highly descriptive. In contrast, papers since the 1970s have become increasingly focused on hypothesis testing – organised so as to answer why animals engage in behaviour X. The concerns that were raised by Tinbergen (1963) and Lorenz (1973) on the need for, and demise of, behavioural description have been largely ignored. This has led to many papers in the modern era providing brief and relatively arbitrary definitions of the behavioural markers to be scored for the quantitative testing of the hypothesis being proposed. The core of the rationale provided in most studies concerns the hypothesis being tested and the sampling methods and statistics used. In this book, we make the case that developing and using behavioural markers is itself a hypothesis – a hypothesis that the chosen measures are appropriate reflections of the

behavioural phenomenon being studied. Therefore, rather than being the accepted starting points of a study, the chosen behavioural markers should be subject to empirical testing, just like any other hypothesis. For several reasons, that is not routinely the case.

Amassing and analysing more data Over the past 30 years or so, modern digital and computer technologies have revolutionised not only how data are collected, but also how much can be collected and how they can be analysed. For example, for field-based studies, using handheld devices that tap into the global positioning system have been a boon in accurately tracking the movement of animals and their inter-individual spatial relationships (e.g., Tomkiewicz, Fuller, Kie & Bates, 2010). Digital video and audio recordings have become cheap and easy to use in both field and laboratory contexts. Computer-based analysis systems (e.g., *The Observer* from Noldus, *RavenPro* from Cornell University) have now been refined to the point that large quantities of data can be collected in real time. Furthermore, new computational methods for analysing large amounts of quantitative data derived from traditional statistical methods, or the more recent Bayesian approaches, have been developed (e.g., Casarrubea et al., 2018; Garamszegi et al., 2009; Kline, 2013). Combining large data sets with new computational techniques can lead to novel insights, hitherto unreachable (e.g., Anderson & Perona, 2014; Brown & de Bivort, 2018).

One unfortunate consequence of this trend, however, has been to tolerate poorer quality data, since one can always add another factor in a linear mixed model to rule out statistically the influence of some presumed confound. This is not necessarily a bad thing, particularly at the early exploratory phases of a study, when patterns of association are being sought for identifying material worthy of more detailed study – an approach to which we are not averse (e.g., Burke, Kisko, Euston & Pellis, 2018; Stark et al., 2020). Where this becomes a problem is when some broad statistical pattern becomes confused with real understanding of the biological organisation of the system.

Confounding levels of behavioural organisation Irrespective of the quantity of data collected, it needs to be borne in mind that how data are collected greatly influences how those data can be analysed and what questions can be answered (Gomez-Marin et al., 2014; Leonelli, 2019). A good example of how measuring regimes need to be tailored to the

behavioural question of interest concerns the duration of the behavioural events to be measured (Altmann, 1974). If you need to know how much time during the day animals are engaged in different activities, the duration of the bouts of the different activities greatly influences how they are best sampled. For example, to estimate the amount of time a goose spends foraging relative to scanning for predators, that is, whether it has its head down cropping grass or has its head up directing its gaze to the horizon can be recorded at some set interval (e.g., every 10 minutes). That is, at the onset of the time interval, you look at the animal and score whether it is grazing or scanning; then at the beginning of the next interval, the scoring procedure is repeated. Once sampling for the day, say from dawn to dusk, is completed, the number of intervals containing grazing and scanning can be tallied and so the proportion of the samples devoted to each activity can be calculated. Such sampling provides an estimate of how much of the day the animals spend in these activities (Pellis & Pellis, 1982). But such instantaneous scan sampling only works well for behaviours like grazing and scanning, bouts of which last for many seconds or even minutes.

For behaviour patterns like the same goose scratching its head with its hind foot or engaging in a brief courtship encounter, such a sampling technique is inadequate. Given the low frequency of occurrence (which may happen from a couple of times to a few dozen times a day) and short duration (from just a few seconds to less than a second), the chances that the behaviour is caught in the snapshot of an instant when it is the time to sample is highly unlikely. Thus, for rare and short-duration behaviours, a more suitable approach is to sample, continuously, throughout the day, recording them whenever they occur (Pellis, 1982). There are well-established guidelines for taking such factors into account for developing scoring schemes that can appropriately sample different kinds of behaviours (see Dawkins, 2007; Martin & Bateson, 2007).

Whether of short duration or long duration, what all these behaviours have in common is that they are mutually exclusive. From a practical point of view, an observer is unlikely to mistake scratching for grazing or grazing for scanning. Moreover, such behaviours have undeniable biological relevance; eating, avoiding being eaten and grooming are all essential for maintaining life. Also, because these behaviours cannot co-occur, there is no ambiguity in scoring them as independent events and in studying their sequential organisation. Engaging in aggression can be similarly fitted

into these scoring schemes and assessed for its occurrence and relative juxtaposition with the other behaviours of interest. However, a closer examination of fighting reveals a level of analysis at which the kind of numerical scoring considered above becomes less informative.

Consider a pair of animals, such as two bull elephants, fighting. The sequence of action can be analysed by determining whether behaviour A follows behaviour B more frequently than expected by chance (e.g., Clark & Moore, 1994; Donaldson et al., 2018; Lerwill & Makings, 1971). But how are behaviours A and B abstracted from the stream of behaviour observed? Most of the movements made by the two animals overlap and continually influence one another in a bidirectional manner (Geist, 1978). Yet, despite these empirical problems, most papers simply state 'my definitions of A and B are . . .,' give the heuristic criteria for how they were measured and provide no further rationale. The 'behavioural markers' that are selected for quantification are snapshots of what researchers presume reflect the underlying organisation of the behavioural phenomenon being studied. Researchers' biases in selecting behavioural markers in highly dynamic situations such as fighting are likely to have a greater influence than in selecting those from less dynamic contexts, such as reflected in scoring the scanning and grazing of geese. What are likely to be selected are behavioural actions that are readily identifiable and commonly present in the interactions, but these easy-to-score markers may not be a good reflection of the organisation of the behaviour. Many examples will be explored in the pages that follow.

Confusing agreement with biological reality Increasingly, justification for the validity of arbitrarily selected behavioural markers is how robustly they can be recognised and scored repeatedly by the same observer and by independent observers. While intra- and, especially, inter-observer reliability is an important part of characterising useful measures that can be widely used (Burghardt et al., 2012), by itself, it is an insufficient criterion with which to establish whether an abstracted behavioural marker is a valid description of the behavioural phenomenon in question.

Martin and Bateson (2007) use shooting at a bull's-eye to help conceptualise the reliability of scoring behavioural markers between observers and within the same observer. This is also a helpful metaphor with which to think about the quality of the behavioural marker being scored. The close clustering of bullet holes in Figure 1.1a would represent high inter-observer

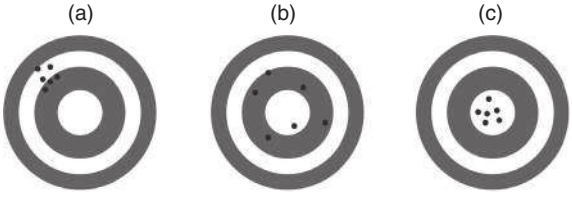

Figure 1.1 A bull's-eye is shown as a representation of how close a behavioural marker is to the underlying organisation of the behaviour with the centre being the closest. Inter-observer (or within observer) consistency is reflected in how close the bullet holes are clustered together. (a) Precision (highly reliable scoring, but in this case, off the target). (b) Accuracy (closer to the target, but in this case, with poor reliability). (c) Precision and accuracy (close to the target and highly reliable).

reliability compared to the looser clustering in Figure 1.1b. But the clustering in Figure 1.1a is further from the centre than that in Figure 1.1b. In terms of shooting a target, Figure 1.1a has higher precision (i.e., less variation among shots), but lower accuracy (i.e., further from the target) than Figure 1.1b. Simply relying on measures of inter-observer reliability would lead researchers to view Figure 1.1a as superior to Figure 1.1b. However, with regard to the biology of the system being measured, Figure 1.1b is more informative than Figure 1.1a. Of course, in the best of all possible worlds, Figure 1.1c, which has both precision and accuracy, would represent the superior measurement scheme. But more often than not, in a messy world, the actual choice is between Figure 1.1a and b, and current standards would favour Figure 1.1a because it has higher inter-observer reliability.

When the level of behavioural organisation under investigation becomes more prone to subjective judgements as to what should be measured, increasing the quantity of data collected or relying on inter-observer reliability are poor criteria for passing judgement on what is measured. What is critical is that the behavioural markers are selected because the researcher believes that they reflect something important about how the behavioural phenomenon is organised. In this regard, selection of what to measure is, in itself, a hypothesis of the underlying organisation of the behaviour, and as such, should be amenable to being tested. More often than not, the rationale for selection is not made explicit; only how to measure what is selected is

explicitly stated. But the most critical question that needs to be answered is how closely does the abstracted 'behaviour pattern' or 'behavioural marker' reflect the organisation of the behavioural phenomenon under study. Most currently available books on methods in the study of animal behaviour tend to focus on providing guidance on *how* to measure behaviour (e.g., Dawkins, 2007; Martin & Bateson, 2007), not on *what* to measure. In this book, we provide a framework to make the selection of behavioural markers explicit and so more readily subject to testing.

Some Lessons from Righting

When an adult rat is laid on its back on a flat surface, it will rotate so that it goes from supine (on its back) to prone (with all four of its paws on the ground), that is, it rights itself. Typically, the rotation to prone is cephalocaudal, starting with the head and ending with the pelvis (Magnus, 1926). Compared to adults, newly born rats are much slower in gaining the prone position, engaging in many, seemingly irrelevant movements that are not present in adults. A simple behavioural marker for assessing how quickly over the course of development animals can achieve the adult-typical form is to measure the time it takes for them to go from supine to prone. This can be done simply: take a video record of the righting and count the number of frames, starting at the frame at which the animal is released, and ending when all four of its paws contact the ground. The number of frames can then be converted to seconds. A complication is that, because righting at early ages can sometimes be very slow, or can even fail to occur altogether, researchers have often chosen some cut-off, such as ending the trial if the animal, after it has been released, has not righted by 15 or 30 seconds. Irrespective of the exact criterion for ending a righting trial, what such studies show is that, with age, animals are increasingly likely to right, and do so with increasing speed, until the timing is indistinguishable from that of adults (e.g., Almli & Fisher, 1977; Altmann & Sudarshan, 1975; Cowan, 1981; Markus & Petit, 1987). There is a practical advantage to this approach, but it comes with a biological disadvantage.

As we discovered by training students to score righting in rats, naïve observers can be quickly taught to count video frames and the scores from multiple students exhibit high inter-observer reliability. Thus, from a practical perspective, this is a 'good' measure; it is reliable across scorers,

can be taught easily, can be scored with high precision, and its simplicity allows large samples of animals to be scored rapidly. However, the cost comes with the assumption of what researchers think the measure reveals about the underlying organisation of righting and how that organisation changes with age. By using the time it takes for the animal to right to prone ('time-to-right') as the cardinal marker for the development of righting, it is explicitly or implicitly assumed that righting is a unitary phenomenon that, with maturation of the animal's sensory and motor skills, improves with age. That is, the measure is based on a particular hypothesis about how righting is organised. The problem is, what if that hypothesis is incorrect?

When we first began working in Philip Teitelbaum's laboratory on animal models of Parkinson's disease, righting became one of the behaviours on which we focused. The reason is simple: people with Parkinson's disease have impaired postural reflexes, including the ability to right themselves (Lakke, 1985). Using rats, our goal was to evaluate the behaviour of animals with damage to the neural circuits that are compromised in Parkinson's disease. Naturally, given the literature of the time, we started by using the time-to-right as the relevant behavioural marker. However, this turned out to be an unsatisfactory approach when applied to adult rats with bilateral electrolytic lesions at the level of the hypothalamus. Such lesions damage the ascending dopaminergic neurons, and the disruption of dopamine input to the basal ganglia, anterior to the hypothalamus, results in immobility and catalepsy, symptoms comparable to those of patients with Parkinson's disease. Initially, after the damage, the rats do not right themselves, but with recovery they begin to do so. The recovery involves a complex array of movements, in which there are shifting patterns in how they are integrated until the animals right normally (Pellis et al., 1989). The complexities in how the animals righted through recovery could not be captured by simply scoring the time it took them to right. Since the atypical patterns of righting present in brain-damaged rats could reflect novel compensatory manoeuvres to overcome the Parkinsonian deficiencies, we could not be sure how they related to normal righting.

Teitelbaum's earlier work showed that, both with regard to movement systems and patterns of ingestion, recovery from lateral hypothalamic damage parallels normally occurring development (Teitelbaum, Cheng, & Rozin, 1969; Teitelbaum, Wolgin, De Ryck & Marin, 1976). We therefore

turned our attention to how movements occurring during righting changed in rats during ontogeny (V. Pellis, Pellis & Teitelbaum, 1991). Starting with scoring the time the animal took to right to prone, our data concurred with that in the literature that, on average, the speed of righting increased with age. However, the range was extraordinarily large. From the day of birth, infant rats could sometimes right almost as fast as adults. On other occasions, they made repeated movements with their limbs and torso, but failed to gain the prone position. In other cases, they did manage to right to prone, but this took longer than is typical for an adult. The hypothesis that righting involves a unitary pattern of cephalocaudal rotation that improves with age due to the maturation of sensory and motor capabilities did not account for why a newborn could, on occasion, right as fast as an adult! How is that possible if improved righting merely reflects changes in sensorimotor skills?

Working mostly with adult cats, Magnus (1924, 1926) showed that, when falling supine in the air, righting is triggered by the vestibular system (the sensory organ for balance located in the inner ear), or, in the absence of vestibular signals, by vision. When righting on the ground, in addition to those two forms of righting, tactile contact on the upper body triggers righting by the forequarters and contact on the lower body triggers righting by the hindquarters. Since Magnus, another form of tactile righting has been described that involves the trigeminal nerve (a cranial nerve that projects over the face) (Troiani, Petrosini & Passani, 1981). Of the righting systems known to Magnus, he showed that in adults there is a hierarchy. When righting on the ground, irrespective of other sensory signals, vestibular ones have priority access to righting movements. If vestibular signals and vision are blocked, then tactile signals preferentially trigger forequarter righting, with tactile-induced hindquarter righting only occurring if forequarter righting is prevented.

The anomalies in using the time-to-right measure could have arisen from a complex interweaving of these different righting modalities with age. Therefore, following the pioneering work of Magnus (1924, 1926) and others (e.g., Tilney, 1933; Windle & Fish, 1932), we devised ways of testing the capability of each sensory system in triggering righting independently of the influence of other systems in the young of small mammals, including rats (Pellis, Pellis & Nelson, 1992; V. Pellis, Pellis & Teitelbaum, 1991). When doing so, it became apparent that some

unique patterns of movement are utilised by each righting system. From birth to weaning, three aspects of righting change until the fully adult pattern is present. First, there is an order of emergence of the different righting systems. Unlike cats, for rats and some of the other small mammals that we studied, vision is not capable of triggering righting at any age. In rats, while all other forms of righting are present from birth, they differ in how closely they resemble the fully adult form. Only trigeminal righting has the typically adult pattern from its first appearance. Second, at earlier ages, unlike fully mature righting, the hierarchy among the different righting systems is incomplete, leading to a simultaneous co-activation of multiple types of righting. Third, vestibular and body tactile forms of righting systems undergo a set of characteristic changes in the combination of head, body, and limb movements that are used over the first three weeks of development until their adult typical forms are consolidated.

Trigeminal righting involves cephalocaudal rotation of the body axis, starting with the head and neck, followed by the shoulders and, finally, the pelvis; this is the same order in infants as it is for adults. Furthermore, this form of righting is completed as successfully in infants as it is in adults. For successful trigeminal righting, the lower part of the infant's snout must maintain contact with the ground, so that its body rotates around that point of contact (Figure 1.2). However, the co-activation of other righting systems and the movements used in the early stages of development can produce actions in the pup that interfere with the successful completion of trigeminal righting. For example, at the onset of tactile-induced forequarter righting, the rat pup's forepaws reach for the ground and pull their forequarters to prone. The immaturity of the pup's forelimbs can lead to a failure to right, so that before the lower side of its face contacts the ground, its forelimbs may lose their grip with the ground and extend upward, away from the ground. This forelimb movement can rotate the pup's shoulders away from the ground and so pull its face away from the ground, interfering with the trigeminal input needed to complete the trigeminal form of righting.

The difficulties are compounded if tactile-induced hindquarter righting is simultaneously triggered with forequarter righting. As in the tactile-triggered forequarter righting that occurs early in development, tactile-triggered hindquarter righting involves reaching and pulling actions by the

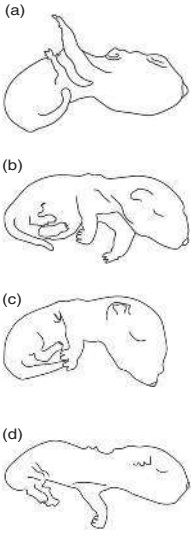

Figure 1.2 A sequence of drawings shows a young marsupial carnivore (the Northern quoll) engaging in trigeminal righting at 40 days of age when righting behaviours begin to emerge. After making side-to-side movements with its head and grasping movements with its forepaws (a), it makes contact with the ground with the anterior of its snout (b and c). Maintaining snout contact with the ground, the quoll rotates to the prone position (d). Adapted from Pellis, Pellis, and Nelson (1992) with permission (Copyright © 1992 John Wiley & Sons, Inc.)

animal's hind paws. The failure of the rat pup to grasp the ground successfully leads to an upward extension of its hind limbs and thus also contributes to the animal rotating its body away from the side of the ground to which its face is closer. Indeed, the co-activation of tactile-triggered forequarter righting and tactile-triggered hindquarter righting can lead to conflicting back-and-forth rotations of the longitudinal axis of the pup's body that is reminiscent of a 'corkscrew' and can prevent either type of righting from being successful in righting the animal's body to prone (Figure 1.3).

The early onset of vestibular righting makes matters worse for the pup, not better. Initially, the rat pup's head is thrust upwards, away from the ground, not towards the ground (Figure 1.4). This reduces the likelihood of trigeminal contact with the ground and makes a successful purchase on the