

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

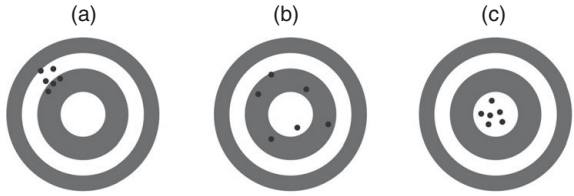


Figure 1.1 A bull's-eye is shown as a representation of how close a behavioural marker is to the underlying organisation of the behaviour with the centre being the closest. Inter-observer (or within observer) consistency is reflected in how close the bullet holes are clustered together. (a) Precision (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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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,

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

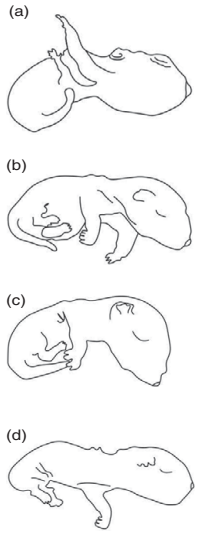


Figure 1.2 A sequence of drawings shows a young marsupial carnivore (the Northern quoll) engaging in trigeminal righting at 40 days of age when righting behaviours begin to emerge. After making side-to-side movements with its head and grasping movements with its forepaws (a), it makes contact with the 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

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

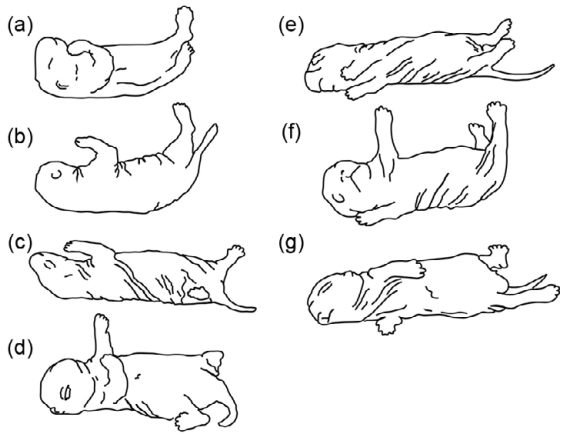


Figure 1.3 A sequence of drawings shows a two-day-old rat pup exhibiting the corkscrew pattern of righting. The pup begins by rotating its forequarters towards the ground with grasping movements of its forepaws. This rotation is accompanied by its righting movements with the hind leg closest to the ground (a and b). However, its hind foot fails to gain traction on the ground and as its leg extends across the midline, it rotates its hindquarters in the opposite direction of its head (c). The pup then rotates its forequarters to the same side as its hindquarters (d), but again, failure by its hind foot to gain traction leads to an extension of its hind leg across its mid-line (e and f). The simultaneous righting movements by the pup's fore- and hind limbs produce a conflicting rotation of its body, which produces the characteristic corkscrew configuration of the body (c and e). In this case, after several failed righting movements such as this, the pup was fatigued and lay inert (g). Adapted from V. Pellis, Pellis, and Teitelbaum (1991) with permission (Copyright © 1991 John Wiley & Sons, Inc.)

ground by the pup's forepaw more difficult. Later in development, both vestibular and tactile-triggered righting by the pup's forequarters switches to the adult-typical form, in which the animal's limbs are tucked into its body and its upper torso rotates around its longitudinal axis. These versions of vestibular and tactile-triggered forequarter righting facilitate, rather than counteract, trigeminal righting. Also, at these later stages, the hierarchy among the righting systems begins to be established, further reducing the likelihood that multiple types of righting are co-activated. As described above, co-activation of multiple righting systems can lead to the animal making movements that are not coordinated with one another,

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

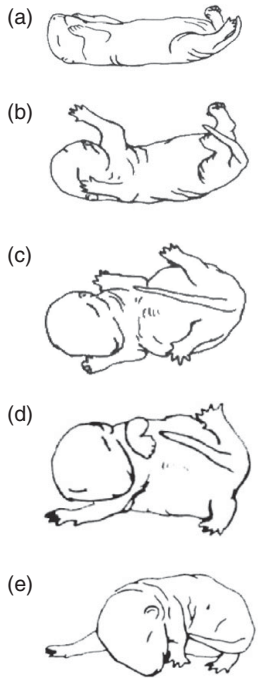


Figure 1.4 A sequence of drawings shows a rat pup on the day of birth exhibiting the U form of righting. When released on its back, the pup ventrally flexes its body (a) and then falls to one side (b). Once its forepaw makes contact with the ground (c), the pup rotates its head (d) and body to prone (e). Adapted from V. Pellis, Pellis, and Teitelbaum (1991) with permission (Copyright © 1991 John Wiley & Sons, Inc.)

leading to delayed or failed righting. Given these complex interactions among righting systems, it is no wonder that the time-to-right measurement fails to capture the changes that occur in rats in righting over development or in recovery from the effects of brain damage in adulthood (Pellis, 1996).

Finding Suitable Behavioural Markers to Measure

Thus, although the time-to-right measure is a simple one, and is quick and easy to teach and can achieve a high precision among observers

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

(Figure 1.1a), as it turns out, it is also one that poorly reflects the organisation of righting and how that organisation changes with age and during recovery from brain damage. As evidenced by the level of training needed to ensure that naïve students can acquire the level of skill needed to recognise and so score complex sequences of righting movements (e.g., Field, Watson & Pellis, 2005; Martens, Whishaw, Miklyaeva & Pellis, 1996), increased accuracy comes with costs in the efficiency of scoring (fewer animals can be scored per unit time) and in the precision of scoring across observers (Figure 1.1b). But how to gain accuracy while increasing precision as much as possible (Figure 1.1c), and, in doing so, have behavioural markers that are not onerous to teach or to use? With regard to the development of righting, this has been achieved in two ways.

First, each type of righting system can be tested separately. For example, to determine the effects of early vestibular sensory experience on the development of the vestibular righting system, the postnatal righting of rat pups was tested in a bath of warm water, in which the sole form of righting possible is that triggered by vestibular stimulation, thus avoiding the confound of triggering multiple righting systems that may be activated when the rat is placed in contact with the ground (V. Pellis, Pellis & Teitelbaum, 1991). Using this more specific testing paradigm, it was shown that, after housing gravid female rats in microgravity (in space flight), their young, when born, were delayed in the onset of vestibular righting, but not in the tactile forms of righting (Ronca & Alberts, 2000). Second, movements that are highly correlated with a particular stage of development can be scored to determine if the rate of maturation is affected by some treatment. For example, in one scheme, body postures typically present at different ages during contact righting (V. Pellis, Pellis & Teitelbaum, 1991) are used to evaluate the rate of maturation of righting. Unidirectional cephalocaudal rotation in which the head is kept away from contact on the ground (and so not involving trigeminal righting) can be used as a marker for the mature, adult-typical form of righting on the ground (Figure 1.5a, axial rotation). An immature stage of righting, in which there is co-activation of forequarter and hindquarter tactile-triggered righting can be readily recognised by the opposing directions of rotation by the fore- and hindquarters (Figure 1.5b, corkscrew). An even earlier appearing immature stage of righting triggered by the vestibular system can be recognised by the initial ventroflexion of the whole body (Figure 1.5c, U-posture). Two studies show

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

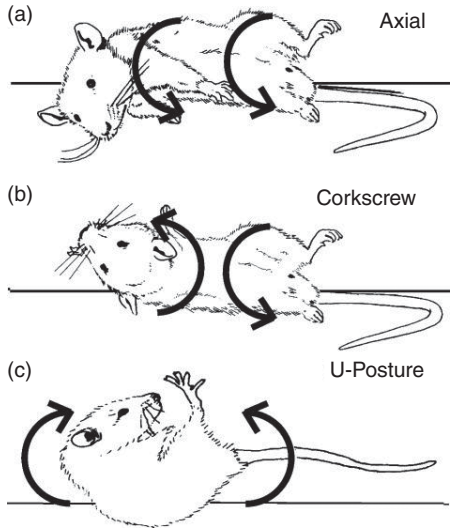


Figure 1.5 Three typical postures associated with different stages of maturation of contact righting are shown. (a) Cephalocaudal axial rotation as is typical of adult vestibular or forequarter tactile-triggered righting. (b) Counterproductive opposing axial rotation as forequarter and hindquarter tactile-triggered righting are co-activated. (c) The initial whole body ventroflexion typical of vestibular-triggered righting typical of the early stages of onset of vestibular righting. Reprinted from Walton et al. (2005) with permission (Copyright © 2005 John Wiley & Sons, Inc.)

how using this simple scoring scheme can reveal the effects of experimental manipulations on the development of righting.

In one study, also involving microgravity, infant rats were exposed to space flight for up to 14 days after birth. Complete, unidirectional cephalocaudal rotation of the body axis reflects the mature form of righting not only of trigeminal righting, but also of vestibular and body tactile forequarter righting, and all these forms are highly likely to occur in their fully mature form during the third week of life (V. Pellis, Pellis & Teitelbaum, 1991). The difference is that trigeminal righting has a relatively mature pattern on first emergence, so unidirectional axial rotation involving snout contact with the ground does not necessarily reflect a mature developmental state (Figure 1.2). However, unidirectional axial rotation in which the head is not in contact with the ground does (Figure 1.5a), and so can be

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

used as a marker for developmental maturity. The rats that were raised in microgravity were less likely to use this pattern than the controls and the difference was permanent in those exposed to microgravity the longest (Walton et al., 2005). Similarly, in the first few days after birth, when rat pups are placed on their backs, vestibular input not only stimulates the head to be thrust upwards, but also the whole body bends ventrally, so that the rat pup adopts a 'U' posture (Figure 1.5c). As the upward vestibular influence wanes, the body tactile forequarter righting becomes predominant, but simultaneously, so does body tactile hindquarter righting. The co-activation of these two forms of body tactile righting leads to the counterproductive back-and-forth movements of the body. This pattern of righting can be readily identified by the corkscrew configuration of the body (Figure 1.5b). As the U posture is more frequent than the corkscrew posture in the first couple of days after birth, the former can be thought of as representing an immature stage of development (V. Pellis, Pellis & Teitelbaum, 1991). Comparing the relative frequency of the U and the corkscrew postures in rats that were born prematurely compared to normal gestational length revealed that the premature rats developed the more mature form of righting (i.e., the corkscrew) more rapidly (Roberto & Brumley, 2014).

With regard to understanding how righting develops, scoring time-to-right yields precision but poor accuracy (Figure 1.1a), and scoring whole sequences of righting yields accuracy but less precision (Figure 1.1b). However, by persisting with the option shown in Figure 1.1b, the sensory and motor patterns that reflect the developmental organisation of righting were identified (V. Pellis, Pellis & Teitelbaum, 1991; Pellis, Pellis & Nelson, 1992). As shown by the examples mentioned, behavioural markers that are relatively easy to learn to identify and score could be abstracted to assess particular aspects of the organisation of righting, such as the relative maturity of animals reared in particular environments. Scoring these measures yields both accuracy and precision (Figure 1.1c). Even though we have used the time-to-right measure as a foil to illustrate how behavioural markers should be selected to reflect the organisation of the behaviour being studied, if this measure is pegged to an appropriate aspect of organisation, it too can yield accurate as well as precise measurements.

For example, in trigeminal righting, a righting system that involves the same sequence of movements across all developmental stages, the

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

time-to-right measure can be useful to assess whether sensorimotor changes with age affect the execution of righting. Video records taken at a higher frame-rate than we used could numerically assess the magnitude of the small age-dependent improvement that we noticed. Similarly, individual variation is another potential factor that can affect performance. For example, Madagascar hissing cockroaches from the same colony were grouped into bold and timid animals based on their responses to various threats. Righting to prone when placed supine on the ground was tested, with the bold animals righting faster than the timid ones. However, this is a composite measure that combines both the latency to begin to right and the time it takes to right once righting is initiated. When these two components of timing were scored separately, it was found that timid cockroaches had a longer delay in beginning to right, but once they began, they right themselves just as quickly as bold animals do (Logue et al., 2009). That is, the timid animals are more inhibited in commencing movements. The same difference in behavioural inhibition has been demonstrated in rodents and has been linked to variation in spontaneously occurring levels of neurotransmitters, such as dopamine (Cools et al., 1990; Pellis & McKenna, 1992). Suppressing such neurotransmitters experimentally delays the onset of initiating righting, but does not affect the speed of righting (Martens, Whishaw, Miklyaeva & Pellis, 1996). The lesson to be learned is that, if the time taken to right to prone is used to measure an appropriate aspect of the underlying behavioural organisation, then it too can be both an accurate and precise measurement.

Some Thoughts on Hypotheses, Hypothesis Testing, and Predictions

A central theme of this book is that proposed behavioural markers are hypotheses about behavioural organisation. However, we have detected some sloppiness in the current literature in how behavioural hypotheses are framed and tested. Therefore, we think it worthwhile to look at what constitutes hypotheses more closely.

In an interesting paper, Hempel and Oppenheim (1948) draw the distinction between an *explanandum*, a Latin term for the phenomenon to be explained, and the *explanans*, the Latin term for the proposed explanation of that phenomenon. The authors use an example to clarify

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

this distinction. One person asks, 'why is there smoke?' and another responds with 'because there is a fire.' In this case, 'smoke' is the *explanandum*, and 'fire' is the *explanans* – the presence of smoke is explained by fire. Any of the behavioural examples used in this book can be viewed from this perspective. For example, in this chapter, we discuss the phenomenon of counterproductive righting movements by the fore and hindquarters in young rats (see Figure 1.3). The question of 'why do young rats perform this corkscrew action?' is the *explanandum*, the phenomenon that needs to be explained. The next phase in the process is to find an *explanans*, an explanation for why the phenomenon occurs.

Before we began to study the development of righting, one explanation for the corkscrew was that, given their sensorimotor immaturity, this was an efficient pattern of righting in infant mammals (Horwich, 1972). That is, the *explanans* is the hypothesis that the corkscrew is an infantile adaptation for righting. In contrast, as described in detail above, our explanation was that due to the immaturity of the central control mechanisms, two forms of righting – one involving tactile-triggered righting by the forequarters and another involving tactile-triggered righting by the hindquarters – are co-activated, leading to counterproductive movements that impede righting. That is, our *explanans* is the hypothesis that the corkscrew is a by-product of immaturity, not an adaptation to overcome the limitations of immaturity (V. Pellis, Pellis & Teitelbaum, 1991). Having two clearly articulated alternative *explanans*, that is, hypotheses, sets the stage for testing their relative explanatory utility. Tests are usually framed in concrete ways by generating predictions that follow logically from the hypotheses.

In this case, if the 'corkscrew as an adaptation' hypothesis were true, then one prediction should be that mammals in which righting is developed *in utero* should not have this pattern of righting. The reason is that if this pattern of righting were an adaptation, only infants that need to right would have this capacity. In contrast, if the 'corkscrew as a by-product of immaturity' hypothesis were true, this would not be the case, as the maturation of the different forms of righting should proceed in a similar manner irrespective of whether the infant lives in an environment that requires it to right or not.

We tested this prediction by examining the development of righting in a small carnivorous marsupial, the Northern quoll (Pellis, Pellis & Nelson,

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

1992). As a marsupial, quolls are born at an immature stage, comparable to that of a foetus in placental mammals. The newly born quoll then migrates to the pouch area on its mother's ventrum and attaches itself to a teat, where it completes its development until it can move about freely, apart from its mother (Nelson & Gemmell, 2004). At the age at which righting is being developed in foetal quolls, they are firmly attached to one of their mother's teats – skin grows around the edges of their mouths so they cannot be dislodged. Their forepaws are well developed with protruding claws that enable them to hold onto the thick hairs on the mother's ventrum. By removing such infants from their mothers at different ages and testing them on the ground, in the air and in water, as we did with young rats (Pellis & Pellis, 1994; V. Pellis, Pellis & Teitelbaum, 1991), we could evaluate the development of righting in quolls.

Even though when attached to the mother, infant quolls never have an occasion to need to right to prone, when tested, they nevertheless undergo the same pattern of maturation of the different righting systems as rats. Furthermore, in the quolls, when tactile forms of forequarter and hindquarter righting begin to mature, they are co-activated, resulting in the corkscrew form of righting. This is counter to the prediction made by the 'corkscrew as an adaptation' hypothesis, but is consistent with the 'corkscrew as a by-product' hypothesis. This example highlights two aspects about the relationship between hypotheses and predictions that we think needs to be closely considered when using the hypothesis testing method.

First, both when examining graduate theses and in reviewing papers for scientific journals, we have found that there is often confusion between hypotheses and predictions. A statement such as 'we hypothesise that A should follow B' does not constitute a hypothesis for the simple reason that there is no explanation as to why A should follow B. A hypothesis needs to contain an *explanans*, otherwise the proposed relationship between A and B are simply statements about the relationship between components of the phenomenon being studied, the *explanandum*. This confusion is more likely to be found in descriptive studies that derive behavioural markers for quantitative scoring. What should be in the forefront of the researcher's mind is whether what is being posited contains an explanation for the phenomenon being studied. If so, then the prediction about the relationship between A and B should follow from that explanation. In the example of the corkscrew, because of the different causal processes

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

invoked by the two hypotheses, different predictions can be made about the expected presence of the corkscrew dependent on the environment in which the young develops. Consequently, the causal processes in a hypothesis should explain why A should follow B, and, if so, the statement that A should follow B can be posited as a prediction. Measuring whether A follows B in the real world would then be a test of the hypothesis.

Second, and again, from reviewing many submitted papers, time and again we have noticed that many predictions are weakly pegged to the posited hypothesis. The conclusion is often in the form of 'the predictions are consistent with our hypothesis'. However, as reviewers, our reaction is, 'sure, but they are also consistent with three other hypotheses that we can posit.' The most convincing predictions are ones that are unique to particular hypotheses, so that, if they are supported or contradicted, they provide strong evidence for whether a particular hypothesis has merit or not. If both the 'corkscrew as an adaptation' hypothesis and the 'corkscrew as a by-product' hypothesis predicted that all young animals in all contexts should exhibit the corkscrew, then a positive finding that all young animals perform the corkscrew, even though consistent with both hypotheses, would not be a particularly useful test of either hypothesis. We should note that we are not posing as overly insightful researchers in giving this advice. Often, it is the constructive criticisms from reviewers that lead us to develop more refined hypotheses and mutually exclusive predictions. Finding the flaws in the work of others is often much easier than finding the flaws in one's own work. Let us illustrate with an example.

Spider monkeys shake their heads from side to side so as to facilitate amicable social contact. This occurs frequently during vigorous play fighting, and so it is common during the juvenile period. Occasionally, juvenile spider monkeys use headshakes during non-social locomotor play. The phenomenon needing explanation was the occurrence of these seemingly out-of-context headshakes (the *explanandum*). Based on several prevailing views about why animals sometimes perform out-of-context behaviours during play, we developed three hypotheses (the *explanans*). One hypothesis was derived from the view that play trains animals to deal with unpredictability, another hypothesis was derived from the view that immaturity can result in misdirected signalling, and the final hypothesis was derived from the view that animals may use signals to self-regulate their own emotional state (Pellis & Pellis, 2011). With three competing

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

hypotheses, it was a challenge to develop predictions that would not simply be consistent with more than one hypothesis but predictions that were mutually exclusive, so that, if they turned out in a particular way, they would not only provide confirmatory evidence for one hypothesis but would also disconfirm the other two. It took three rounds of revision with critiques and suggestions from reviewers to finally develop a set of predictions that had this highly discriminative structure (see Table 1.1).

The two hypotheses that we thought would be more likely to be involved – the training for the unexpected and the misdirection due to immaturity – turned out to have the least support; rather, it was the self-regulation hypothesis that received the strongest support. Without strong predictions that provided mutually exclusive tests of the three hypotheses – with some of those predictions suggested to us by diligent reviewers – very different conclusions would likely have emerged from this study. For another insightful example, see Waterman (2010). In the course of the book, we endeavour to use this strong version of behaviour testing to test particular hypothesised behavioural markers.

The Way Forward

It should be noted that the problem of identifying suitable behavioural markers to score is the same whether we are dealing with two or more animals engaging in an interaction or if the body parts of a single animal are combined in various ways. The first principle to digest and use then is that a behavioural marker is a reflection of the hypothesis that the researcher has in mind as to how that marker reflects the underlying organisation of the behaviour to be studied. As such, the hypothesis needs to be made explicit and tested. After all, the marker could be a poor reflection of that organisation, or, alternatively, the hypothesis about the organisation of the behaviour may be wrong or deficient.

This book will explore principles that can guide the researcher in characterising features of the organisation of the behavioural phenomenon being studied and so identify markers that would be suitable for scoring quantitatively. We do not pretend to provide an exhaustive list of principles. Indeed, in the last chapter, we offer clues to new principles that will likely need to be included in the future. Nonetheless, the principles that we

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

Table 1.1 Three hypotheses and their associated predictions regarding the non-social headshakes (HS) by spider monkeys are detailed. Adapted from Pellis & Pellis (2011) with permission (Copyright © 2011, American Psychological Association)

	Immaturity	Increasing variability	Self-regulation
Hypotheses	HS in a non-social context arise from being misdirected to incorrect targets.	HS in a non-social context are performed to experience unpredictability.	HS in a non-social context are used to motivate action in unfamiliar contexts.
Predictions	**1a.** HS should occur frequently when playfully mouthing both social partners and inanimate objects, especially when very young. **2.** HS in non-social contexts should be absent in adults. **3.** Social and non-social HS should be structurally similar. **4.** No differences predicted about actions followed or not by non-social HS.	**1a.** HS should occur frequently in all contexts, especially when the young are more independent of their mothers. **2.** HS in non-social contexts should be absent in adults. **3.** Social and non-social HS should be structurally similar. **4.** No differences predicted about actions followed or not by non-social HS.	**1a.** HS should occur in a contextually correct manner. **1b.** Non-social HS should mostly occur when young monkeys are making their first forays away from their mothers. **2.** HS in non-social contexts should be present in adults. **3.** Neither similarity nor difference predicted between social and non-social HS. **4.** There should be hesitancy in actions performed in association with non-social HS.

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press

do provide have proven to be valuable guides in our own work and that of many others. Most importantly, their utility has stood the test of time. Once the four main principles are developed over the next three chapters, Chapter 5 will show how they can all be brought together so as to aid in understanding the organisation of a complex behaviour and to identify useful behavioural markers that can be abstracted for quantitative studies. However, in all cases, in applying one or more of these principles, we stress that the resultant markers are hypotheses about the underlying organisation and so need to be tested. In the final chapter, we will also explore some novel methodological approaches for testing such hypotheses.

https://doi.org/10.1017/9781108650151.002 Published online by Cambridge University Press