Practical mechanical threshold estimation in rodents using von Frey hairs/Semmes-Weinstein monofilaments: Towards a rational method

Title page for submission to Journal of Neuroscience Methods

Article type: Review

Title: Practical mechanical threshold estimation in rodents using von Frey hairs / Semmes-Weinstein monofilaments: towards a rational method

Authors: Matthew J G Bradman[a], Francesco Ferrini[a], Chiara Salio[a], Adalberto Merighi[a]

Affiliation: [a]University of Turin, Department of Veterinary Science, Largo Paolo Braccini 2, 10095 Grugliasco (TO), Italy

Corresponding author: Matthew Bradman

Telephone: (+44) 7720713836

Email: mjgbradman@yahoo.co.uk

Text pages: 22

Tables: 2

Figures: 3

Highlights

Updating von Frey methodology to apply monofilaments more efficiently

What the test measures: theoretical considerations

Practical usage of von Frey's hairs in the laboratory environment

Tightening sensory threshold estimates from raw data

Statistics of (group) threshold estimation

Practical mechanical threshold estimation in rodents using von Frey hairs / Semmes-Weinstein monofilaments: towards a rational method

Abstract

Here, we reconsider the status quo in testing mechanical sensitivity with von Frey's hairs. The aim is to improve paw withdrawal estimates by integrating current psychometric theory, and to maximise the clinical relevance and statistical power of mechanosensory models. A wealth of research into human tactile stimulus perception may be extended to the quantification of laboratory animal behaviour. We start by reviewing each step of the test, from its design and application through to data analysis. Filament range is assessed as a whole; possible test designs are compared; techniques of filament application to mice and rats are considered; curve fitting software is introduced; possibilities for data pooling and curve fitting are evaluated. A rational update of classical methods in line with recent advances in psychometrics and supported by open source software is expected to improve data homogeneity, and Reduce and Refine animal use in accord with the '3R' principles.

Key words: Von Frey hairs; Semmes-Weinstein monofilaments; Esthesiometer; Threshold estimation; Allodynia; Rodent behaviour

1 Introduction

Touch perception in non self-reporting subjects represents a major neuroscientific challenge. Because perception is individual and internal, certain critical issues arise: Is it possible for an external observer to measure somebody else's perception? How does the magnitude of a stimulus come to be perceived? In the late nineteenth century, these questions were investigated empirically by von Frey, Weber, Fechner, Blix and others (Gardner and Martin 2000). Weighted horsehairs were initially used by Blix to analyse human cutaneous perception, and von Frey adapted the method, producing a mechanical stimulator which became widely used in animal research (Norrsell et al. 1999). Thus psychometrics may inform physiology. Monofilaments are now used extensively in medicine (Armstrong et al. 1998; Jerosch-Herold 2005) and preclinical neuroscience (Le Bars et al. 2001), often to assess pain in the hope of therapeutic translatability (Mogil 2009); yet the quantification of rodent behaviour by hair applications has fallen behind recent advances in human psychometrics (Carandini and Churchland 2013) and analysis is commonly obsolete (Milligan et al. 2004). For one discipline to inform the other, it makes practical sense that methodologies be reintegrated. Already, strong criticism has been made of hair application and operator bias (Bove 2006); over the following pages, practicable solutions already dispersed throughout the literature are drawn together, clarified, and recommended as a corrective where appropriate.

The usual primary aim of a bioassay is to derive a numerical description of how input (dose, stimulus, time etc) relates to output. In the case of von Frey hairs, this number is usually a 50% withdrawal threshold (Chaplan et al. 1994). The investigator asks: What strength of stimulus is required to elicit clear responses from this subject (or group of subjects) in half of the instances when it is applied? To answer, a relevant dataset is needed, with input stimuli spanning from below to above the dynamic range of sensation; then a 50% point may be estimated, by plotting response frequency against log stimulus intensity (Lawless and Heymann 2010). The resulting graph, at once both a dose-response curve and a psychometric function, is sigmoid in shape (example in Fig 1). It shows the 50% withdrawal point closely coincident with α, defined as the point of steepest gradient at the inflection (Pentland 1980; Wichmann and Hill 2001a).

Currently, a staircase assay known as the up-down method (Dixon 1965; Chaplan et al. 1994) is most widely used (Mills et al. 2012). This method works, but perpetuates historical limitations imposed by the physical tools, and other accumulated inaccuracies, which together increase the variance (see Section 3). Fortunately, these problems are easily resolved. We review current practices first to facilitate improved data collection (Section 4). The mathematical estimation of α (or 50% withdrawal) represents a science in its own right (Klein 2001; Wichmann and Hill 2001a), so pertinent issues are addressed, and suitable curve-fitting freeware is recommended (Section 5). Alternative techniques for quantifying mechanosensitivity are reviewed and compared (Section 6). Improved data homogeneity enhances statistical power (Dell et al. 2002), such that fewer animals are required to achieve a given confidence level (Festing and Altman 2002); thus the recommendations which follow (summarised in Table 2) may contribute to Reduction and Refinement.

2 Conceptual background

Between stimulus and effect lies the nervous system in full. Therefore, before dealing with the practical methodological task, we sketch the linking of cause to sensation; work necessarily pioneered in humans. We then make the return step from sensation to behavioural response and quantification in animals.

2.1 From stimulus to sensation

The perceived strength of a stimulus (e.g. volume of a sound or a weight in the hand) has been described by Weber's law. Laying "the foundation stone of experimental psychology" (Titchener 1915), Weber stated that for a given reference stimulus, the *increase* required in stimulus strength for a change to be perceptible is related to the magnitude of the *initial stimulus* (Weber 1846). This was expanded by Fechner (1860) to suggest that perception is on a logarithmic scale:

$$\Psi = k \log \Phi \qquad \text{Equation 1}$$

where Ψ is perceived sensation ratio, Φ is stimulus intensity, and k a constant reflecting the function of the sensory system. Fechner called Equation 1 "Weber's Law"; this is what publications on monofilaments using the term generally intend, though it is also referred to as the Weber-Fechner Law.

That the relationship between stimulus and perception approximates a logarithmic function in different sensory systems likely represents an evolutionary adaptation of the nervous system to respond over a working intensity range, and implies that firing activity of certain types of sensory neurons follows a similar intensity/response ratio (Muniak et al., 2007). Alternative relationships have subsequently been proposed (i.e. Stevens power law; Stevens, 1960), but the specific powers relating different sensory modalities and stimulus types is beyond the scope of the present review. It has been noted that k is an approximation and remains constant only across mid-range stimulus values (Stevens 1957; Savage 1970). However, this is the dynamic part of the range, where quantification is rightly focused.

With the benefit of this conceptual model (Equation 1), some measure of a sensation can be plotted as a function of its log causatory stimulus. The result is a psychometric curve (Fig 1). The most important parameters for our purposes are α and β, respectively the inflection point and its gradient. The sigmoid form of psychometric functions is mainly familiar to biologists as a dose-response curve; the similarity has been noted (Weiner et al. 2012), especially when estimating parameters from binomial datasets (Foster and Bischof 1991; Marin et al. 1991; Zychaluk and Foster 2009). The 50% withdrawal region is the most helpful region to sample from a sigmoid function because the maximum amount of information can be extracted where the slope of the curve is steepest (Pentland 1980). What is more, much behavioural work already uses this value as accepted currency; therefore its retention is both logical and expedient.

When afferent input is processed to sensation, the descriptive "transducer function" of the intervening nervous system appears similar between different senses such that only an exponent need be varied (Stevens 1957). Interestingly, this exponent may have adaptive significance for organisms and senses exposed to differing environmental ranges of experience (Lawless and Heymann 2010). A straightforward log scaling as in Equation 1 assumes an afferent exponent function = 1, which is a good approximation to measured values of 0.9 to 1.0 for glabrous skin of the human hand (Greenspan and Bolanowki 1996). The exponent is approximately halved in hairy skin (Jones 1960; Mountcastle 1967).

2.2 Returning to behaviour

The inaccessibility of perceived sensation (Ψ) in non self-reporting subjects mandates an indirect approach through observable responses.

Explicitly psychometric approaches are increasingly being applied to rodent perception (Carandini and Churchland 2013), including mechanosensitivity (Song et al. 1999; Milligan et al. 2005; Morita et al. 2011; Ghanouni et al. 2012). But do monofilaments actually test sensation, or is it merely the case that we may use tools and analyses of human perception towards our aim of quantifying animal behaviour? Any connection between mechanical stimulus and a psychometric-type analysis of subsequent behaviour might be contingent; this being allowed, the analysis remains applicable.

In humans, a number of classical and recent psychometric tests exists (Gescheider 2013), of which many might be applied to non-communicating subjects. Behaviour can be categorised as yes-no, or might be assigned a numerical intensity (Bi and Ennis 1996). In the process of doing so, a level of subjectivity must be introduced whereby an observer is ultimately responsible for deciding the result. Published data support this contention; a prime source of variability in yes-no rodent behavioural assay exists between human operators (Chesler et al. 2002; Bonin et al. 2014; Sorge et al. 2014), especially in control animals (Chaplan et al. 1994). Indeed, in a strong sense, what is under test is equally the *operator's* perception in response to the stimulus of a behaving rodent! Therefore, the whole experimental approach needs to be conceived considering operator, as well as subject, in order to estimate mechanosensitivity in the most unbiased way.

To minimise the effect of intra-operator subjectivity, simple binary categorisation makes intuitive and statistical sense. Though superseded for human subjects by less biased two-alternative forced choice (2AFC) testing (Gescheider 2013), yes-no testing is well suited to von Frey use in rodents, though very recent theory considers including a third choice of "uncertain response" to be an improvement (García-Pérez and Alcalá-Quintana 2013). For the animal receiving the stimulus, von Frey may be considered a true yes-no test; for the operator it may remain so if failed attempts at monofilament application are recognised as such and discarded.

Finally, subjectivity can be limited by blind testing. The necessity for the operator to be blind to experimental variables is more extensively discussed in Section 3.4. Here, it is worth mentioning that ideally the operator should be also blinded to the applied stimulus of each monofilament. Unfortunately such a condition is not easily applicable under common experimental settings.

3 The physical test

3.1 Semmes-Weinstein filament numbers

Semmes and Weinstein modified von Frey's wood-mounted animal hairs; today a standard esthesiometer set contains 20 nylon monofilaments, summarised in Table 1. Handle markings are dimensionless values, originally log transformations of a $1/10^{th}$ mg scale (Dellon et al. 1993; Weinstein 1993) such that:

Quoted stimulus = Log (*force in g* x 10 000) Equation 2

As well as handle markings, several hair stimulus measures are in general use and are, in fact, more accurate. The "target force" in g inflicted by each hair is not that which may be calculated from the handle markings. Note therefore that handle mark does not accurately describe filament stiffness, so another measure of stimulus should be used when calculating withdrawal threshold.

3.2 Standardising stimulus values

Monofilaments may deliver reproducible forces (Bell-Krotoski and Tomancik 1987), including some filaments aged over twelve years (Haloua etal., 2011). However, stimulus accuracy may conversely undergo significant alteration over time, when repeatedly applied (Lavery et al. 2012). The level of degradation appears related to the intrinsic qualities of nylon used by different manufacturers (Bell-Krotoski et al. 1995; Booth and Young 2000; Yong et al. 2000). Therefore it is recommended that each set is calibrated regularly (Bell-Krotoski et al. 1995; Voerman et al. 1999). It is sufficient to place a piece of paper, tape (Yong et al. 2000; Haloua et al. 2011) or rubber glove (Levin et al. 1978) to prevent the tip from slipping on the smooth surface of a pan balance, and record a series of repeated measurements from buckled hairs in grams. These may then be converted by Equation 2 and the calibrated values may be used in preference to "target force" when estimating withdrawal threshold.

It is worth distinguishing between force (expressed strictly in newtons but commonly in grams) and pressure (Dellon et al. 1993; Stoetling Co 2001) which takes into account the contact area. During filament application, skin is compressed axially under the probe, but also stretched concentrically across a locally diffuse indentation. Speculatively, mechanoceptors in these regions are subjected to stretching and compression (Del Valle et al. 2012). Details of the tip profile as it abuts the skin are of biological relevance, because different contact types varyingly affect surrounding tissue, and may in turn activate different types of mechanoreceptor (Treede et al., 2002). Thus, functionally different fibre types may discern particular aspects of a von Frey application.

To know the effect of a given stimulus in its local biomechanical context would be ideal for quantifying behaviour in terms of 1) specific receptor activation, resulting from 2) intradermal deformation, initiated 3) by some quantified aspect of filament bending. Usually, our knowledge of 1) and 2) is either speculative, or itself a variable under test. Because filament bending concentrates contact towards one edge, the circular tip area is not biologically meaningful. Experimentally, mechanical threshold duly correlates better with tip circumference, rather than area (Greenspan and McGillis, 1991). Taken together, these arguments explain why a log of the bending force is the preferred stimulus measure (but see Levin *et al.,* 1978, who recommend "stress").

3.3 Uneven steps: a potential trip hazard

Monofilament esthesiometer sets, sold as standard, increase in unequal intervals. The actual steps between hairs increase on an approximately logarithmic scale (Werner et al. 2011) in approximate accordance with Weber's law (see Equation 1 above). However the intervals, or steps, between log values are not even (Fig 2), for the good reason that standard kits reflect filament diameters dating from the early availability of nylon (Lambert et al. 2009; Trossarelli 2010). Thus modern von Frey sets reproduce an arbitrary historical variable. This problem is limited if threshold is estimated by a robust analysis method; though commonly, analysis is not so flexible (see Section 5).

In particular, if the sum of two particularly small steps between adjacent hairs is closer to the mean step size across the set, it may be preferable to combine the steps. This occurs twice across the set in the

regions shown in Fig 2: around the 9th, and around the 13th/14th hairs. Using log step size, the coefficient of variation c_v was calculated using target force for all 20 hairs (c_v=0.379) and for sets missing 9th and 13th (c_v = 0.279), and 9th and 14th (c_v= 0.293) hairs. Similar improvements result using our measured forces. We conclude that the 9th (handle mark 4.17) and 13th (handle mark 4.93) hairs should be removed from von Frey sets, especially where the method of Chaplan *et al.* (1994) is used for data analysis. Note that this recommendation is based on target force, but is also reflected by a calibrated filament set in our hands.

3.4 Applying the hairs

The manner in which von Frey hairs are used affects the stimulus they administer, so careful application should increase repeatability. To arrive at consistent stimulus intensity, hairs may be flexed immediately prior to use (Booth and Young 2000). Mechanical sensitivity changes according to activity in live subjects (Massy-Westropp 2002); rodents should not be tested while grooming, as this behaviour greatly increases the withdrawal threshold (Callahan et al. 2008). Limbs under test should be weight bearing on a metal grid rather than hanging free (Kauppila et al. 1998). The support affects measured threshold, but while a flat surface may be preferable despite filament *guiding* through a small diameter straight-sided hole (Pitcher et al. 1999), values quoted for almost all rodent testing are derived from metal grids. A clear view of the feet from below is helpful. Because of the operator's role in determining the stimulus, and the subjective nature of the response, it is important that the operator is blinded to experimental variables (Bove 2006; Begley 2013).

One at a time, filaments are advanced perpendicularly to a subject's skin until they are seen to buckle. They should not be bounced onto the skin, but rather applied smoothly (Bell-Krotoski and Buford 1997; Detloff et al. 2010). Beyond the initial buckle, increased bending yields no change in applied force but may add a lateral component to alter the stimulus (Johansson et al. 1980). In rodents, the hindpaw mid-plantar region is used most, avoiding the raised tori (Chaplan et al. 1994) where epidermal thickness increases and innervation with calcitonin gene-related peptide-immunoreactive fibres is reduced (Duraku et al. 2012). Medial or lateral deviation causes some baseline difference (Hogan et al. 2004), however surgical intervention (spared nerve injury) causes particular increase in the medial-lateral effect of hair placement (Duraku et al. 2012). It seems remarkable that relatively little is known about cranial-caudal sensitivity; some attempts have been made to sample across this variable by testing throughout the plantar aspect (Song et al. 1999; Hogan et al. 2004). Accordingly, we presume variable sensitivity along the cranial-caudal axis which mandates care in selection of test area. Importantly, human plantar testing is concentrated around the hallux and metatarsal heads, with the heel more rarely tested (Baraz et al. 2014).

The length of time a stimulus continues to act upon the nervous system is physiologically important (Lucas 1910). Temporal summation is effected at the synaptic level (Magee 2000) but also, in describing whole-organism sensory systems, the concept of stimulus summation has been variously accommodated into human psychometric models (Watson 1979). Given the importance of stimulation time, the lack of standardisation in duration of hair applications may rightly be severely criticised (Bove 2006). Contact time with rodent feet ranges from two hair applications per second (Seltzer et al. 1990) up to prolonged application of single hairs for 10 s (Millecamps et al. 2007; Metz et al. 2009). This is a variable which has not been addressed in comparing reported values for rodent withdrawal. Human clinical applications last from 1 s (Mueller 1996; Smieja et al. 1999) to 2 s in order to assess integrity of cutaneous sensation (National Diabetes Education Program 2000) and estimate threshold (Rolke et al. 2006). At the very least, internal consistency of application time is important.

3.4.1 Potential pitfalls during testing

The force in a column at buckling was described mathematically by Euler (Oldfather et al. 1933). Its accurate modelling is complex beyond our current scope, but the variables involved are informative. Effective hair (column) length is related to actual length by a constant reflecting *tethering*, i.e. how the column interacts with its support at one end and its load at the other. At the handle end, a von Frey filament is *fixed* because it is glued in place. At the paw it should be *pinned*, such that it is unable to move around, without being physically bonded; but if the hair can slide, this tethering constant more than doubles (Levin et al. 1978), changing the effective filament length. Thus if a hair slips across the paw pad, the interaction between filament and skin changes and the energy delivered is likely very different. Since the force delivered under this condition is unknown, the application should be repeated. If a separate tip is attached to the hair, as in some efforts to standardise the diameter, the tip is no longer *pinned* but may be *guided* or *fixed* and the force once again changed (Huang and Luo 2011), so recalibration with the tip becomes essential.

Humidity, and probably temperature, affect elastic modulus and thus bending force (Andrews 1993; Haloua et al. 2011; Werner et al. 2011). Humidity is the most variable of atmospheric characteristics, and indeed each of these references reports a significant effect within normal humidity ranges. For labs where humidity may change from low to high across the course of an experiment, bending force can decrease by over one third (Haloua et al. 2011); in such instances, we recommend the adjustment scales in the cited references. Alternatively the problem can be circumvented with optical fibres (Fruhstorfer et al. 2001), but this refinement – swapping nylon for glass – appears not to have been widely adopted. There may be good reason not to demand that all testing environments should be identical if animal behavioural studies are to yield robust and repeatable results (Richter et al. 2009), at least where the effects of known variables are understood.

3.4.2 Recording responses

A successful hair application is binary: response, or no response. Paw withdrawal behaviours are widely varied so a paradigmatic positive response is not easily described. The literature focuses on paw lifting but includes shaking, licking and tapping (Choi et al. 1994). Stronger responses include prolonged licking and paw biting, escape attempts, repetitive flicking, and vocalisation (Seltzer et al. 1990). The speed at which the paw is withdrawn must be brisk (Kim et al. 1997), however the response may come after the filament is removed from the paw. Experience suggests that some of the strongest and most prolonged withdrawal behaviours are exhibited a clear second or so after contact ceases. We frequently observe a mild response of hyperextension of the digits; considering this to be neither zero response nor withdrawal, we retest the hair in these cases. In the case of asymmetric (lateralised) models of painful injury, considering only the range of exaggerated responses such as prolonged withdrawal and guarding improves the ability to distinguish lateralised effects (Hogan et al. 2004).

A significant proportion of hair applications are clear failures. The hair may miss-hit the foot or slip, and animal movement (apart from response) frequently occurs. Judging when not to accept an application inevitably brings with it operator bias, but for the animal's behaviour to play a greater role than the investigator's decision process, a high proportion of failures seems preferable to recording excessively from confounded tests. While distinguishing failed application, irrelevant movement and actual withdrawal may appear simple, it is not easy to maintain throughout testing; therefore Fig 3 proposes a simple algorithm for maximum consistency.

4 Test design

4.1 Full battery or staircase?

The ideal design yields a good estimate of withdrawal threshold with the minimum number of applications. The variables in test design are which hairs to apply, and how many times to repeat their application. At least three different test paradigms can be identified: 1) several hairs are all applied an equal number of times, and % withdrawal from each is calculated (Kim and Chung 1992); 2) hairs are applied in ascending order until the first withdrawal is observed (Takaishi et al. 1996); and 3) each response or non-response by the animal under test informs the selection of the next hair (Pentland 1980; Chaplan et al. 1994). This third procedure is widely known as a staircase ("up and down") assay. It was developed to assess explosives (Dixon and Mood 1948), whereby if a substance failed to detonate in response to being dropped, the next sample was logically tested with a higher fall, and *vice versa*. This is clearly useful in psychometric behavioural testing. Sequential stimuli are selected using a simple algorithm (Fig 3), which concentrates the most-repeated measures around a point of greatest dynamic sensitivity. Thus, fewest applications may be used to reach confidence in estimating α.

Staircase testing has been refined and adapted, such that modified forms can be advantageous (García-Pérez 2001); however the underlying concept remains unchallenged in its superiority (Fuh et al. 2003). This is especially true for the "method of constant stimuli" where stimulus values are fixed, as in Semmes-Weinstein filaments. Human psychometrics has discarded non-staircase constant stimulus methods, in favour of exploring variations within the accepted paradigm of up and down stimulus application (García-Pérez 1998; Shen 2013).

Both staircase and non-staircase methods have been proposed to measure mechanical sensitivity in laboratory rodents. Non-staircase repetitive methods typically address the frequency of responses to sub- or supra-threshold stimulus, aimed at sorting animals into allodynic/non-allodynic or hyperalgesic/non-hyperalgesic (Detloff et al. 2010). However, it is preferable to avoid such dichotomization which may at once lead to a loss of relevant information and to an oversimplification of the data (Cohen 1983). Moreover, adopting pre-established pain/non-pain bins seems to favour a circular degradation of the method, justifying poor threshold quantification, which in turn blunts the sorting of experimental animals. Conversely, staircase methods do not detect a normal or abnormal index of sensitivity according to pre-established definitions, but rather allow derivation of an actual threshold estimate for the purpose of quantifying mechanositivity. It is also worth considering that monofilaments evoke withdrawal without aversive pain (Wu et al. 2010; Rigaud et al. 2011). Thus, while we acknowledge that a quantitative change in estimated threshold may be referred to altered mechanosensation, and then be used to define group cutoffs, a leftward shift of the estimate should be considered as distinct from a qualitative change to nociception.

4.2 How many steps make a staircase?

Given the staircase as a preferred paradigm, the question arises: how many individual hair applications should be made per test run (i.e. to contribute to the creation of each data point). It may not be simply the case that more applications per animal create inherently more accurate estimates, because a change in the subject's behaviour part way through a single testing session is likely to confound analysis (Klein 2001; Fründ et al. 2011) and the more testing is carried out, the more likely it becomes that habituation will occur. Between testing sessions, Chaplan et al. (1994) found a threshold reduction in control rats following daily testing for three days, comprising a total of up to 81 filament applications per paw. Repetitions adequate to the requirements of statistical sampling and the ethical benefit of using animals to elicit a

maximum of data must be balanced against behavioural consistency, habituation, and the patience of the investigator.

Dixon considered the application preceding a first change of response, which signals the first change of staircase direction, to be the critical point from which applications should be counted. He used the prime (‘) to differentiate between total applications per sequence, and applications minus the lead-in. For example, the sequence “oooxxoxo” contains N’=8 tests, where ‘ denotes a full count of all hair applications in the string. The same sequence is also meaningfully described as N=6 (note absence of ‘), discounting the same-direction steps preceding the test immediately prior to the first direction change. This is because tests after the first direction reversal are likely to be nearer the 50% point, and form a distinct part of the chain for the purposes of Dixon’s analysis. Regardless of the N’ count, N=6 satisfied his convergence model, and was found to be adequate for chemical dose (LD_{50}) estimation (Dixon 1965, 1980).

Comparing monofilament staircase lengths, Detloff et al. (2010) suggested N’=10 (i.e. a total of ten applications irrespective of reversals of direction) to be adequate in rats, and preferable to Dixon’s N=6. Bonin *et al.* (2014) developed their SUDO method, with a time saving reduction to N’=5 combined with an adjustment factor, capable of reproducing closely rat and mouse thresholds estimated using Dixon’s N=6. To reach N’=10, the increase in sampling beyond N=6 is generally slight, and accommodates the confounding case in which no reversal of staircase direction occurs during the first five steps.

Of course, to stipulate an absolute N’ implies some prior knowledge of a likely region for α, but this is not unreasonable in the laboratory environment. Dixon (1965) showed that N=6 (and by extension SUDO’s N’=5) is adequate to estimate median lethal dose (LD_{50}). We do not argue a special case for 50% withdrawal; however the greater variability inherent in behavioural interpretation may need to be reflected in staircase length. In reality, the number of applications required to estimate with reasonable confidence is specific to each subject under test. That is to say, $N_{(IDEAL)}$ is itself a function of β. Ideally, in-test recalculation of a running threshold will allow every staircase to be stopped once a predetermined confidence value for threshold estimate is reached.

The flexibility of modern analysis (see Section 5.3) allows testing with any given N’ with any known stimulus strengths applied regardless of sequence. Therefore, step sizes might start off at greater than single intervals early in a testing sequence, diminishing later to single hair steps in order to refine the estimate after some reversal is observed (Cornsweet 1962). Thus tactile thresholds may be approached more efficiently (Dyck et al. 1993; Berquin et al. 2010).

4.3 Rats *vs* mice: a note on passive foot lifting

An up-down method for estimating mechanical sensitivity in laboratory animals was first applied to rats (Chaplan et al., 1994) and subsequently mice (Malmberg et al. 1997). Although conventional psychometric principles likely apply to both models, species differences exist which impinge on staircase testing.

In rats especially, increasingly stiff filaments may be applied until such a time as the weight-bearing foot is lifted passively by the force of a stimulus which yet elicits no withdrawal behaviour. In adult Sprague–Dawley rats, Detloff et al. (2010) estimated the frequency of passive lift at approximately 1 in every 60 filament applications. There is a lack of consensus concerning what to do with these events. Some authors have considered them to be positive responses (Pogatzki et al. 2002) or have retested the same hair after waiting (Detloff et al. 2012). It is also common practice to set the first passive lifting stiffness as an upper cutoff for the withdrawal estimate, generally at around 10% of animal weight, variously 15g stimulus

(handle mark 5.18) (Chaplan et al., 1994) up to 28.8 g in adult male rats (Hamity et al. 2010) and even 100 g (handle mark 6.10) (Detloff et al. 2012).

The problem, then, is where on the response curve (example in Figure 1) these passive lifting events might lie. The only way to gather evidence for this is to keep testing around or below the passive lifting point: if repeated measures with the next-lightest hair also very rarely evoke any response, clearly the 50% estimate should lie to the right of this point. Otherwise, the passive lift is of less concern because sampling is feasible from both sides of α (50 % withdrawal) and so curve-fitting can go ahead (see next section). Therefore, in rats responding to a series of retests with a consistent passive lift and no withdrawal, care should be taken to test the next lightest hair repeatedly and the numbers of responses and non-responses recorded. In this way, special cases in which the passively lifting stimulus needs to be recorded as the estimate are strongly contained.

Finally, it should be noted that the upper lapse rate (see Figure 1) may be greater than expected in rats prone to passive lifting and so, to limit bias, a reasonably inflated estimate should be fed into subsequent analyses (Prins 2012). Flexible curve fitting systems do require some responses as well as some non-responses, but a limited number of either does not preclude threshold estimation (see following section). Similarly, in the case of a superbly sensitive animal, an initial response to the lightest available stimulus may also be considered as a mirror of the passive lifting situation; it does not necessarily prevent meaningful data gathering. Ultimately, it is possible to set a minimum threshold value as a cutoff.

5 Analysis of staircases

To generalise empirical data, a model equation may be fitted mathematically, and estimates then derived for various meaningful parameters. With nonlinear relationships such as paw withdrawal, it is usual first to choose a model then to apply an iterative fitting process. Curve fitting by maximum likelihood estimation (MLE) applies the dogma that parameters can be estimated from a sample of a Gaussian distribution (Fisher 1922). Although not unchallenged (Foster and Bischof 1991), MLE persists as a psychometric standard; it underpins Dixon's and other methods discussed below.

5.1 Dixon's modified analysis

The model most used for von Frey data was developed by Dixon (1965) for toxicological estimation of LD_{50} by staircase analysis. It employs the following equation:

$$LD_{50}=x_f+kd \qquad \text{Equation 3}$$

Where x_f is the log of the final dose applied in the staircase, *d* is the step size, and *k* refers to a lookup table of values calculated by MLE from the **Gaussian cumulative** distribution (see Section 5.2). This was well suited to longhand pen-and-paper calculation, and met a basic need for dose-response estimation. An innovation by Chaplan *et al.* (1994) was to combine Equations 2 and 3 such that Dixon's lookup tables could be used to estimate 50% g threshold in von Frey experiments.

Dixon's (1948) staircase protocol remains useful for data *collection*, but there are limitations inherent in the *analysis* (Dixon, 1965) as follows:

1) Dixon specified equal sized steps (in log units), and his analysis assumes constant steps, expressed as ***d*** in Equation3. This became ***δ*** in Chaplan's modified method, defined as *mean* difference (Chaplan et al. 1994). However, usable means differ between e.g. the range of hairs applied per animal, or the mean of the experimental testing range (in our animals frequently 14% variability of δ from per-animal range), or

perhaps the entire hair set (see Table 1 & Fig 2). The up-down method was intended for instances where log steps could be accurately applied, for example serial dilutions in toxicity testing. Dixon states clearly that while log stimulus gradient may deviate from standard deviation (SD) by as much as 50% (i.e. the slope should lie between 2/3 SD and 3/2 SD), the steps should be equal. Modern monofilament sets contain uneven steps, but this analysis method sees *only* the final hair value; an estimate of mean step size (***δ***) sets the entire test range. Therefore applying the Dixon-Chaplan method to von Frey hairs forces an over- or under-estimation of *every* calculated 50% threshold. This is avoidable by analysis methods which manage individual stimulus values for each hair.

2) The lookup tables do not extend to any start run of five or more values without reversal (e.g. ooooox...). For such instances, and also for test runs where N>6, either data must be discarded or else a separate equation is provided (Dixon 1965). However, the concordance between values derived from the two equations appears to be poor. Where animals are refractory to staircase testing, for example responding frequently to the lightest hair (handle mark 1.65) or passive foot lifting, an arbitrary value must be assigned (see Section 4.3); otherwise alternative analysis methods must be sought.

3) Adjacent experimental sequences and their resulting threshold estimates are not always internally consistent. For example, depending on step size (i.e. mean interval throughout the set versus mean interval across the hairs applied), values for “oxooxo” and “oxooox” can cross over one another so that “oxooxo” (which logically implies a more sensitive state) produces a withdrawal estimate *higher* than “oxooox” (which logically implies less mechanosensitivity). In other words, the calculated threshold estimates resulting from a similar pair of sequences become logically inverted relative to one another. Such inconsistencies depend on the step size which, contrary to Dixon’s founding assumption, is not equal across the range and may fluctuate as a constant between different test sequences.

5.2 Probability distributions: model choices

Fitting a mathematical function to real-world data represents a science in itself and in this respect the following paragraph is far from exhaustive. The purpose here is to provide an overview of mathematical models that are often applied to psychometric data, in order to foster a rational choice of analytical method and curve fit. An excellent and accessible tutorial for readers unfamiliar with estimation methods is available (see Myung 2003).

Before fitting a sigmoid function to psychometric data, an estimate of parameters (α, β, and asymptotic values, Fig 1) can be made by Bayesian inference (Kuss et al. 2005). Once these parameters are estimated, curve fitting by MLE routines may then be applied without the risk, for example, of generating impossible negative estimates (Wichmann and Hill 2001a). How then to choose the best mathematical model to fit the psychometric function? In a broad sense, any model should approximate to the real data distribution. However, as no “correct” data shape is known, even when a model fits the data well, it may not represent well the underlying process, thus yielding poor parameter estimates. The chosen model should combine goodness of fit in specific cases with the possibility to apply that model for fitting all data of the same kind (Myung 2003). At this juncture, several common models compete for favour:

The **logistic function** and the **Gaussian cumulative** are both symmetrical about the centre of the Y-axis, and therefore carry an advantage in fixing α to the 50% point, at least where lower and higher asymptotes reach 0 and 100 (or upper and lower lapse rates are set equally, as in Fig 1). Choosing between Gaussian and logistic models is often a matter of convenience in psychometrics (Tiest and Kappers 2011).

The **Weibull function** is effectively a family of curves (defined by shape parameters), used widely in engineering as well as psychometrics. It explicitly lacks assumptions, its appeal being the widest likely applicability of the simplest equation (Weibull 1951). Indeed in the case of von Frey, a "correct" model to fit is unknown. However, Weibull's 50% point may diverge from α (Strasburger 2001), requiring interpolation to estimate 50% withdrawal.

The **Poisson distribution** is classically applied to rarely (but constantly) occurring events and is recommended for use where the variance is approximately equal to the mean (Forbes et al. 2011). However, it has been used to model flashed light response parameters for visual psychometrics (Hecht et al. 1942).

Nonparametric methods are immune to the effects of heteroscedasticity and poor model selection, but require larger datasets to implement and so are not likely to be useful for individual animal von Frey analysis (Zychaluk and Foster 2009).

5.3 Software for estimating curve parameters

Ideal software for our purposes should compute all stimulus information in order to maximise usefulness of each test (Pentland 1980) and thus minimise animal numbers. Where animals are either too sensitive or unresponsive for strict staircase data collection, input should allow for this (see for example Section 4.3). It should allow some flexibility of fitting different models and parameter estimates with reasonable priors, and be capable of estimating α at the 50% point (for backwards-compatibility with past studies).

Much free software for threshold estimation is available (Harvey 1997; Wichmann and Hill 2001a; Miller and Ulrich 2004; Peirce 2007; Zychaluk and Foster 2009). Of these, Psychofit (Harvey 1997) has already been recommended for monofilament testing (Milligan et al. 2004) and used to estimate sensory thresholds in rats (Milligan et al. 2000, 2005) and humans (Linschoten et al. 2001). It can fit the Gaussian function, which underlies the Dixon method modified by Chaplan (1994), and is also easy to use; therefore we recommend it here. Other software may supercede it; curve-fitting routines are under constant development for Matlab (MathWorks, Inc.) and the R-project (GNU public license). PsychoPy (Pierce 2007) is powerful, uses an intuitive GUI, and is programmable to run flexible adaptive staircases; but it does not readily accommodate pre-defined stimuli, as in the case of monofilament sets.

The Psychofit program is available at http://psych.colorado.edu/~lharvey/html/software.html. Raw data are input by means of a text file included with the download; Milligan *et al.* (2004) show an example modified for von Frey, which we reproduce and extend in our Supplementary Material. Note that input is not limited to the x and o Markov chain; rather any number of hair applications can be made at any stimulus level and in any order. This allows for flexibility of step size, and also for nearest-hair repeat tests in cases of passive foot lifting or extreme responsiveness. Preset "flags" of 1 or 0 switch on and off the available curve models, and also turn on parameters such as α and β estimation. Excellent instructions for using Psychofit are in Milligan *et al.* (2004, pp 84-86). We may add that starting estimate of α should be adapted to reflect the data, and that when inputting the data filename, the .txt extension is required for the program to recognise the input file (which needs to be stored in the same folder as the executable program).

5.4 Subsequent statistics: analysing experimental groups

Whatever presumed function relates Ψ to paw withdrawal, it appears to fit a logarithmic approximation, or equally an exponent of approximately 1 (see Section 2.1). The intervals between Semmes-Weinstein monofilaments are *approximately* logarithmic steps (Fig 2), and it is essential that withdrawal estimates are

handled in their log state to avoid bias across the range (Mills et al. 2012). The insight of Mills and collaborators is of particular interest especially when von Frey testing informs pharmacology, for two main reasons: 1) Graphical representations of drug responses are right-shifted in linear scale; thus misleading ED_{50} values are produced unless log-scale axes are properly used. 2) Expressing threshold values in log brings information derived from animal studies closer to clinical observations, therefore favouring a correct translation of experimental findings into clinical settings. Once statistical comparisons are made, it is through habit that paw withdrawal estimates may be converted to grams. For comparison with previous studies, it may be safest to convert grams into log values and not *vice versa*.

Are withdrawal estimates from multiple animals best analysed by parametric or non-parametric means? This is usually the first question posed by statisticians regarding any dataset, but there exists in the literature on esthesiometers some confusion. Theoretically, we may distinguish between the discrete/continuous nature of the filament range, and the discrete/continuous nature of estimates of α. Ideally the hairs should be discrete tests, while the stimulus they aim to model is continuous. But do the discrete stimuli act effectively as narrow bins, or rather as samples from a continuous spectrum of overlapping, normally distributed curves? And, once raw stimuli are transformed into response and an estimate of α (generally by parametric curve fitting routines), might these derived data then be compared by standard parametric statistics?

Short staircases clearly sort animals into a smaller number of discrete bins, while longer test runs dilute this effect as bins become narrower. Yes-no test categories are nominal, but estimates of α represent interval-type data (regardless of bin size) and so, up to this point, there exists no theoretical barrier to comparing group means of staircase-derived withdrawal estimates by parametric tests (Mills et al. 2012). Beyond these criteria, if the population distribution is not theoretically normal, or if the actual sample fails a normality test, then clearly a non-parametric comparison must be performed.

5.4.1 Pooling data – how far to go?

Most commonly, a threshold is estimated per animal. We may then calculate mean and variance within a study group for standard statistical analyses ("fit-then-pool"). Yet the plantar surface of a foot is treated as if homogenous; in whole-body models plantar surfaces are typically combined into one, either as a mean of left and right paws or by pooling data. The commonest diabetic neuropathy, for example, is symmetrical (Llewelyn 2003; Aring et al. 2005; Bansal et al. 2006). In lateralised neuropathy models, data from left or right might rationally be pooled. Thus crystallises a question: Is it best to restrict curve fitting to per-animal estimates of withdrawal threshold, or to estimate threshold from one function fitting data pooled from several animals ("pool-then-fit")?

To combine hair applications within a group and fit one psychometric function to several animals avoids the problem of small data sets giving relatively poor estimates of α (Wallis et al. 2013). Also if the number of animals within a group is low, the reliability of standard deviation as a measure of variance is also poor (Foster and Bischof 1991; Kingdom and Prins 2010). The preferred method of estimating error from whole-group pooled data is the bootstrap method, using Monte Carlo simulations of repeated sampling from the data set. This does not give variability of the parameter within the external population, rather the variability of estimation; its advantage is that a significance of difference between groups may be estimated (Foster and Bischof 1997; Wichmann and Hill 2001b). The disadvantages are that if threshold genuinely varies between individuals, β may be underestimated, while a combination of intra-animal and inter-animal error, known as "overdispersion", causes underestimates of variance by analyses based on classical functions such as the Gaussian or logistic (Bi and Ennis 1996). A pool-then-fit approach can incorporate

overdispersion using a beta-binomial distribution (Ennis and Bi 1998; Young-Xu and Chan 2008). Comparing overdispersed pooled experimental groups using the beta binomial is beyond the scope of this paper; for such packages see for example 'VGAM' and 'aod' at http://cran.r-project.org/.

6 Future directions

There have been additions to the battery of behavioural tests available and which overlap with von Frey. We may reasonably ask whether von Frey may be out of date. For example, automated gait analysis using the Catwalk system (Noldus Information Technology) represents an elegant method for assessing the extent of loaded contact between paws and a glass floor (Vrinten and Hamers 2003). It works well with unilateral lesions which change the gait, and in this scenario correlates well with von Frey, but might be less sensitive to bilaterally elevated pain resulting from systemic changes. Also, what is measured is subtly different; not a single-point external stimulus which may reveal a leftward-shifted curve (sensitisation), rather a willingness to weight bear.

A more closely related development has been the addition of an electronic force transducer to a point stimulus. The resulting devices are commercially available in two forms: a mechanically advancing probe, which records time to withdrawal along with force applied at the moment of withdrawal (dynamic plantar aesthesiometer), and a handheld nylon stimulator of fixed diameter which the investigator advances manually, and which records applied force (marketed as Electronic von Frey) similar to a spring-gauge algometer (Wallas et al. 2003). In a human clinical setting, traditional von Frey may be slower to apply and less repeatable than handheld electronic von Frey (Tena et al. 2012). In rats, electronic von Frey was reported to be more sensitive, but by comparison against a poor method which attributed withdrawal threshold to the lightest Semmes-Weinstein filament to evoke any response (Vivancos et al. 2004). In a comparison using three established neuropathic pain models, von Frey hairs applied with the Dixon-Chaplan method detected differences in all models, whereas a dynamic planter aesthesiometer fared less well, presumably due to postural change (Nirogi et al. 2012).

What may be overlooked in comparing methods is that the underlying premise of electronic von Frey represents a classical psychometric paradigm quite different from staircase testing, namely the "ascending method of limits" (Gescheider 2013). Handheld electronic hairs might usefully be developed to standardise application time; the underestimated extent to which stimuli accrete towards a response is discussed above (Section 3.4). Beyond cost/benefit (de Sousa et al. 2014), innovations need be assessed against older methods in biologically and statistically comparable ways in order to identify which methods save time or improve repeatability, and whether they are in fact testing comparable stimuli.

Monofilament testing retains appeal for clinical use in humans (Jerosch-Herold 2005; Keizer et al. 2008); it may be that continued improvement to von Frey brings rewards without financial burden. Routines such as Parameter Estimation by Sequential Testing (PEST) and QUEST exist to create idealised staircases, in which accumulating response data are used to predict the best intensity for each successive stimulus (Pentland 1980; Watson and Pelli 1983; Linschoten et al. 2001). This level of efficiency lends itself to continuous variables, and has been used to test rodent olfaction (Clevenger and Restrepo 2006). For PEST or QUEST to be applied via a mechanical stimulus would represent a clear improvement; what is more, the duration of application could be controlled. The operator may advantageously be blinded to the applied stimulus forces, thus removing one source of bias. What is more, testing could be stopped in each animal or group as soon as a predefined confidence limit was reached.

7 Summary of recommendations

Behavioural data will reflect the variability inherent in animal and experimental models. Consequently, it is important that testing itself does not contribute unnecessarily to variance, and thus to using unnecessary animal numbers to achieve confidence limits. Considering von Frey's test, Table 2 emphasises fundamental considerations which may readily improve methodology.

Acknowlegements

We would like to thank Lewis O. Harvey, Jr. for his advice and for pointing out the range of software available for psychometric analysis. Marco del Giudice provided helpful comments on the manuscript. This work was supported by a grant from Compagnia San Paolo (Fondi di Ateneo 2011).

Figure legends

Figure 1: Example of a Gaussian cumulative function fitted to von Frey staircase data; a dose-response curve, or a psychometric function of the animal/observer? Fitted by MLE (Harvey, 1997) to combined data from 42 mice tested with von Frey filaments (each point shows proportion of responses per hair). α = point of inflection, here fixed to the 50% withdrawal point by equal lapse rates and the symmetry of the Gaussian function. β= gradient at α where the function becomes linear. γ = lower lapse rate, at which subjects show a withdrawal response when tested with zero stimulus, or where the investigator records a withdrawal when there is none. δ = lapse rate at top of function (false negatives). Note: α and β estimation applied here for best fit; for best α, we concur with Milligan *et al.* (2004) in leaving β unflagged.

Figure 2. Magnitude of theoretical and measured steps between hairs; measures derived from Table 1. Ordinal numbers from complete Semmes-Weinstein set. Y-axis shows buckling force intervals in log units (by Equation 2). Interval between handle marks of 1^{st} – 2^{nd} filaments is above the scale (=0.71). Curly brackets show region of successive small steps where a hair may be removed. Coefficient of variation revealed greatest improvement in the absence of 9^{th} and 13^{th} hairs (marked 4.17 and 4.93); see Section 3.2 for details.

Figure 3. Flow chart for testing by the up-down, or staircase, method. The output is a Markov string of responses and non-responses. Step sizes may be varied (adaptive staircase testing), starting as larger steps then diminishing as a central value is approached, where each stimulus value is known and threshold estimation method is suitably flexible.

References

Andrews K. The effect of changes in temperature and humidity on the accuracy of von Frey hairs. J Neurosci Methods. 1993 Oct;50(1):91–3.

Aring AM, Jones DE, Falko JM. Evaluation and prevention of diabetic neuropathy. Am Fam Physician. 2005 Jun 1;71(11):2123–8.

Armstrong DG, Lavery LA, Vela SA, Quebedeaux TL, Fleischli JG. Choosing a practical screening instrument to identify patients at risk for diabetic foot ulceration. Arch Intern Med. 1998 Feb 9;158(3):289–92.

Bansal V, Kalita J, Misra UK. Diabetic neuropathy. Postgrad Med J. 2006 Feb;82(964):95–100.

Baraz S, Zarea K, Shahbazian HB, Latifi SM. Comparison of the accuracy of monofilament testing at various points of feet in peripheral diabetic neuropathy screening. J Diabetes Metab Disord. 2014 Jan;13(1):19.

Le Bars D, Gozariu M, Cadden SW. Animal models of nociception. Pharmacol Rev. 2001 Dec;53(4):597–652.

Begley CG. Six red flags for suspect work. Nature. 2013 May 23;497(7450):433–4.

Bell-Krotoski J, Tomancik E. The repeatability of testing with Semmes-Weinstein monofilaments. J Hand Surg Am. 1987 Jan;12(1):155–61.

Bell-Krotoski JA, Buford WL. The force/time relationship of clinically used sensory testing instruments. J Hand Ther. Hanley & Belfus, Inc.; 1997 Oct;10(4):297–309.

Bell-Krotoski JA, Fess EE, Figarola JH, Hiltz D. Threshold detection and Semmes-Weinstein monofilaments. J Hand Ther. 1995;8(2):155–62.

Berquin AD, Lijesevic V, Blond S, Plaghki L. An adaptive procedure for routine measurement of light-touch sensitivity threshold. Muscle Nerve. 2010 Sep;42(3):328–38.

Bi J, Ennis DM. SENSORY THRESHOLDS: CONCEPTS AND METHODS. J Sens Stud. 1996;13(1998):133–48.

Bonin RP, Bories C, De Koninck Y. A simplified up-down method (SUDO) for measuring mechanical nociception in rodents using von Frey filaments. Mol Pain. 2014 Jan;10:26.

Booth J, Young MJ. Differences in the performance of commercially available 10-g monofilaments. Diabetes Care. 2000 Jul;23(7):984–8.

Bove G. Mechanical sensory threshold testing using nylon monofilaments: the pain field's "tin standard". Pain. 2006 Sep;124(1-2):13–7.

Callahan BL, Gil ASC, Levesque A, Mogil JS. Modulation of mechanical and thermal nociceptive sensitivity in the laboratory mouse by behavioral state. J Pain. 2008 Feb;9(2):174–84.

Carandini M, Churchland AK. Probing perceptual decisions in rodents. Nat Neurosci. 2013 Jul;16(7):824–31.

Chaplan SR, Bach FW, Pogrel JW, Chung JM, Yaksh TL. Quantitative assessment of tactile allodynia in the rat paw. J Neurosci Methods. 1994 Jul;53(1):55–63.

Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. Neurosci Biobehav Rev. 2002 Dec;26(8):907–23.

Choi Y, Yoon YW, Na HS, Kim SH, Chung JM. Behavioral signs of ongoing pain and cold allodynia in a rat model of neuropathic pain. Pain. 1994 Dec;59(3):369–76.

Clevenger AC, Restrepo D. Evaluation of the validity of a maximum likelihood adaptive staircase procedure for measurement of olfactory detection threshold in mice. Chem Senses. 2006 Jan;31(1):9–26.

Cohen J. The cost of dichotomization. Applied Psych Measurement. 1983;7(3):249-253

Cornsweet TN. The Staircase-Method in Psychophysics. Am J Psychol. 1962 Sep;75(3):485.

Dell RB, Holleran S, Ramakrishnan R. Sample size determination. ILAR J. 2002;43(4):207-13.

Dellon AL, Mackinnon SE, Brandt KE. The markings of the Semmes-Weinstein nylon monofilaments. J Hand Surg Am. 1993 Jul;18(4):756–7.

Detloff MR, Clark LM, Hutchinson KJ, Kloos AD, Fisher LC, Basso DM. Validity of acute and chronic tactile sensory testing after spinal cord injury in rats. Exp Neurol. Elsevier B.V.; 2010 Oct;225(2):366–76.

Detloff MR, Fisher LC, Deibert RJ, Basso DM. Acute and Chronic Tactile Sensory Testing after Spinal Cord Injury in Rats. JoVE 2012;(62):3247.

Dixon W. The up-and-down method for small samples. J Am Stat Assoc. 1965;60(312):967–78.

Dixon W. Efficient analysis of experimental observations. Annu Rev Pharmacol Toxicol. 1980;20:441–62.

Dixon W, Mood A. A method for obtaining and analyzing sensitivity data. J Am Stat Assoc. 1948;43(241):109–26.

Duraku LS, Hossaini M, Hoendervangers S, Falke LL, Kambiz S, Mudera VC, et al. Spatiotemporal dynamics of re-innervation and hyperinnervation patterns by uninjured CGRP fibers in the rat foot sole epidermis after nerve injury. Mol Pain. Molecular Pain; 2012 Jan;8(1):61.

Dyck PJ, O'Brien PC, Kosanke JL, Gillen DA, Karnes JL. A 4, 2, and 1 stepping algorithm for quick and accurate estimation of cutaneous sensation threshold. Neurology. 1993 Aug;43(8):1508–12.

Ennis DM, Bi J. THE BETA-BINOMIAL MODEL: ACCOUNTING FOR INTER-TRIAL VARIATION IN REPLICATED DIFFERENCE AND PREFERENCE TESTS. J Sens Stud. 1998 Dec;13(4):389–412.

Fechner GT. The fundamental formula and the measurement formula. Elem der Psychophysik (Trans Langfeld, 1912). 1860.

Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. 2002;43(4):244-58.

Fisher R. On the mathematical foundations of theoretical statistics. Philos Trans R Soc London Ser A. 1922;222:309–68.

Forbes C, Evans M, Hastings N, Peacock B. Statistical Distributions. Wiley; 2011.

Foster DH, Bischof WF. Thresholds from psychometric functions: Superiority of bootstrap to incremental and probit variance estimators. Psychol Bull. 1991;109(1):152–9.

Foster DH, Bischof WF. Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions. Spat Vis. 1997;11(1):135–9.

Fruhstorfer H, Gross W, Selbmann O. von Frey hairs: new materials for a new design. Eur J Pain. 2001 Jan;5(3):341–2.

Fründ I, Haenel NV, Wichmann FA. Inference for psychometric functions in the presence of nonstationary behavior. J Vis. 2011 Jan;11(6).

Fuh C, Lee J, Liaw C. The Design Aspect of the Bruceton Test for Pyrotechnics Sensitivity Analysis. J Data Sci. 2003;1:83–101.

García-Pérez MA. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Res. 1998 Jun;38(12):1861-81.

García-Pérez MA. Yes-no staircases with fixed step sizes: psychometric properties and optimal setup. Optom Vis Sci. 2001 Jan;78(1):56-64.

García-Pérez MA, Alcalá-Quintana R. Shifts of the psychometric function: distinguishing bias from perceptual effects. Q J Exp Psychol (Hove). 2013 Jan;66(2):319–37.

Gardner E, Martin J. Coding of sensory information. In: Kandel E, Schwartz J, Jessell T, editors. Princ neural Sci. New York: McGraw-Hill; 2000. p. 411–29.

Gescheider G. Psychophysics: the fundamentals. 3, revised. Psychophys. Fundam. Psychology Press; 2013. p. 448.

Ghanouni P, Behera D, Xie J, Chen X, Moseley M, Biswal S. In vivo USPIO magnetic resonance imaging shows that minocycline mitigates macrophage recruitment to a peripheral nerve injury. Mol Pain. 2012 Jan;8:49.

Greenspan JD, McGillis SL. Stimulus features relevant to the perception of sharpness and mechanically evoked cutaneous pain. Somatosens Mot Res.1991;8(2):137-47. PubMed PMID: 1887724.

Greenspan J, Bolanowki S. Perceived intensity of nonvibratory forms of stimulation and correlated mechanoreceptive afferent response properties. In: Kruger L, editor. Pain Touch. Academic Press; 1996. p. 48–53.

Haloua MH, Sierevelt I, Theuvenet WJ. Semmes-weinstein monofilaments: influence of temperature, humidity, and age. J Hand Surg Am. Elsevier Inc.; 2011 Jul;36(7):1191–6.

Hamity MV, White SR, Hammond DL. Effects of neurokinin-1 receptor agonism and antagonism in the rostral ventromedial medulla of rats with acute or persistent inflammatory nociception. Neuroscience. 2010 Feb 3;165(3):902-13.

Harvey LO. Efficient estimation of sensory thresholds. Behav Res Methods, Instruments, Comput. 1986 Nov;18(6):623–32.

Harvey LO. Efficient estimation of sensory thresholds with ML-PEST. Spat Vis. 1997 Jan;11(1):121–8.

Hecht S, Shlaer S, Pirenne M. Energy, quanta, and vision. J Gen Physiol. 1942;819–40.

Hogan Q, Sapunar D, Modric-Jednacak K, McCallum JB. Detection of neuropathic pain in a rat model of peripheral nerve injury. Anesthesiology. 2004 Aug;101(2):476–87.

Huang Y, Luo Q-Z. A simple method to determine the critical buckling loads for axially inhomogeneous beams with elastic restraint. Comput Math with Appl. 2011 May;61(9):2510–7.

Jerosch-Herold C. Assessment of sensibility after nerve injury and repair: a systematic review of evidence for validity, reliability and responsiveness of tests. J Hand Surg Br. 2005 Jun;30(3):252–64.

Johansson RS, Vallbo a B, Westling G. Thresholds of mechanosensitive afferents in the human hand as measured with von Frey hairs. Brain Res. 1980 Feb 24;184(2):343–51.

Jones F. Some Subjective Magnitude Functions for Touch. Symp Cutan Sensit. 1960. p. 63–72.

Kauppila T, Kontinen VK, Pertovaara A. Weight bearing of the limb as a confounding factor in assessment of mechanical allodynia in the rat. Pain. 1998 Jan;74(1):55–9.

Keizer D, Fael D, Wierda JMKH, van Wijhe M. Quantitative sensory testing with Von Frey monofilaments in patients with allodynia: what are we quantifying? Clin J Pain. 2008 Jun;24(5):463–6.

Kim KJ, Yoon YW, Chung JM. Comparison of three rodent neuropathic pain models. Exp Brain Res. 1997 Feb;113(2):200–6.

Kim SH, Chung JM. An experimental model for peripheral neuropathy produced by segmental spinal nerve ligation in the rat. Pain. 1992 Sep;50(3):355–63.

Kingdom F, Prins N. Psychophysics: a practical introduction. Academic Press; 2010.

Klein SA. Measuring, estimating, and understanding the psychometric function: a commentary. Percept Psychophys. 2001 Nov;63(8):1421–55.

Kuss M, Jäkel F, Wichmann F. Bayesian inference for psychometric functions. J Vis. 2005;5(5):478–92.

Lambert G a, Mallos G, Zagami AS. Von Frey's hairs--a review of their technology and use--a novel automated von Frey device for improved testing for hyperalgesia. J Neurosci Methods. 2009 Mar 15;177(2):420–6.

Lavery LA, Lavery DE, Lavery DC, Lafontaine J, Bharara M, Najafi B. Accuracy and durability of Semmes-Weinstein monofilaments: what is the useful service life? Diabetes Res Clin Pract. 2012 Sep;97(3):399–404.

Lawless HT, Heymann H. Physiological and Psychological Foundations of Sensory Function. Sens Eval Food. New York: Springer New York; 2010. p. 19–56.

Levin S, Pearsall G, Ruderman RJ. Von Frey's method of measuring pressure sensibility in the hand: an engineering analysis of the Weinstein-Semmes pressure aesthesiometer. J Hand Surg Am. 1978 May;3(3):211–6.

Linschoten MR, Harvey LO, Eller PM, Jafek BW. Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. Percept Psychophys. 2001 Nov;63(8):1330–47.

Llewelyn J. The diabetic neuropathies: types, diagnosis and management. J Neurol Neurosurg Psychiatry. 2003;74(II):15–9.

Lucas K. Quantitative researches on the summation of inadequate stimuli in muscle and nerve, with observations on the time-factor in electric excitation. J Physiol. 1910 Mar 8;39(6):461-75.

Magee JC. 2000 Nat Rev Neurosci. 2000 Dec;1(3):181-90. Dendritic integration of excitatory synaptic input.

Malmberg AB, Chen C, Tonegawa S, Basbaum AI. Preserved acute pain and reduced neuropathic pain in mice lacking PKCgamma. Science. 1997 Oct 10;278(5336):279-83.

Marin AB, Barnard J, Darlington RB, Acree TE. SENSORY THRESHOLDS: ESTIMATION FROM DOSE-RESPONSE CURVES. J Sens Stud. 1991 Dec;6(4):205–25.

Massy-Westropp N. The effects of normal human variability and hand activity on sensory testing with the full Semmes-Weinstein monofilaments kit. J Hand Ther. 2002;15(1):48–52.

Metz AE, Yau H-J, Centeno MV, Apkarian AV, Martina M. Morphological and functional reorganization of rat medial prefrontal cortex in neuropathic pain. Proc Natl Acad Sci U S A. 2009 Feb 17;106(7):2423–8.

Millecamps M, Centeno M V, Berra HH, Rudick CN, Lavarello S, Tkatch T, et al. D-cycloserine reduces neuropathic pain behavior through limbic NMDA-mediated circuitry. Pain. 2007 Nov;132(1-2):108–23.

Miller J, Ulrich R. A computer program for Spearman-Kärber and probit analysis of psychometric function data. Behav Res Methods Instrum Comput. 2004 Feb;36(1):11–6.

Milligan E, Zapata V, Schoeniger D, Chacur M, Green P, Poole S, et al. An initial investigation of spinal mechanisms underlying pain enhancement induced by fractalkine, a neuronally released chemokine. Eur J Neurosci. 2005 Dec;22(11):2775–82.

Milligan ED, Maier SF, Watkins LR. Sciatic inflammatory neuropathy in the rat: surgical procedures, induction of inflammation, and behavioral testing. Methods Mol Med. 2004 Jan;99:67–89.

Milligan ED, Mehmert KK, Hinde JL, Harvey LO, Martin D, Tracey KJ, et al. Thermal hyperalgesia and mechanical allodynia produced by intrathecal administration of the human immunodeficiency virus-1 (HIV-1) envelope glycoprotein, gp120. Brain Res. 2000 Apr 7;861(1):105–16.

Mills C, Leblond D, Joshi S, Zhu C, Hsieh G, Jacobson P, et al. Estimating efficacy and drug ED50's using von Frey thresholds: impact of weber's law and log transformation. J Pain. Elsevier Ltd; 2012 Jun;13(6):519–23.

Mogil JS. Animal models of pain: progress and challenges. Nat Rev Neurosci. 2009 Apr;10(4):283–94.

Morita T, Kang H, Wolfe J, Jadhav SP, Feldman DE. Psychometric curve and behavioral strategies for whisker-based texture discrimination in rats. PLoS One. 2011 Jan;6(6):e20437.

Mountcastle V. The problem of sensing and the neural coding of sensory events. Neurosci A study Progr. The Rockefeller University Press; 1967. p. 393–408.

Mueller MJ. Identifying patients with diabetes mellitus who are at risk for lower-extremity complications: use of Semmes-Weinstein monofilaments. Phys Ther. 1996 Jan;76(1):68–71.

Muniak MA, Ray S, Hsiao SS, Dammann JF, Bensmaia SJ. The neural coding of stimulus intensity: linking the population response of mechanoreceptive afferents with psychophysical behavior. J Neurosci. 2007 Oct 24;27(43):11687-99. PubMedPMID: 17959811.

Myung IJ. Tutorial on maximum likelihood estimation. J Math Psychol. 2003 Feb;47(1):90–100.

National Diabetes Education Program. Feet can last a lifetime: A Health Care Provider's Guide to Preventing Diabetes Foot Problems. 2000.

Nirogi R, Goura V, Shanmuganathan D, Jayarajan P, Abraham R. Comparison of manual and automated filaments for evaluation of neuropathic pain behavior in rats. J Pharmacol Toxicol Methods. Elsevier Inc.; 2012 Jul;66(1):8–13.

Norrsell U, Finger S, Lajonchere C. Cutaneous sensory spots and the "law of specific nerve energies": history and development of ideas. Brain Res Bull. 1999 Mar 15;48(5):457–65.

Oldfather W, Ellis C, Brown D. Leonhard Euler's elastic curves. Isis. 1933;20(1):72–160.

Peirce JW. PsychoPy--Psychophysics software in Python. J Neurosci Methods. 2007 May 15;162(1-2):8–13.

Pentland A. Maximum likelihood estimation: The best PEST. Attention, Perception, Psychophys. 1980;28(4):377–9.

Pitcher GM, Ritchie J, Henry JL. Paw withdrawal threshold in the von Frey hair test is influenced by the surface on which the rat stands. J Neurosci Methods. 1999 Mar;87(2):185–93.

Pogatzki EM, Gebhart GF, Brennan TJ. Characterization of Adelta- and C-fibers innervating the plantar rat hindpaw one day after an incision. J Neurophysiol. 2002 Feb;87(2):721-31.

Prins N. The psychometric function: the lapse rate revisited. J Vis. 2012 Jan;12(6).

Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009 Apr;6(4):257–61.

Rigaud M, Gemes G, Abram SE, Dean C, Hopp FA, Stucky CL, Eastwood D, Tarima S, Seagard J, Hogan QH. Pain tests provoke modality-specific cardiovascular responses in awake, unrestrained rats. Pain. 2011 Feb;152(2):274-84.

Rolke R, Baron R, Maier C, Tölle TR, Treede R-D, Beyer A, et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain. 2006 Aug;123(3):231–43.

Savage C. The Measurement of Sensation: A Critique of Perceptual Psychophysics. University of California Press; 1970.

Seltzer Z, Dubner R, Shir Y. A novel behavioral model of neuropathic pain disorders produced in rats by partial sciatic nerve injury. Pain. 1990;43:205–18.

Shen Y. Comparing adaptive procedures for estimating the psychometric function for an auditory gap detection task. Atten Percept Psychophys. 2013 May;75(4):771-80.

Smieja M, Hunt DL, Edelman D, Etchells E, Cornuz J, Simel DL. Clinical examination for the detection of protective sensation in the feet of diabetic patients. J Gen Intern Med. 1999 Jul;14(7):418–24.

Song XJ, Hu SJ, Greenquist KW, Zhang JM, LaMotte RH. Mechanical and thermal hyperalgesia and ectopic neuronal discharge after chronic compression of dorsal root ganglia. J Neurophysiol. 1999 Dec;82(6):3347–58.

Sorge RE, Martin LJ, Isbester KA, Sotocinal SG, Rosen S, Tuttle AH, et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat Methods. 2014 Jun;11(6):629–32.

De Sousa MVP, Ferraresi C, de Magalhães AC, Yoshimura EM, Hamblin MR. Building, testing and validating a set of home-made von Frey filaments: A precise, accurate and cost effective alternative for nociception assessment. J Neurosci Methods. 2014 Jul 30;232:1–5.

Stevens SS. On the psychophysical law. Psychol Rev. 1957;64(3):153–81.

Stevens SS. The psychophysics of sensory function. Am. Scientist 1960; 48, 226– 253

Stoetling Co. Touch Test ™ Sensory Evaluators - Semmes Weinstein Von Frey Aesthesiometers - Operation Manual. Wood Dale, Illinois: Stoetling Co.; 2001.

Strasburger H. Converting between measures of slope of the psychometric function. Percept Psychophys. 2001 Nov;63(8):1348–55.

Takaishi K, Eisele JH, Carstens E. Behavioral and electrophysiological assessment of hyperalgesia and changes in dorsal horn responses following partial sciatic nerve ligation in rats. Pain. 1996 Aug;66(2-3):297–306.

Tena B, Escobar B, Arguis MJ, Cantero C, Rios J, Gomar C. Reproducibility of Electronic Von Frey and Von Frey monofilaments testing. Clin J Pain. 2012 May;28(4):318–23.

Tiest WMB, Kappers AML. An antisymmetric psychometric function on a logarithmic scale. Perception. 2011 Jan;40(1):99–100.

Titchener EB. Vol. 2, Pt. 2: Instructor's manual. Exp Psychol a Man Lab Pract. London: MACMILLAN & CO Ltd.; 1915. p. p. 417–422.

Treede RD, Rolke R, Andrews K, Magerl W. Pain elicited by blunt pressure: neurobiological basis and clinical relevance. Pain. 2002 Aug;98(3):235-40.Review. PubMed PMID: 12127024.

Trossarelli L. The History of Nylon [Internet]. 2010. Available from: http://www.caimateriali.org/index.php?id=32

Del Valle ME, Cobo T, Cobo JL, Vega J a. Mechanosensory neurons, cutaneous mechanoreceptors, and putative mechanoproteins. Microsc Res Tech. 2012 Aug;75(8):1033–43.

Vivancos GG, Verri W a, Cunha TM, Schivo IRS, Parada C a, Cunha FQ, et al. An electronic pressure-meter nociception paw test for rats. Braz J Med Biol Res. 2004 Mar;37(3):391–9.

Voerman VF, van Egmond J, Crul BJ. Normal values for sensory thresholds in the cervical dermatomes: a critical note on the use of Semmes-Weinstein monofilaments. Am J Phys Med Rehabil. 1999;78(1):24–9.

Vrinten DH, Hamers FFT. "CatWalk" automated quantitative gait analysis as a novel method to assess mechanical allodynia in the rat; a comparison with von Frey testing. Pain. 2003 Mar;102(1-2):203–9.

Wallas TR, Winterson BJ, Ransil BJ, Bove GM. Paw withdrawal thresholds and persistent hindlimb flexion in experimental mononeuropathies. J Pain. 2003 May;4(4):222–30.

Wallis S a, Baker DH, Meese TS, Georgeson M a. The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision. Vision Res. Elsevier Ltd; 2013 Jan 14;76:1–10.

Watson AB. Probability summation over time. Vision Res. 1979 Jan;19(5):515–22.

Watson AB, Pelli DG. QUEST: a Bayesian adaptive psychometric method. Percept Psychophys. 1983 Feb;33(2):113–20.

Weber E. Der Tastsinn und das Gemeingefühl “The Sense of Touch and the Common Sensibility.”1846;

Weibull W. A statistical distribution function of wide applicability. J Appl Mech. 1951;293–7.

Weiner IB, Healy AF, Proctor RW. Handbook of Psychology, Experimental Psychology, Volume 4 of Handbook of Psychology. 2nd ed. John Wiley & Sons; 2012.

Weinstein S. Fifty years of somatosensory research. J Hand Ther. Hanley & Belfus, Inc.; 1993 Jan;6(1):11–22.

Werner MU, Rotbøll-Nielsen P, Ellehuus-Hilmersson C. Humidity affects the performance of von Frey monofilaments. Acta Anaesthesiol Scand. 2011 May;55(5):577–82.

Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. Percept Psychophys. 2001 a Nov;63(8):1293–313.

Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence intervals and sampling. Percept Psychophys. 2001 b Nov;63(8):1314–29.

Wu HE, Gemes G, Zoga V, Kawano T, Hogan QH. Learned avoidance from noxious mechanical simulation but not threshold semmes weinstein filament stimulation after nerve injury in rats. J Pain. 2010 Mar;11(3):280-6.

Yong R, Karas TJ, Smith KD, Petrov O. The durability of the Semmes-Weinstein 5.07 monofilament. J Foot Ankle Surg. American College of Foot and Ankle Surgeons; 2000 Jan;39(1):34–8.

Young-Xu Y, Chan KA. Pooling overdispersed binomial data to estimate event rate. BMC Med Res Methodol. 2008 Jan;8:58.

Zychaluk K, Foster DH. Model-free estimation of the psychometric function. Atten Percept Psychophys. 2009 Aug;71(6):1414–25.

Table1

Table 1. Comparison of published and measured stimulus strength for complete sets of Semmes-Weinstein monofilaments. Our measurement is mean of 10 repetitions (method after Levin *et al.,* 1978) performed by one operator (MB); repetitions showed homoskedasticity. In a comparison of co-authors (MB, FF, CS) calibrating hairs 2.36 through 4.17, mean difference between us was 1.33 % (data not shown). Exponential series are shown in grey text; all other values are properly comparable (see Mills et al., 2012 on the importance of using log values).

Theoretical or supposed values					Observed values	
Ordinal: full set filaments	Target force g	Pressure calculated by Dellon (1993) g/mm2	Handle mark (not equivalent to target force)	Target force converted Equation 2	Measured by Levin (1978), converted Equation 2	Measured by us, converted Equation 2
1	0.008	1.5	1.65	1.903	1.602	1.893
2	0.02		2.36	2.301	1.973	2.340
3	0.04	3.3	2.44	2.602	2.531	2.621
4	0.07	4.9	2.83	2.845	2.959	2.876
5	0.16		3.22	3.204	3.049	3.202
6	0.4	17.7	3.61	3.602	3.328	3.575
7	0.6		3.84	3.778	3.750	3.718
8	1.0		4.08	4.000	3.990	3.984
9	1.4		4.17	4.146	4.199	4.108
10	2.0	33.1	4.31	4.301	4.267	4.295
11	4.0		4.56	4.602	4.449	4.576
12	6.0		4.74	4.778	4.497	4.724
13	8.0	60.9	4.93	4.903	5.025	4.861
14	10.0		5.07	5.000	5.230	4.935
15	15.0		5.18	5.176	5.270	5.119
16	26.0	107	5.46	5.415	5.348	5.342
17	60.0		5.88	5.778	5.865	5.676
18	100	243	6.1	6.000	5.937	5.859
19	180		6.45	6.255		6.132
20	300	439	6.65	6.477		Off scale

Table 2. Summary of Recommendations

Recommendation	***Advantage***
Operators should be blinded to experimental group	*Reducing confirmation bias*
Operator ideally constant per experiment	*Reducing variability due to operator sex, odour, and experience*
Remove 9th and 13th hairs (marked 4.17 and 4.93) from standard von Frey set	*Reducing bias of poor step selection (especially to assist Chaplan's method)*
Standardise time per application (2 sec in recent clinical guidelines)	*Assisting comparability within and between datasets*
Test by a staircase method, starting from mid-range (e.g. hair 3.22 mice, 4.31 rats).	*Concentrates on dynamic region, optimising contribution of each stimulus*
Selection of adequate staircase length and "adaptive testing" can be flexible	*Preventing excessive testing; avoiding dogmatism as thresholds become clear*
Thresholds should be calculated and analysed using log stimulus values	*Avoiding bias across the range*
Use suitable software (e.g. Psychofit) for curve fit and threshold estimation	*Flexible analysis using all available data and minimising noise*
Where possible, pool responses/non-responses per datapoint (e.g. per animal)	*Improves curve fit without masking true variance*

Figure1

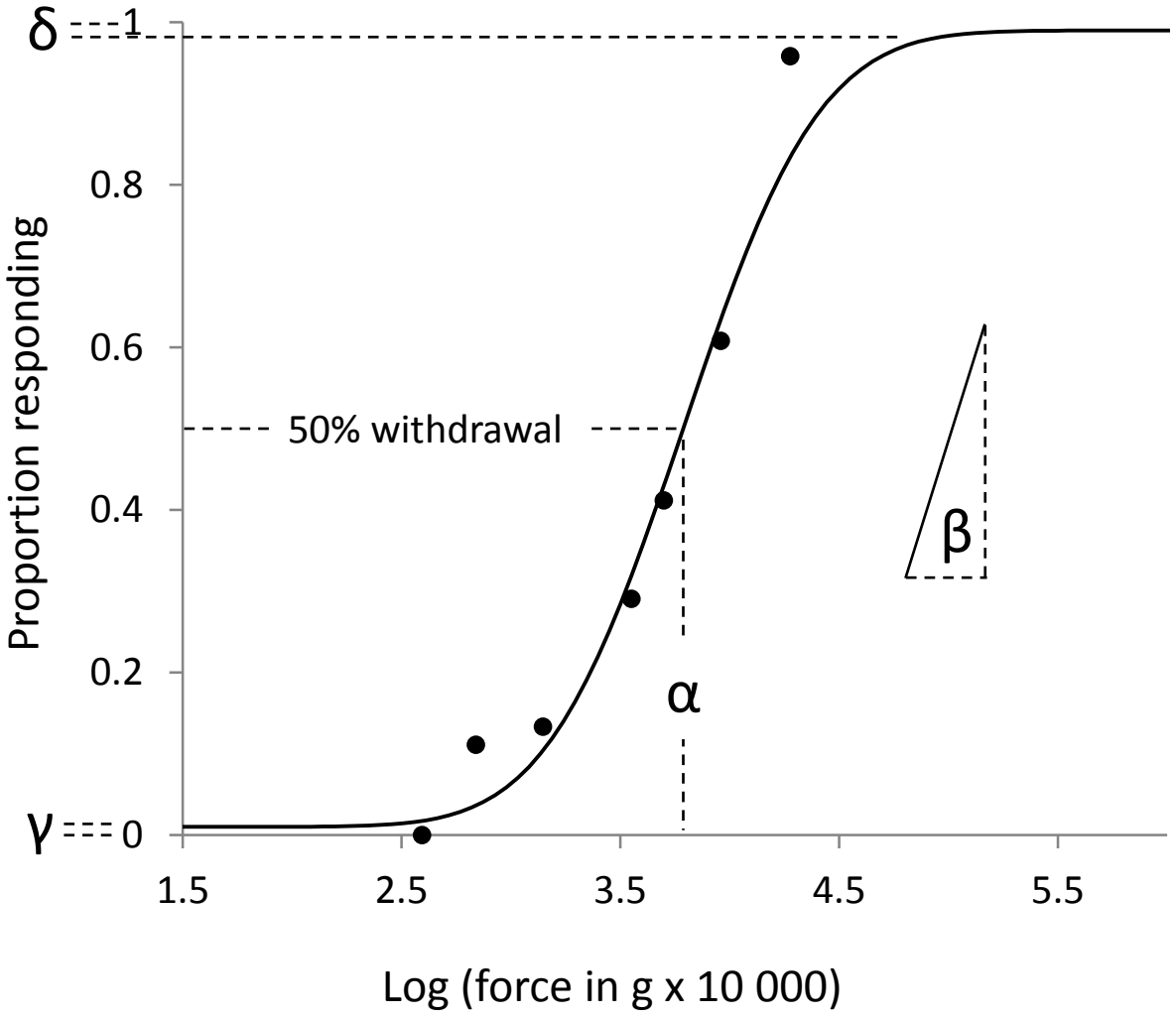

δ
1
0.8
0.6
0.4
0.2
γ
0
Proportion responding
50% withdrawal
α
β
1.5
2.5
3.5
4.5
5.5
Log (force in g x 10 000)

Figure2

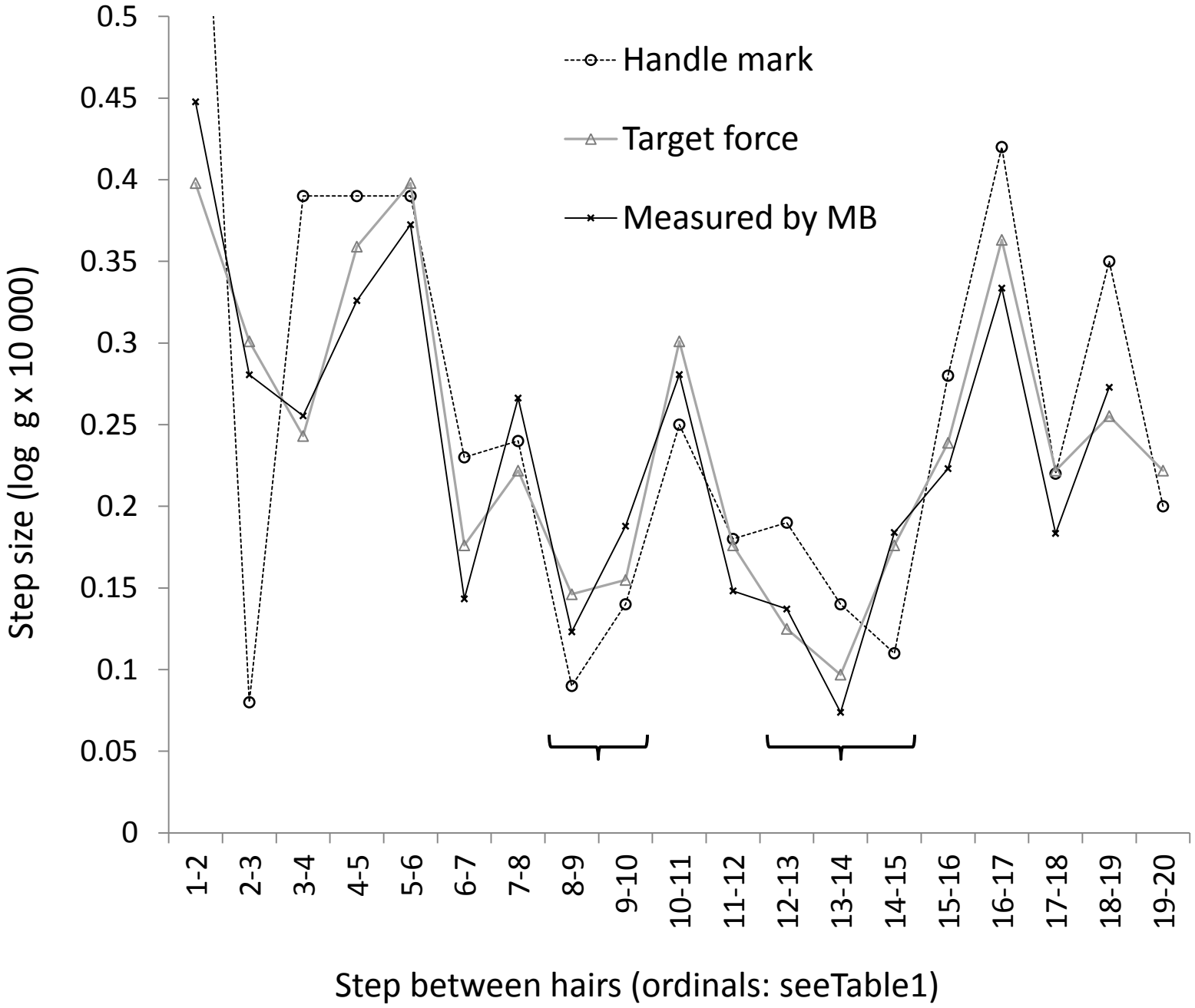
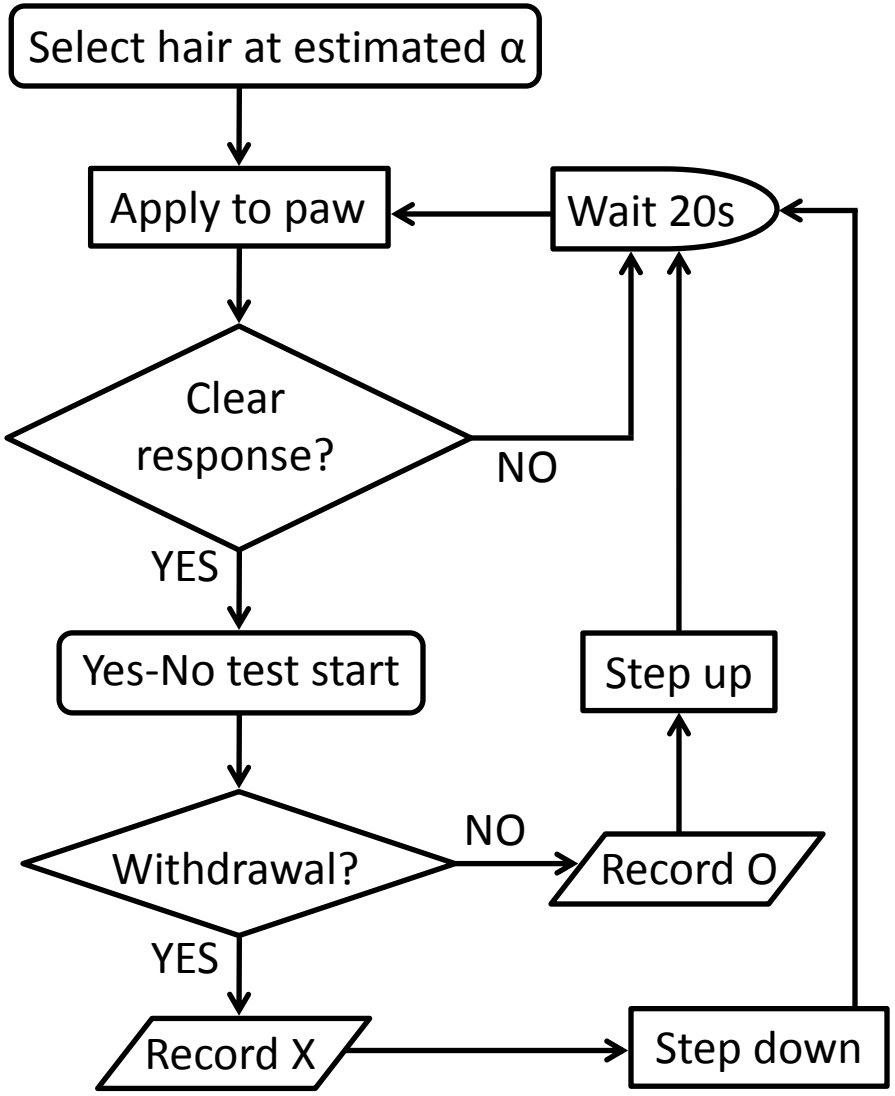

Handle mark
Target force
Measured by MB
Step size (log g x 10 000)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
10-11
11-12
12-13
13-14
14-15
15-16
16-17
17-18
18-19
19-20
Step between hairs (ordinals: seeTable1)

Figure3


Select hair at estimated α
Apply to paw
Wait 20s
Clear response?
NO
YES
Yes-No test start
Step up
Withdrawal?
NO
Record O
YES
Record X
Step down