DISCREDITING THE “DISCREDITING” OF PSYCHOPHYSICS: H.K. BEECHER VERSUS THE HARDY-WOLFF-GOODELL DOLORIMETER

Lance Nizami and Claire S. Barnes
Independent Research Scholars, Bedford, MA 01730 lancenizamiphd@comcast.net

pp. 94-99, Fechner Day 2022: Proceedings of the 38th Annual Meeting of the International Society for Psychophysics, Lund, Sweden, 23-26 August 2022

Abstract

In 1947, Hardy, Wolff, and Goodell achieved a psychophysics milestone: they built a putative sensation-growth scale, for skin pain, from pain-difference limens. Limens were found using the “dolorimeter”, a device first made by Hardy & co. to evoke pain for pain-threshold measurements. Scant years later, though, H.K. Beecher (MD) discredited the pain scale – according to Paterson (2019), citing the historian Tousignant. Yet Hardy & co. receive approval in the literature. Intrigued, we scrutinized their methods, then Beecher’s critiques, and Tousignant’s history of threshold dolorimetry. Beecher decried dolorimetry as irrelevant, favoring clinical trials of pain relief. *But he failed to discredit dolorimetry.*

Writing in the journal *Body & Society*, Paterson (2019) emphasizes that pain is distinct from touch. Paterson also emphasizes the philosophy that *internal* pain, not to be confused with skin pain, is part of a homeostasis mechanism. Paterson (p. 118) notes that a scale for *skin* pain was constructed by Hardy, Wolff, and Goodell in the 1940’s (citations below); but subsequently, “The Hardy–Wolff–Goodell scale was discredited in 1957 by the distinguished clinical anaesthesiologist Henry K. Beecher, who became Chair of the Committee on Drug Addiction and Narcotics and who established a Laboratory of Anaesthesia (Tousignant, 2014: 122)”. No further details emerge. The cited Tousignant (2014) is a book chapter, derived from a lengthier peer-reviewed paper (Tousignant, 2011) which was a history of the use of a skin-pain-evoking device built and used by Hardy et al. – the very same device used to construct their “discredited” pain scale. But Tousignant denies “discrediting” that pain scale (personal communication, 1 April 2022). Nonetheless, Tousignant’s (2011) review does indeed (and prominently) feature the above-mentioned Henry K. Beecher, MD. His numerous publications sometimes concerned military medicine, but not psychophysics. How, then, did Beecher discredit the psychophysics of Hardy et al.? Here, we describe their psychophysics, and its clinical use, and Beecher’s attempts to dismiss it in favor of clinical *trials*. We begin with Hardy et al.’s pain thresholds.

Measuring the threshold for pain, with and without pain reduction by drugs

Establishing a pain-detection threshold

Hardy et al. (1940) wished to evoke controlled pain in human skin. Historically, the stimuli were “mechanical, chemical, electrical, and thermal” (Hardy et al., p. 649). Hardy et al. favored thermal, by radiation, thanks to (1) relatively easy construction of equipment, (2) accurate measurement of stimulus intensity, (3) “sharply defined” sensory thresholds (Hardy et al., p. 649), (4) manipulability of exposure duration and of skin condition, (5) ability to choose skin patches (regardless of texture) of different size, and (6) “The stimulus can be repeated in rapid succession without injury to the skin surface tested” (p. 649). Also (p. 649), “The sensation produced is sharp, a “bright pain”, and is to be distinguished from an ache or deep pain”.

Hardy et al. (p. 650) described the apparatus and method as follows: “The light from a 1000 watt lamp was focussed [*sic*] by a condensing lens through a fixed aperture onto the blackened forehead of the subject. The surface of the forehead to be tested was thoroughly blackened with India ink. This measure was taken to insure total absorption of the radiation, regardless of pigmentation of the skin, and to eliminate possible effects due to the penetration of the rays below the skin surface. The stimulus could thus be considered as purely thermal". Further (Hardy et al., p. 650), "The intensity of the radiation was controlled by means of a rheostat. Immediately in front of the lamp was mounted an automatic shutter, which was arranged to allow the radiation to pass through to the subject for exactly 3 seconds". Pain threshold was then established by exposing 3.5 cm^2 of blackened forehead-skin. If the subject felt no pain, then the light intensity was increased, the exposure being repeated after 30-60 sec. As Hardy et al. (p 650) note, "This procedure was followed until the subject just felt pain at the end of the exposure. This threshold pain was easily recognizable even by untrained subjects. The sensation was that of heat finally "swelling" to a distinct, sharp stab of pain at the end". A radiometer was placed where the subject's forehead had been, to measure the pain-threshold radiation intensity (expressed in gram calories per second per square centimeter, abbreviated gm. cal./sec./$cm.^2$). Remarkably, half of the thresholds agreed within 2% over re-tests (Hardy et al., p. 651); here and elsewhere, Hardy, Wolff, and Goodell were their own research subjects.

Measuring skin-pain-threshold elevation by analgesics

Hardy, Wolff, and Goodell immediately applied their pain-threshold-detection method to assessing "the action of analgesic agents". Their motivation: "No adequate method for assaying their effects on the pain threshold in man has been available. However, since the prime purpose of an analgesic drug concerns its action in man, it is desirable to measure accurately its effect on man's pain threshold" (Wolff et al., 1940, p. 659). Note well that Hardy, Wolff, and Goodell were doing classic *psychophysics* on healthy volunteers (themselves, using their above methods); they were not doing *pharmacology* with clinical patients. They established the "control" pain threshold; then, "an analgesic agent was administered [intramuscularly] and observations of the pain threshold were made at 10-minute intervals until the threshold had returned to the control level, that is, until all pain threshold-raising action had ceased. The height of the pain threshold-raising effect was expressed in per cent elevation above the control level" (Wolff et al., 1940, p. 659). The latter elevation was plotted versus time for a "time-action curve". These were obtained for opioids, with special attention to morphine sulphate; its time-action curve rose, peaked, and fell, reaching greater height (i.e., pain relief) with greater dosage. Supplementary doses prolonged the relief. Psychophysics had proven useful to the pain clinic.

Measuring skin-pain-threshold elevation by analgesics during "internal" pain

Hardy and colleagues then embarked upon a far more dangerous program. That is, "prolonged pain was introduced as a variable since in this way the action of morphine could be appraised more nearly in terms of its common therapeutic use" (Wolff et al., 1940, p. 672). Three methods were used: inflating a blood-pressure cuff to 200 mm Hg pressure on the upper arm, or "swallowing a catheter to which was attached a balloon which was distended with water when it reached the duodenum" (Wolff et al., p. 672), or clamping the trapezius and biceps muscles. All three treatments lasted 40 *minutes*, and each produced a different kind of "deep, aching pain". The actual administration of morphine sulphate and the pain-threshold measurements (radiation on the forehead, as above) proceeded thus: "After the control readings, which preceded the morphine injection, the painful procedure was begun: (1) 46 minutes before injection; (2) 1 minute after the injection; (3) 50 minutes after the injection; (4) 120 minutes after the injection. Pain-threshold readings were made every 10 minutes throughout the subsequent 6 to 7 hours" (Wolff et al., p. 672). Compared to the case *without* deep aching pain, whose time-action plot lasted 7 hours, the general finding was that 0.015 grams of morphine sulphate caused the time-action plot to remain the same for case (1), to peak sooner and lower but to decline faster for case (2), to generally shrink for case (3), and to peak at the same level but decline faster for case (4). (These changes are difficult to picture, but there is no space for pictures.) Hardy and colleagues had shown that analgesia could be quantified psychophysically.

Building a scale for pain from just-noticeable pain differences

The just-noticeable difference (jnd) in sensation (*difference limen*) is a standard psychophysical measure. In principle, jnds of warmth, heat, and pain are empirically obtainable. But Paterson (2019, p. 118) disparages jnds of pain as "an arbitrary pain intensity unit". Units imply a *scale*.

Herget et al. (1941) used the Hardy-Wolff-Goodell apparatus to obtain jnds. They employed a two-alternative two-interval method-of-limits, alternating 2 sec. light-exposures between the right and left halves of the blackened area of the forehead, keeping one side at a fixed intensity (0-35,000 $\times 10^{-5}$ gm. cal./sec./cm.^{2}) while increasing the illumination of the other side step-wise until the subject could just detect a difference between left and right. The research subjects were one or more of Herget et al. themselves (otherwise unclear). The jnds (ΔI) follow three stages, each consisting of a rise that decelerated to a plateau. Herget et al. (p. 651) labeled the stages for lower, middle, and high intensities respectively as "warmth" ("mild, pleasant, diffuse"), "heat" ("sharper and sometimes stinging"), and "pain" ("sharp, biting, and granular").

Using the same equipment, but a somewhat different method, Hardy et al. (1947) carefully obtained jnds for pain. Notably, "In the series of experiments with stimuli greater than 500 millical./sec./cm.^{2} [units of 0.001 gram cal./sec./cm.^{2}], considerable tissue damage was produced. For this reason, a second test area, the blackened volar surface of the forearm, was chosen. This area had the same pain threshold as the forehead and was more easily cared for when blistered" (Hardy et al., p. 1153). The discovered jnds (ΔI) were approximately constant for 220-320 millical./sec./cm.^{2}, then increased roughly monotonically over 320-680 millical./sec./cm.^{2}, the last illumination providing pain saturation. Hardy et al. found 21 jnds between threshold and saturation. They then adopted Fechner's classic postulate, that each jnd represents an equal sensation change. Thus, plotting jnd count from 0 to 21 versus the radiation intensity produced a pain-growth plot. It became a pain-growth *scale*, a psychophysics ideal, when Hardy et al. declared that any two adjacent pain jnds constituted one pain unit, the *dol*.

Hardy, Wolff, and Goodell encounter Henry K. Beecher

The Hardy-Wolff-Goodell apparatus was dubbed the "dolorimeter". Tousignant (2011) details the history of its clinical use; she reports that "By 1950, over twenty research teams had published data generated by a dolorimeter; it had entered laboratories across the United States, and traveled to Britain and Canada" (Tousignant., p. 147). This popularity was highest in the late 1940's during a "burst of commercial interest in synthetic analgesics" (Tousignant, p. 164). However, "From 1950, the number of articles reporting the evaluation of analgesic drugs with the dolorimeter – its most popular usage – leveled off. In 1953, the validity of dolorimetric analgesic tests was under attack in the pages of *Science*" (Tousignant, p. 148).

Indeed it was. The attacker was Henry K. Beecher MD. His offensive started in 1952, in a six-page salvo promoting clinical-trials research. Beecher (1952, p. 159) stated that "During work on pain in 1947, we were led to postulate that there is a fundamental difference in what can be learned in studying "natural" pain which arises in a pathological focus (disease or trauma are defined here as "natural" cause) from that produced experimentally (heat to forehead, pin pricks, electric shocks, or heat to teeth, pain deliberately produced with a tourniquet, and so on). The basis for this postulate had its beginning in our attempts to use the Hardy-Wolff-

Goodell technique". Such attempts had evoked some complaints. As Tousignant (2011, p. 153) notes, "Although Hardy, Wolff, and Goodell initially claimed that their instrument worked with untrained subjects, they later admitted the importance of practice and instruction (they used the term familiarization rather than training) to obtain more consistent pain thresholds". Further (Tousignant, p. 165), "Even at the height of its popularity, few denied that using the Hardy–Wolff–Goodell method required adjustments and compromises. It was found to lack sensitivity to the effects of weaker analgesics such as aspirin. A more significant issue was the difficulty in replicating the consistency in threshold values obtained by Hardy, Wolff, and Goodell".

Beecher (1952, p. 159) opined that "It requires little imagination to suppose that the sickbed of the patient in pain, with its ominous threat against his happiness, his security, his very life, provides an entirely different milieu (*and reaction*) than the laboratory, with its dispassionate and unemotional atmosphere" (original italics). Of course, we might dispute whether laboratory-induced pain involves an unemotional atmosphere. But back to Beecher (p. 159): "In the experimental pain experience, the relatively short duration of the stimulation and the experimental situation make the experience primarily one of pain sensation. We do not believe that the pathological pain situation with all the diffuse associations of illness, disease, and pain can be satisfactorily reproduced in the laboratory". Hardy and his co-authors had already evoked pain in themselves that was pseudo-pathological, lacking "the diffuse associations" (above). Beecher (1952, p. 160) further declared that "Since pain is almost always a consequence of disease or pathological trauma, the study of pathologic pain seems to us the more direct and logical approach to an understanding of the pain experience and its relief".

Hardy, Wolff, and Goodell respond to Beecher

Hardy, Wolff, and Goodell (1953, p. 165) responded promptly: "Pain *sensation* is far more difficult to investigate when an individual is extremely frightened, inattentive, obtunded, prostrated, "sick," or exhausted. On the other hand, these would be ideal circumstances for the assay of an agent designed to make the patient "more comfortable." The bedside method is the only one that will ultimately establish whether a given analgesic has a place in clinical medicine. On the other hand, the separately studied effects of an agent on the pain threshold, pain intensity, and reactions to noxious stimulation, local and general, are of vital interest to the investigator and therapist" (original italics). In other words, psychophysics could co-exist with clinical medicine. Beecher (1953a, p. 166) replied: "It is plain that the reaction of the man in a sickbed, where his pain may be a warning of disaster, will not be the same as the reaction of a well and comfortable man in the laboratory subject to a momentary pricking sensation", emphasizing that "it seems to us reasonable to separate pain on the basis of its origins and significance to the subject; that is, experimental or pathological". But the two pains could be endured simultaneously (Wolff et al., 1940), on which Beecher (1952, 1953a) had no comment.

Beecher's approach: precision through population

How Beecher measured pain

Later in 1953, Beecher re-stated his preferred approach to measuring pain: "Measurement of pain depends upon how much analgesic is required *to relieve the pain*. To be sure this is indirect, but no more so than determination of the acidity of a solution by the quantity of standard alkali used to neutralize it. We depend upon average pain, defined as the *average response elicited from 25 or more individuals in pain* (postoperative)" (Beecher, 1953b, p. 323; italics added). Tousignant (2011, p. 152) explains: "Beecher's subjects – postoperative patients – were asked to estimate the relief of their pain as "none, slight, moderate or complete"; later they would be asked only to choose between more than 50 percent relieved or not". Note well that psychophysical measurement of pain *threshold* of laboratory staff is now replaced by expressions of pain *relief* amongst a cohort of patients – a cohort, not only because of the crudeness of the data, but also because, amongst other things, "Sound design of the experiment requires that willing, cooperative, undistracted subjects be used in sufficient numbers to cancel out normal mood swings above and below par. The body's diurnal temperature swings, with their demonstrable effects on performance, also require controls" (Beecher, 1952, p. 160).

Tousignant synopsizes (2011, p. 172) the data analysis: "Statistical analysis, standardized observation procedures, and access to large amounts of subjects were essential given Beecher's conceptualization of the experimental subject as a collective one, whose main virtue was abundance rather than stability, certainty, or detachment. The precision and accuracy of the analgesic clinical trial did not rely on individual judgments of pain relief but on their aggregation". Of course, "more" is not "more credible"; a study's credibility depends upon the credibility of each outcome, not their sheer number. But Beecher (1952, pp. 160-161) declared that "Mathematical validation of any supposed [treatment-caused] differences is essential". This is ignorance. Math cannot unequivocally validate differences (e.g., Rozeboom, 1960; Dunnette, 1966; Abbott, 2013; Smith, 2018; Nizami, 2019; Kuorikoski, 2021); results deemed "significant" may fail attempts at replication, particularly in psychology (e.g., Yong, 2018).

Complications in clinical trials

Beecher (1952, p. 160) rationalized clinical trials, so: "We are concerned incidentally, of course, with simplicity. A method that can function with no apparatus other than a notebook and pencil is manifestly more desirable and more broadly useful, other things being equal, than one that requires complex and delicate apparatus which needs calibration by a well-trained physicist". But the replacement of psychophysical measures by clinical ones brought new complications. Now, the drug-giver had to be ignorant of what was given – whether it was the active ingredient or merely a placebo, used as a control. Placebos, too, caused complications: "We are interested in studying the pharmacology of a new drug. We try it out on a group of patients; a third to a half of this group will be relieved of their symptoms by a placebo; they react favorably to the syringe regardless of what it contains. Thus they dilute the significant data derived from the other half or two thirds of the group that react only to the drug contained in the syringe. We are not, in studying a new drug, interested in the pharmacology of syringes; we are nonetheless obliged to take into account the placebo reactors; *we must screen them out* …" (Beecher, 1952, p. 161; italics added). And stimulus presentation had to be randomized. Further, treatment outcomes could depend upon who administered the substances, and who recorded their effects!

Not surprisingly, then, "the analgesic clinical trial was more expensive, time-consuming, and difficult to coordinate than the dolorimetric method. Using simple interrogation – "fairly primitive questions" as one of Beecher's colleagues would later say – to obtain comparisons of pain-relieving efficacy required a large team of observers, subjects, and consultants to collect and manipulate information under appropriate conditions, and on a sufficiently large scale. Observers and statistical experts needed to be paid salaries and consultant fees; while recruiting subjects required the authority needed to access a clinical setting, coordinate clinical staff, and oversee patient treatment" (Tousignant, 2011, pp. 156-157). All this is true today.

Conclusions

Paterson (2019, p. 118) states that "The Hardy–Wolff–Goodell [pain] scale was discredited in 1957 by the distinguished clinical anaesthesiologist Henry K. Beecher". Paterson implicates Hardy et al.'s measurement of just-noticeable pain differences (pain jnds) in Beecher's critique.

But Dr. Beecher (MD) criticized Hardy et al. rather for their method of assessing pain mitigation by drugs, namely, the elevation of the pain-detection threshold, measured psychophysically using Hardy et al.'s *dolorimeter* – also used by Hardy et al., a few years later, for pain jnds. Beecher disparaged psychophysics in favor of clinical trials. But dolorimetry and clinical trials had already been recognized as different means to similar ends; Beecher merely made the dolorimeter a "straw man" and then burnt it down. Tousignant (2011, p. 166) offers a final perspective: "The dolorimeter fit into a gap between animal studies (convenient, relatively cheap and easy to standardize, but with limited translatability to humans) and clinical ones (perhaps more authentic, but much more difficult, and expensive, to standardize)".

References

Abbott, D. (2013). The reasonable ineffectiveness of mathematics. *Proceedings of the IEEE,* 101, 2147-2153.

Beecher, H.K. (1952). Experimental pharmacology and measurement of the subjective response. *Science*, 116, 157-162.

Beecher, H.K. (1953a). *Response*: Pain – controlled and uncontrolled. *Science*, 117, 166-167.

Beecher, H.K. (1953b). A method for quantifying the intensity of pain. *Science*, 118, 322-324.

Dunnette, M.D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist,* 21, 343–352.

Hardy, J.D., Wolff, H.G., & Goodell, H. (1940). Studies on pain: a new method for measuring pain threshold: observations on spatial summation of pain. *Journal of Clinical Investigation*, 19, 649-657.

Hardy, J.D., Wolff, H.G., & Goodell, H. (1947). Studies on pain: discrimination of differences in intensity of a pain stimulus as a basis of a scale of pain intensity. *Journal of Clinical Investigation*, 26, 1152-1158.

Hardy, J.D., Wolff, H.G., & Goodell, H. (1953). Pain – controlled and uncontrolled. *Science*, 117, 164-165.

Herget, C.M., Granath, L.P., & Hardy, J.D. (1941). Thermal sensation and discrimination in relation to intensity of stimulus. *American Journal of Physiology*, 134, 645-655.

Kuorikoski, J. (2021). There are no mathematical explanations. *Philosophy of Science*, 88, 189-212.

Nizami, L. (2019). Information Theory is abused in neuroscience. *Cybernetics & Human Knowing*, 26, 47-97.

Paterson, M. (2019). On pain as a distinct sensation: mapping intensities, affects, and difference in 'interior states'. *Body & Society*, 25, 100-135.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.

Smith, G. (2018). *The AI Delusion*. New York, NY: Oxford University Press.

Tousignant, N. (2011). The rise and fall of the dolorimeter: pain, analgesics, and the management of subjectivity in mid-twentieth-century United States. *Journal of the History of Medicine and Allied Sciences*, 66, 145-179.

Tousignant, N. (2014). A quantity of suffering: measuring pain as emotion in the mid-twentieth-century USA. In: R. Boddice (Ed.), *Pain and Emotion in Modern History*. New York, NY: Palgrave Macmillan, pp. 111-129.

Wolff, H.G., Hardy, J.D., & Goodell, H. (1940). Studies on pain: measurement of the effect of morphine, codeine, and other opiates on the pain threshold and an analysis of their relation to the pain experience. *Journal of Clinical Investigation*, 19, 659-680.

Yong, E. (19 Nov 2018). Psychology's replication crisis is running out of excuses. *The Atlantic*.