American Pain Society
RESEARCH
EDUCATION
TREATMENT
ADVOCACY
PUBLISHED BY ELSEVIER

The Journal of Pain, Vol 20, No 4 (April), 2019: pp 472.e1–472.e12
Available online at www.jpain.org and www.sciencedirect.com

Online Exclusive

Was That Painful or Nonpainful? The Sensation and Pain Rating Scale Performs Well in the Experimental Context

Victoria J. Madden,[*,†] Peter R. Kamerman[‡,§] Valeria Bellan,[†] Mark J. Catley,[†] Leslie N. Russek,[†,¶] Danny Camfferman,[†] and G. Lorimer Moseley[†]

[*]*Departments of Anaesthesia and Perioperative Medicine and Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa,* [†]*School of Health Sciences, University of South Australia, Adelaide, Australia,* [‡]*School of Physiology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa,* [§]*School of Biomedical Sciences, Faculty of Health Sciences, Curtin University, Perth, Australia,* [¶]*Department of Physical Therapy, Clarkson University, Potsdam, New York*

Abstract: **In experiments on pain, participants are frequently exposed to nonpainful and painful stimuli; however, the conventional pain-rating scales lack a nonpainful range and a clear point of transition from nonpainful to painful events. The Sensation and Pain Rating Scale (SPARS) assesses the full stimulus intensity range, extending from no sensation (rating: −50) to worst pain imaginable (rating: +50), and it explicitly identifies pain threshold (rating: 0). Here, we tested the SPARS in 2 experiments by using laser heat stimuli to establish its stimulus–response characteristics (Experiment 1, N = 19, 13 stimulus intensities applied 26 times each across a 1–4 J range), and compared it to 0 to 100 scales that assess nonpainful (0: no sensation, 100: pain) and painful (0: no pain, 100: worst pain imaginable) events (Experiment 2, N = 7, 9 stimulus intensities applied 36 times each across a 1.5–4.5 J range). Despite high inter- and intraindividual variations, we found a reasonably consistent curvilinear stimulus–response relationship (the curve flattens around pain threshold), with stable response characteristics across the range of the scale. The SPARS ratings transformed to a 0 to 100 range tended to be lower than the 0 to 100 pain rating scale in the noxious stimulus intensity range and greater than the 0 to 100 nonpainful sensation scale in the non-noxious stimulus range, likely reflecting differences in scale dimensionality. The SPARS overcomes limitations in scale range inherent to conventional pain rating scales. As such, it is well suited to experimental studies that must quantify a wider range of perceptual intensity or distinguish between painful and nonpainful events.**

Perspective: ***This article presents the stimulus–response characteristics of a new scale designed to allow participants to rate a range of nonpainful and painful stimuli. The scale could be useful for research that involves exposing participants to a range of stimulation intensities or requires a clear distinction between nonpainful and painful events.***

Key words: *Self-report, pain assessment, pain threshold, healthy volunteers, multilevel analysis.*

Received August 2, 2018; Revised October 5, 2018; Accepted October 15, 2018.

V.J.M. was supported by the Oppenheimer Memorial Trust and is now supported by the National Research Foundation of South Africa. V.J.M. receives speaker's fees for lectures on pain and rehabilitation. G.L.M. has received support from Pfizer, Kaiser Permanente, Workers' Compensation Boards in Australia, Europe and North America, the International Olympic Committee, the Port Adelaide Football Club, and the Arsenal Football Club. He receives royalties for books on pain and speaker's fees for talks and professional development courses on pain and rehabilitation.

The authors have no conflicts of interest to declare.

Supplementary data accompanying this article are available online at www.jpain.org and www.sciencedirect.com.

Address reprint requests to Victoria J. Madden, PhD, D23.30 Department of Anaesthesia and Perioperative Medicine, Groote Schuur Hospital, Anzio Road Observatory, Cape Town 7925, South Africa. E-mail address: torymadden@gmail.com

1526-5900/$36.00

https://doi.org/10.1016/j.jpain.2018.10.006

Pain intensity is best assessed by self-report. Clinicians and researchers attempt to capture another's pain by using cross-modality matching tasks in which pain is matched to a rating scale.[31] Numerical rating scales (NRSs) and visual analog scales (VASs) anchored at opposite extremes with "no pain" and something akin to "most intense pain you can imagine" are widely used, and have evidence for being robust assessments of pain in clinical and experimental contexts.[8] Indeed, for both of these scale types, the stimulus–response relationship has been described by a classical psychophysical power relationship, and the VAS has been said to have ratio properties.[27,29]

Conventional pain-rating NRS and VAS have 3 important limitations with regard to the experimental context. First, participants are frequently exposed to stimuli that span the range from nonpainful to painful; however, these scales have neither clear nor standardized capacity to accommodate that range. Second, these scales do not allow for precise identification of the transition between nonpainful and painful stimuli, even though identifying this "pain threshold" is necessary for many experimental paradigms. Third, the structure of the scales predisposes to error for stimuli around the pain threshold. For example, experimental participants used up to the first 15 points on the 0 to 100 NRS to rate stimuli that they also reported as "not painful" on a verbal report scale.[19] Additionally, our own pilot trials show that patients often use ratings of up to 3 on a 0 to 10 scale to indicate percepts that, on further questioning, are not so much painful as uncomfortable or mildly unpleasant [Moseley, Unpublished data]. These limitations lead to the lower end of scales showing measurement insensitivity that could result in erroneous findings, particularly in studies that investigate pain threshold.

To overcome these limitations, we developed a new scale, the Sensation and Pain Rating Scale (SPARS) (previously called FEST[22]), that locates pain threshold explicitly and allows participants to report the intensity of a range of nonpainful and painful experiences. Pilot studies suggested that participants could readily distinguish and report on the different percepts of "painful" and "nonpainful" if they lay on opposite sides of an intuitively neutral point. Therefore, we retained the 0 to 100 range of a conventional NRS or VAS but structured the SPARS to lie between -50 (anchor: "no sensation") and +50 (anchor: "most intense pain you can imagine"), with pain threshold—the intuitively neutral point—represented by 0 (anchor: "the exact point at which what you feel transitions to pain"). In addition, the range from −50 to 0 was labeled "nonpainful," whereas the range from 0 to +50 was labeled "painful" (Fig 1).

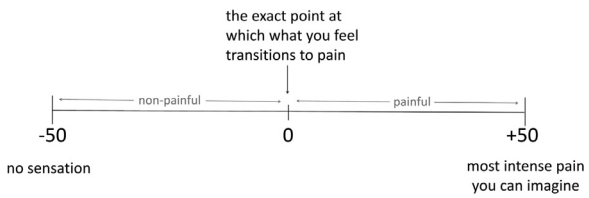

Figure 1. SPARS.

The scale was balanced to reflect the assumption that incremental increases in nonpainful and painful percept are uniform across the scale. Here, we report on 2 experiments undertaken to assess the scale's characteristics. In Experiment 1, we determined the stimulus–response profile of the SPARS. In Experiment 2, we compared the SPARS to 2 NRSs that separately covered the nonpainful and painful ranges spanned by the SPARS.

Methods

Participants

Convenience samples of healthy volunteers between 18 and 65 years of age were sought via flyers, social media, and word of mouth. Volunteers were eligible if they spoke English comfortably; had no history of sensory deficits or chronic pain (defined as daily pain for 3 months or longer[4]); did not have pain at the time of testing; were not using medication that might affect skin sensitivity, pain, or healing; did not have an elevated risk of laser-induced skin damage (eg, thinned skin); and had not been diagnosed with diabetes mellitus or a neurological, peripheral vascular, or psychiatric disease. Participants received study information electronically and in print and gave written informed consent. They were compensated AU$20 per hour. The study conformed to the Declaration of Helsinki and was approved by the University of South Australia's Human Research Ethics Committee (protocols 31880 and 36008).

Stimulus

A cutaneous heat stimulus was applied using a neodymium-doped yttrium aluminum perovskite laser (Nd:YAP; Stimul 1340, Deka, Italy), which is thought to activate only nociceptors.[26] The location of successive stimuli (trials) was slightly shifted to prevent skin damage and progressive alterations in sensitivity,[12,36] and all procedures were consistent with our previous recommendations on safe use of experimental laser stimuli.[23] Stimulus intensity, diameter, duration, and interstimulus interval of the laser differed between the 2 experiments and are described separately under the specific methods for each experiment.

Experiment 1

Aim

To determine the stimulus–response characteristics of the SPARS by using laser heat stimuli applied to skin on the lower back of healthy adults.

Methods

Procedure

Participants were briefed on the use of the SPARS. Then, they lay prone on a plinth. A 3500-mm^2 stimulation area was marked on their backs (lower back, over the

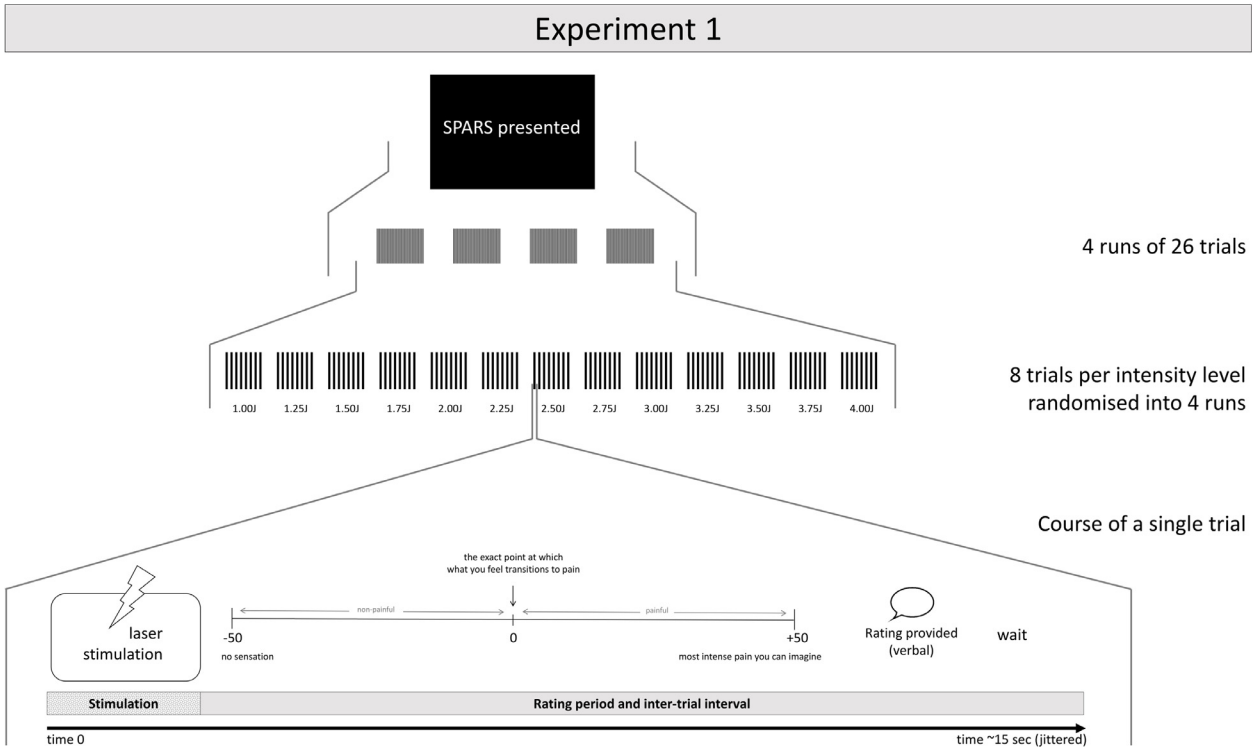

Figure 2. Procedure for Experiment 1.

right paraspinal area), and excess hair was trimmed (not shaved) using clippers. A copy of the SPARS was positioned within the participants' view. Laser stimuli (spot diameter of 4 mm, pulse duration 6 ms, interstimulus interval ~15 s) were delivered to the skin. Participants provided verbal SPARS ratings for 8 trials at each of 13 energy levels (1–4 J, at .25 J intervals) in a randomized order, during 4 blocks of 26 trials (104 trials in total) (Fig 2). Ratings were not restricted to whole numbers. For safety, trials of energy >3.5 J were omitted from the sequence of trials when participants showed skin reddening or reported pain greater than +30, at the discretion of the investigator, as per ethical approval requirement. Participants listened to repetitive ocean wave noises (White Noise Sleep Pillow app for iPhone; Fitness22 Ltd, Petach Tikvah, Israel) through headphones to mask the sound of the laser machine's foot pedal.

Data Handling and Statistical Analysis

Data tidying and statistical analyses were done in R v3.5.0 (The R Foundation for Statistical Computing, Vienna, Austria), in RStudio v1.1.453 (RStudio Inc, Boston, MA). The following packages were used: magrittr,[1] tidyverse,[35] patchwork,[25] ggridges,[37] robustlmm,[16] lme4,[2] lqmm,[10,11] HLMdiag,[20] sjPlot,[21] car,[9] lmerTest,[17] and influence.ME.[24] All data cleaning and analysis scripts, along with their outputs, are available online.[15] Participants did not consent to release their data publicly, and therefore, data files are available only on request. The main analysis consisted of 7 steps:

1. The most suitable measure of central tendency for repeated trials (N = 13) at each stimulus intensity for each participant was chosen on the basis of a visual comparison of the arithmetic mean, median, geometric mean, and Tukey trimean overlaid on scatter and density plots of SPARS ratings.
2. Visual inspection of scatterplots with superimposed, locally weighted scatterplot smoothing (loess, span [alpha parameter] = .75, fit family = Gaussian) curves was used to assess for order effects at a trial level (ie, does the intensity of the preceding stimulus affect the rating of a stimulus?) and at a block level (ie, does the rating of a given stimulus intensity change across experimental blocks?) within each participant, for each stimulus intensity.
3. The stimulus–response relationship was visualized at the individual level and at the group level, using loess curves. These visualizations indicated an approximately linear shape, with a slight curve. Linear mixed model regression was used to describe the relationship between the SPARS rating and stimulus intensity. To allow for a curvilinear relationship between stimulus intensity and rating, we expressed the predictor (stimulus intensity) as 1st (linear)-, 2nd (quadratic)-, and 3rd (cubic)-order orthogonal polynomials. For each polynomial expression, we modeled the random effects as random intercept only (grouping factor: participants) and as random intercept and slope (stimulus intensity). The random intercept only and random intercept and slope models were compared using the likelihood ratio test, and the better model was taken forward in each case.
4. Model diagnostics were run to confirm that the "best fit" model was specified properly. We examined level 1 residuals (conditional and fixed effects) and level 2 residuals (random effects) and influence points.[20] Quantile regression (quantiles: .025, .25, .5, .75, and .975) was used to confirm that the response characteristics of the "best model" were consistent across the quantiles of the SPARS.
5. The "stability" of the model (ie, whether linear, quadratic, or cubic) was examined using a bootstrap procedure. Briefly, we examined the stimulus–response relationship

of the SPARS under 5 study sample size scenarios (n = 6, 9, 12, 15, and 18). Under each sample size scenario, we generated 500 random samples (with replacement) and fit linear mixed models to each of the 500 samples with the predictor (stimulus intensity) modeled as 1st (linear)-, 2nd (quadratic)-, and 3rd (cubic)-order orthogonal polynomials. The likelihood ratio test was then used to decide which of the 3 models had the best fit for each sample, and the proportion of the best models fitting each of the polynomial forms was plotted for each study sample size scenario.

6. To examine the scale's properties, we assessed whether the relationship between the change in rating intensity and change in stimulus intensity between successive stimuli was affected by the direction of the change in stimulus intensity (ie, whether the preceding stimulus was greater than or less than the current stimulus). To make the evaluation, we compared robust linear mixed models (with random intercept) with and without interaction between change in stimulus intensity and the direction of the change in stimulus intensity by using 95% Wald confidence intervals (CIs) of the fixed effect estimates.
7. Finally, we assessed the individual (intraindividual) variability and group (interindividual) variability in scoring on the SPARS at each intensity. To calculate intraindividual variance, we calculated the median absolute deviation (MAD, a robust measure of dispersion) by using the ratings from repeated stimuli at each stimulus intensity. We also used the MAD to calculate interindividual variance for each scale and at each intensity but used the median ratings by individuals at each intensity as the input (see Supplementary File 9 for the calculation).

Experiment 2

Aim

Having established the stimulus–response characteristics of the SPARS, we compared the SPARS scale to 2 NRSs that separately covered the nonpainful and painful ranges spanned by the SPARS.

Methods

Participants underwent a single 4-hour testing session, during which they rated stimulation trials on 3 different scales.

Stimuli

Laser stimuli (spot diameter of 6 mm, pulse duration 5 ms, interstimulus interval ~30 s) were delivered to the dorsum of the left forearm within a delineated area of 3600 mm^2.

Outcomes

Participants were familiarized with the 3 rating scales before the experimental procedure. During the experiment, the scales were presented on a computer monitor, and participants made their ratings on the screen, using the computer mouse (Affect 4.0 software[32]).

1. SPARS: In the initial presentation of the SPARS, the range from extreme left (−50) to the center (0) was labeled "nonpainful" and the range from the center (0) to extreme right (+50) was labeled "painful" (Supplementary File 1). However, these 2 labels were omitted from the scale on which participants provided their ratings during the trials, so as to prevent excessive amounts of text on the screen.
2. NRS: A 0 to 100 numerical pain rating scale, which was anchored at extreme left with 0, "no pain," and at the extreme right with 100, "most intense pain you can imagine."
3. Numerical sensation rating scale (SRS): A 0 to 100 numerical SRS, which was anchored at extreme left with 0, "no sensation," and at the extreme right with 100, "pain."

Thus, the lower extreme of the NRS ("no pain") and the upper end of the SRS ("pain") overlap at the transition between nonpainful and painful stimulus perceptions.

Procedure

First, safe stimulus intensity parameters were determined for each participant. Two trains of single laser stimuli were delivered within the stimulation zone, in an ascending fashion, beginning at 1 J and increasing in 1 J steps to 4 J, ending in response to focal skin reddening or the participant reporting very intense pain. If neither cue occurred, a maximum intensity of 4.5 J was tested. The results of this procedure were used to estimate the appropriate and safe range of stimulus intensities to be used for each participant.

After determining the safe range, participants underwent testing in 3 blocks, with a break of 1 to 15 minutes between blocks. Each block was allocated to 1 of the 3 scales, such that each scale was presented during just 1 of the 3 blocks. The order of the blocks was counterbalanced across participants. Each block was divided into 4 stimulus runs (Fig 3), broken by short breaks, with 3 trials per intensity (3 × 9 intensity = 27 trials) delivered during each run. Thus, each participant received a total of 36 stimulation trials at each of 9 intensities spanning 1.5 to 3.5 J, 1.75 to 3.75 J, 2 to 4 J, 2.25 to 4.25 J, and 2.5 to 4.5 J (.25 J step between successive intensities, and range determined for each individual during the safety procedure), with 12 trials at each intensity being rated on each scale. Each block was preceded by on-screen information on the scale to be used for that block (Supplementary File 1).

Data Handling and Statistical Analysis

Data tidying and statistical analyses were done using the same statistical software described for Experiment 1, with the addition of the boot[5,6] package. Scripts and their outputs are available online.[15] The data from Experiment 2 were handled differently to those from Experiment 1 with respect to the stimulus intensities. In Experiment 1, all participants received the same range of stimulus intensities (1–4 J). In Experiment 2, all participants received stimuli of 9 laser intensities, with a .25 J step between intensities, but those 9 intensities could lie within different intensity ranges because of the calibration procedure that based the range of intensities on estimated, individualized "safe parameters." To

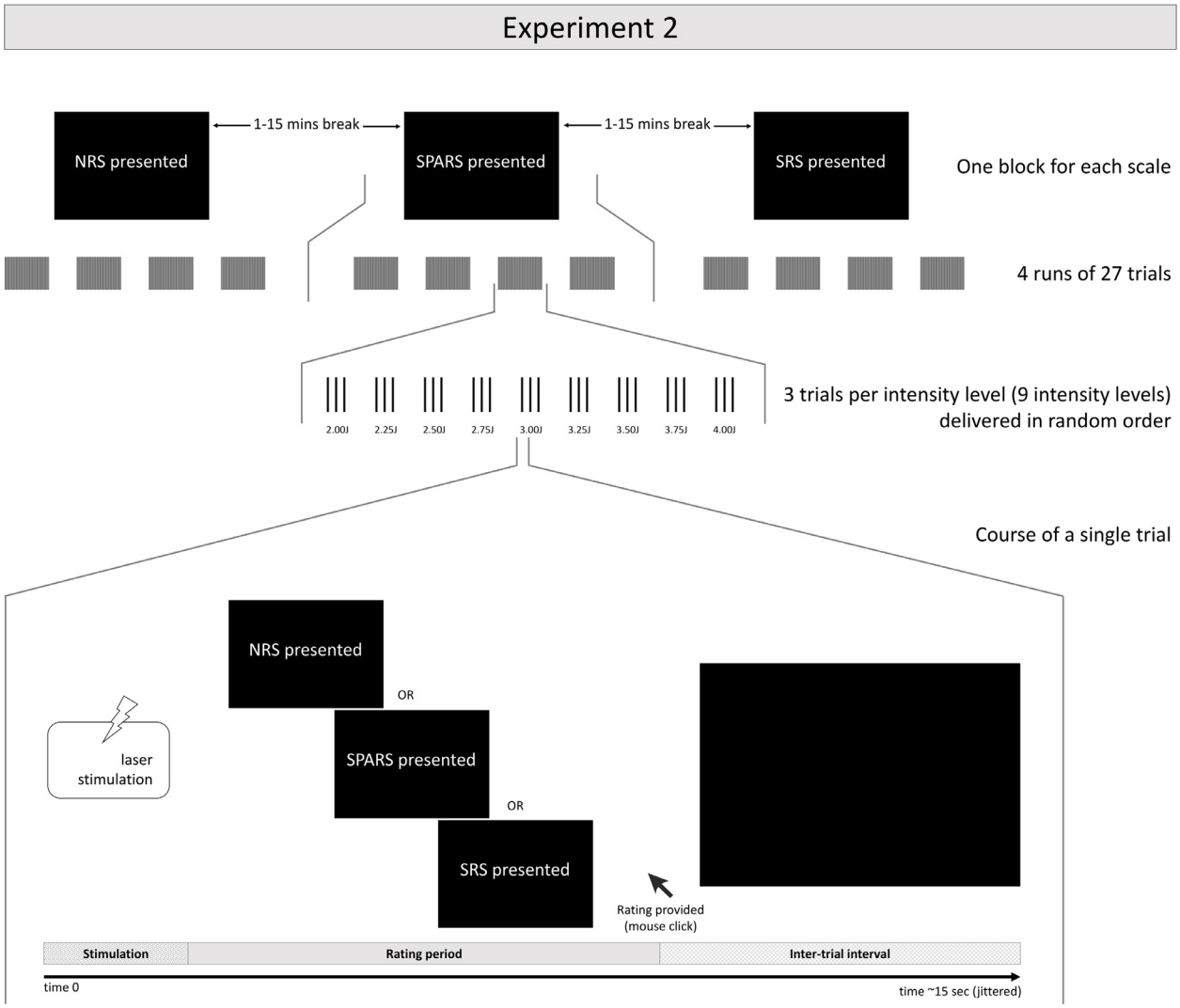

Figure 3. Procedure for Experiment 2.

account for this, we transformed the exposure intensities into a relative scale by ranking (1–9) an ordered list of 9 intensities for each participant. This brought everyone onto the same 1 to 9 scale of relative stimulus intensity. The first 3 steps of the analysis followed a similar structure to the analysis for Experiment 1:

1. In keeping with the results of Experiment 1, the Tukey trimean was used as the measure of central tendency within each participant. Using the Tukey trimean data, the stimulus–response relationship for the SPARS was visualized at the levels of the individual (first, with responses pooled across all 4 runs and then by each of the 4 runs) and at the level of the group, using loess curves. We then proceeded with a linear mixed model analysis, as for Experiment 1 (refer to Step 3 in Experiment 1).
2. Model diagnostics were run on the "best fit" model, and quantile regression was used to confirm that the response characteristics of the "best model" were consistent across the quantiles of the SPARS.
3. The variability in scoring on each scale was calculated for each stimulus intensity rank at both the individual (intraindividual) and the group (interindividual) levels, again using the MAD, with the ratings from repeated stimuli at each stimulus intensity rank (see Supplementary File 13 for calculations).

The next steps differed from those of Experiment 1:

1. Agreement between the SPARS and each of the other 2 scales was assessed by using 50,000 bootstrap samples (sampled with replacement) of each scale's ratings to calculate 95% bootstrap CIs (bias-corrected and accelerated method) for the difference in median ratings for each scale at each stimulus intensity and for each of the resamples. We then subtracted the values as follows: median (SPARS) − median (NRS/SNS). This procedure was chosen because the limits of agreement method[3] is inappropriate when the scales being assessed are not concurrently used to rate each stimulus. The CIs that included 0 were interpreted as indicating agreement between the 2 measurement scales at a particular stimulus intensity. An upper limit of the CI being <0 was interpreted as indicating that the SPARS rating was consistently less than the rating on the other scale and vice versa when the lower bound of the CI was >0.

Table 1. Characteristics of the Best-Fit Cubic Model for the Data From Experiment 1

	Response		
Coefficients	*Estimate*	*95% CI*	*P Value*
Fixed effects (intercept)	-4.67	−10.89–1.56	.159
Stimulus intensity (linear)	205.35	163.69–247.01	<.001
Stimulus intensity (quadratic)	2.12	−10.42–14.67	.744
Stimulus intensity (cubic)	20.95	8.40–33.49	.004
Random effects			
Within-participant variance	40.768		
Between-participant variance	639.311		
Random slope-intercept correlation	−.889		
Number of participants	19		
Observations	244		

Results

Experiment 1

In total, 19 adults (including 11 females, 18–31 years old) met the selection criteria, and each adult participated in a single 45-minute testing session.

Summary results are provided in Supplementary File 2. Response data were missing those trials that were omitted owing to marked skin reddening for 2 participants —for 1 participant, at 4 J only, and for another participant, at 3.5 J and 3.75 J. The results of the 7-step analysis were as follows:

1. There was a significant heterogeneity in the grouping of ratings within and between individuals, with a tendency for data to be left skewed. Of the measures of centrality under consideration, the Tukey trimean and median showed the best stability across all stimulus intensities. As expected, the geometric mean performed poorly on the left-skewed data. The arithmetic mean was strongly affected by the skewed data. When there was a strong skew, the Tukey mean was pulled slightly away from the median and toward the affected tail; otherwise, the 2 measures yielded similar values. Given that the Tukey trimean incorporates information in the distribution of the data but is not excessively affected by extreme values, the Tukey trimean was chosen as the most suitable measure for central tendency for repeated measurements within an individual, at a given stimulus intensity. The Tukey trimean at each intensity for each participant was used in the modeling procedures described under item number 4 described below (Supplementary File 3).
2. There were no systematic order effects between trials or over the 4 blocks (Supplementary File 4).
3. Visualization of the relationship between SPARS rating and stimulus intensity at a group level showed an approximately linear shape, with a slight curve (Supplementary File 5).

Modeling showed that a cubic model with random intercept (grouping factor: participants) and random slope (stimulus intensity) had the best fit to the data (Table 1). The resulting curvilinear response function was steepest at the extremes and flattened out in the mid ranges of stimulus intensity (Fig 4). The details of the other models and the comparisons between the models are given in Supplementary File 6.

1. Diagnostic tests run on the cubic model confirmed that the assumptions were met: Level 1 residuals met the conditions of linearity and homoscedasticity, level 1 and level 2 residuals were normally distributed, and there were no points of sufficient influence to cause concern (Supplementary File 7 provides full details). The cubic model was confirmed to be well specified. Details of the model are provided in Table 1.

Quantile regression showed that there was good stability in the shape of the response characteristics across the quantiles (Fig 5). For all stimulus intensities, the distribution was left skewed (long tail toward lower ratings).

1. Visualization of 5 separate samples (n = 6, 9, 12, 15, and 18) drawn from a bootstrap procedure performed on the current data showed a consistent shape to the stimulus –response curve. Repeated calculation of the best model

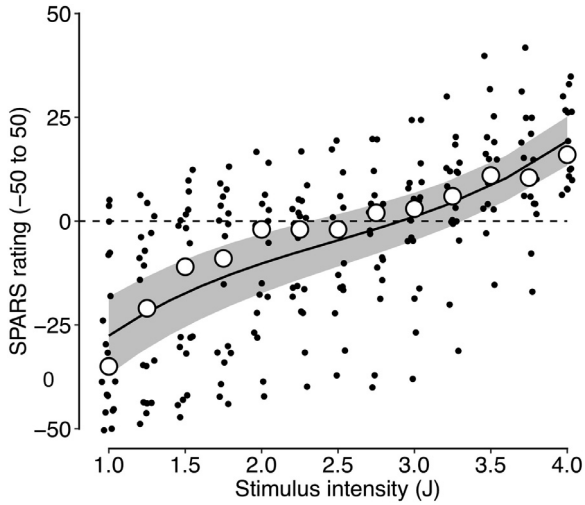

Figure 4. Experiment 1: Stimulus–response plot for the SPARS, with the cubic model visualized. Small dots: participant-level Tukey trimeans; large open circles: group-level median; black line: values predicted by the cubic model; gray band: 95% CI for the cubic model.

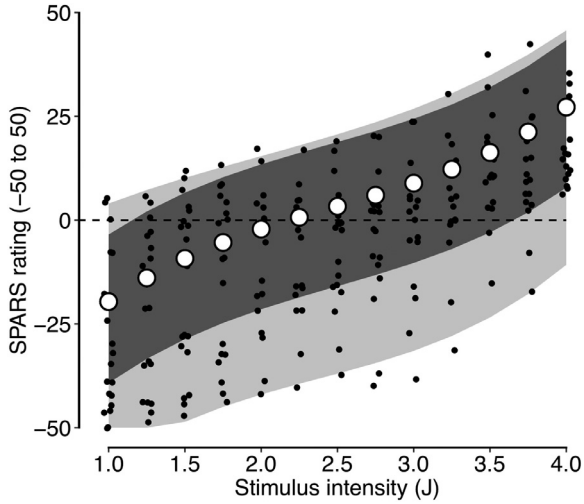

Figure 5. Experiment 1: Quantile regression (with original Tukey trimean data). Open circles: 50th percentile (median); dark band: interquartile range; light band: 95% prediction interval.

to fit 500 resamples (with replacement) of the 5 differently sized samples showed that the cubic model was most commonly the best-fit model, increasing progressively from about 40% of models to about 80% of models as sample size increased from 6 to 18 participants (Fig 6 and Supplementary File 8).

2. The relationship between the change in rating intensity and change in stimulus intensity between successive stimuli was linear and was not affected by the direction of the change in stimulus intensity; it was consistent across increasing and decreasing changes in stimulus intensities (Fig 7). The estimate for the change in stimulus intensity main effect was 13.3 (95% CI, 12.3–14.4), indicating that a 1 J change in stimulus intensity resulted in 12.3 to 14.4 units of change in SPARS rating (Supplementary File 9).
3. The variance analysis and the MAD values are shown in Supplementary File 10. Variance was relatively stable over the full range of the scale. Most intraindividual MAD values were <10% of the scale range, and there was no clear pattern in their distribution. On the whole, interindividual MAD values tended to be higher at lower stimulus intensities, and this variance decreased as stimulus intensity increased.

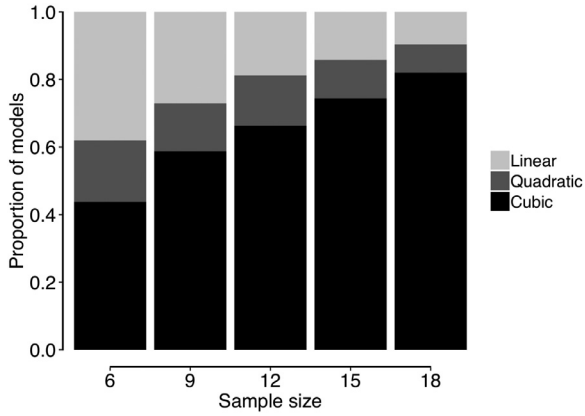

Figure 6. Experiment 1: Proportion of best models for each of 4 study sample size scenarios. Five hundred resamples (with replacement) were taken for each study sample size scenario.

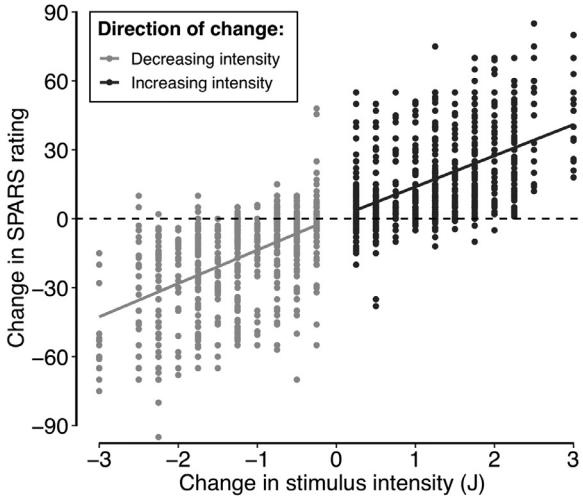

Figure 7. Experiment 1: Group-level plot showing the relationship between the change in SPARS rating and the change in stimulus intensity between 2 successive stimuli. Data have been omitted when successive stimuli were of the same intensity (0 change). Small dots: individual responses for all trials; lines: linear trend for context; light points or line: when a stimulus was of greater intensity than the preceding stimulus (value of the change is negative); dark points or line: when a stimulus was of lesser intensity than the preceding stimulus (value of the change is positive).

Experiment 2

In total, 7 adults (5 females, 21–30 years old) participated. One participant received stimuli within the range of 1.75 to 3.75 J, 3 participants received 2.25 to 4.25 J, and 3 participants received 2.5 to 4.5 J. A software error resulted in the loss of all SRS rating data for 1 participant. The remainder of that participant's data were included in the analysis. The results of the 4-step analysis were as follows:

1. Plots of individual stimulus–response curves for each testing block showed reasonable within-subject consistency in the curve across the 4 testing blocks (Supplementary File 11). Plots of individual participants' stimulus–response curves showed between-subject heterogeneity in curve shape. Visualization of the relationship between SPARS rating and stimulus intensity at a group level showed an approximately linear shape, with a very slight curve (Fig 8 and Supplementary File 11). Modeling of the polynomial regression and comparison between models showed that the quadratic and cubic models with random intercept (grouping factor: participants) and random slope (stimulus intensity) fitted the data well (Fig 9 and Table 2) and provided a superior fit to that provided by the linear model. Because the quadratic and cubic models produced similar fits, we used the simpler quadratic model for the remainder of our analyses. The details of the other models and the model comparisons are given in Supplementary File 11.
2. Diagnostic tests run on the quadratic model confirmed that the assumptions were met: level 1 residuals met the conditions of linearity and homoscedasticity, level 1 and level 2 residuals were normally distributed, and there were no points of sufficient influence to cause concern (Supplementary File 11 provides full details). The quadratic model was confirmed to be well specified. Quantile regression showed

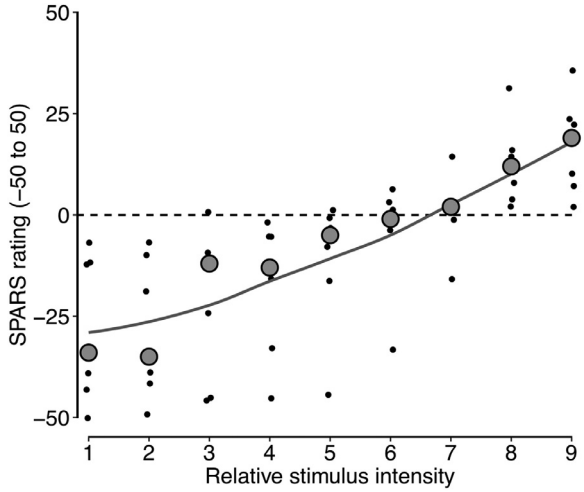

Figure 8. Experiment 2: Group-level stimulus–response plot (relative intensity). Relative intensity was calculated using the rank of the ordered (ascending) stimulus intensities to which each participant was exposed. Small dots: participant-level Tukey trimeans; large shaded circles: group-level median; line: loess curve.

that there was a good stability in the shape of the response characteristics across the quantiles (Fig 10). The prediction interval was extremely broad, which is unsurprising, with data from only 7 participants.

3. Comparative plots of the stimulus–response relationship for each scale are shown in Fig 11 and Supplementary File 12. The variances in responses for each scale are reported in full in Supplementary File 13. Again, the MAD values for the SPARS were mostly <10% of the full range of the scale. There was no clear pattern in intraindividual variance in intensity rating between the 3 scales or within each scale, and the variance was frequently high. Similarly, there was no clear pattern in interindividual variance in intensity rating between the 3 scales; however, again, the variance could be high. The SPARS variance followed a "U-shaped" response (minimum value at stimulus rank = 7), and the NRS had low variance at low stimulus intensities. The SRS did not show a clear pattern. Both intra- and interindividual variances were low at lower stimulus intensities for the NRS, likely because the NRS assesses painful stimuli.
4. Intra-individual agreement (Supplementary File 13) between the SPARS and the NRS differed notably. The SPARS ratings tended to be less than the NRS ratings across all stimulus intensities. This tendency was most apparent at lower intensities, which is likely an artefact of the 2

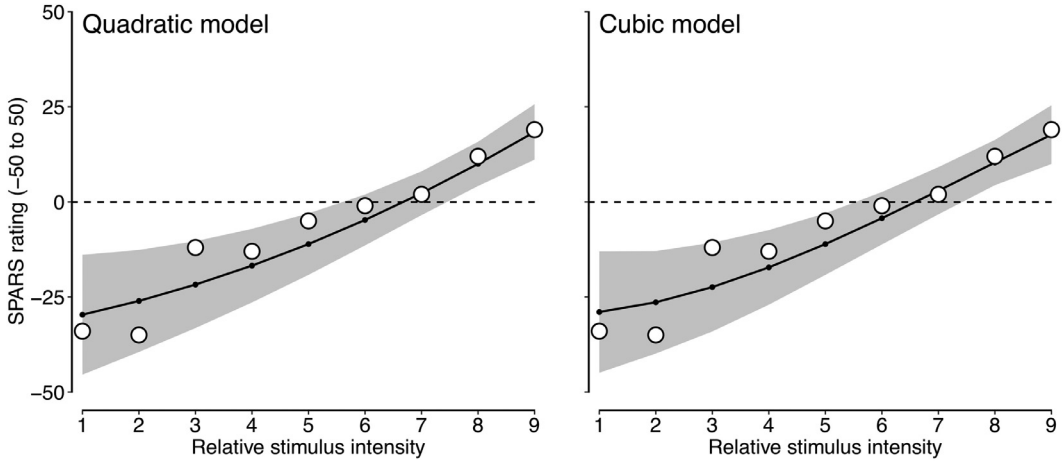

Figure 9. Experiment 2: Quadratic (left) and cubic (right) models: Predicted SPARS ratings versus relative stimulus intensity. Relative intensity was calculated using the rank of the ordered (ascending) stimulus intensities to which each participant was exposed. Black dots or line: predicted values; large open circles: group-level median; gray band: 95% CI.

Table 2. Parameters of Quadratic (Best-Fit) Model for Experiment 2

	Response		
Coefficients	*Estimate*	*95% CI*	*P Value*
Fixed effects (intercept)	−8.83	−16.47–1.18	.064
Stimulus intensity (linear)	123.16	78.03–168.29	.002
Stimulus intensity (quadratic)	15.87	1.42–30.31	.075
Stimulus intensity (cubic)			
Random effects	54.304		
Within-participant variance	512.022		
Between-participant variance	−.957		
Random slope-intercept correlation	7		
Observations	63		

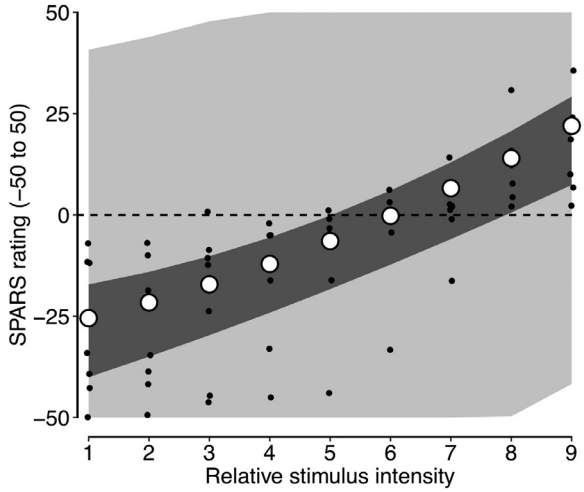

Figure 10. Experiment 2: Quantile regression (with original Tukey trimean data). Large open circles: 50th percentile (median); dark band: interquartile range; light band: 95% prediction interval.

instruments: the SPARS continues to record nonpainful stimuli (using a negative range), whereas the NRS reaches a floor value of 0 when the stimulus intensity is no longer painful. Any agreement tended to be in the upper ranges of stimulus intensity. There was also noteworthy heterogeneity in interscale agreement across participants. In keeping with the intraindividual level, interindividual (group) agreement between the SPARS and the NRS was poor at the low end of the intensity range. At higher stimulus intensities, the upper limits of the 95% CIs tended to approach 0, but except for the 8th most intense stimulus, the intervals did not include 0. (We also believe that the 8th stimulus finding results from sampling variation rather than being a true effect.) Across the range of stimulus intensities, the width of the interval was large and reflects high interindividual variation in the SPARS and NRS ratings at each stimulus intensity.

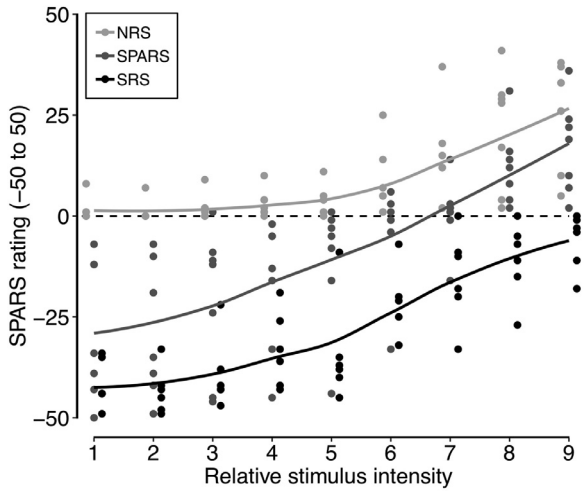

Figure 11. Experiment 2: Scaled group-level stimulus–response ratings on the pain NRS, SPARS, and SRS. Points (dodged for clarity): Tukey trimeans; curves: loess lines. NRS: 0 (no pain) to 50 (worst pain you can imagine); SRS: -50 (no sensation) to 0 (pain); SPARS: -50 (no sensation), 0 (pain threshold), +50 (worst pain you can imagine).

Intraindividual agreement between the SPARS and the SRS also differed notably; but here, the problem was most apparent at higher stimulus intensities. The SPARS ratings were consistently higher than the SRS ratings at the 2 highest ranks of stimulus intensity, likely because the SRS reaches a ceiling value when the stimulus intensity is first perceived as painful, whereas the SPARS continues to allow ratings into the painful range. Here, any agreement tended to be in the lower ranges of stimulus intensity. Again, there was a significant heterogeneity across participants. Interindividual (group) agreement between the SPARS and the SRS was better than that between the SPARS and the NRS (more intervals included 0); but the intervals were wide and the inclusion of 0 was marginal in most cases. In general, the SPARS ratings tended to be greater than the SRS ratings for a given stimulus intensity.

Discussion

We aimed to establish the utility and features of the SPARS, a scale that assesses somatosensation over the full range (no sensation to worst pain). Experiment 1 showed that the stimulus–response relationship was reasonably consistent and curvilinear, its slope was steeper at the extremes than in the middle range of stimulus intensity, and the response characteristics were stable across the scale's range. The SPARS showed no hysteresis; change in stimulus intensity had a consistent linear relationship to change in rating intensity between successive stimuli, regardless of the direction of the change in stimulus intensity (Fig 7). Ratings varied by <10% of the scale range at most stimulation intensities. Experiment 2 replicated the SPARS's curvilinear stimulus–response relationship and stability of responses across the stimulus range. Our assessment of agreement between the SPARS and the other 2 scales showed significant heterogeneity in ratings within each scale and between the scales at each stimulus intensity. In general, ratings provided on the SPARS were lower than those on the NRS and greater than those on the SRS.

Previous studies on rating scales for pain have found a stimulus–response relationship characterized by a power curve,[27–29] in keeping with the stimulus–response curves typical of other psychophysical measurement systems that use cross-modality matching.[33] The first validation studies on the widely used 0 to 100 scales for pain plotted participants' ratings on a "relative" scale. Although the meaning of "relative" was not specified, the authors seem to have followed the lead of the early psychophysics texts, for example, Stevens,[33] by redistributing or "stretching" ratings across the range of ratings used by an individual participant—an intraindividual relativity. For example, if a certain individual gives his or her highest rating of 60 of 100 for a certain stimulus intensity, his or her relative rating for that stimulus intensity would be 100 (ie, 100% of his or her maximum rating). Relative ratings were used to calculate group average ratings for each stimulus intensity. Group averages were plotted, yielding a power curve. Although that approach

facilitates comparison across psychophysical modalities, it needlessly loses information in 2 steps. The first step is to convert raw ratings into relative ratings. A maximum rating that lies halfway across a scale for pain has a very different meaning from a maximum rating that is close to the scale's maximum; however, the relative rating for both is 100—the distinction between them is lost. The second step is to pool data across participants, which ignores individual variability and yields a curve representing only group averages. Pain in response to a given stimulus varies widely across individuals and within an individual across time, and group averaging loses the heterogeneity in stimulus–response data. Indeed, the current findings reflect high heterogeneity between participants, in stimulus–response curve shape, variance of ratings, and agreement between scales at each intensity. Therefore, there is a strong case for analytical approaches that account for individual variability, particularly when there is no immediate need for comparison between psychophysical modalities. The intraindividual variance in the current data likely reflect the high perceptual variability that we have observed in this and previous studies in response to laser stimulation, perhaps owing to signal processing factors[18,36] and differences in heating rate caused by site-specific variations in pigmentation and nociceptor receptive fields.[26] Notably, intraindividual variance in perceptual response to laser stimulation is reported rarely[7,14,34,36]; however, refer to the work of Iannetti et al.[13] Therefore, we cannot compare the variance found here with previous literature. However, the findings of the current MAD analyses suggest that the variance in SPARS ratings was generally <10% of the scale's full range and is therefore unlikely to be problematic.

The SPARS has a curvilinear stimulus–response relationship, rather than the power relationship demonstrated previously for the conventional NRS or VAS. The reason could be that the SPARS provides a range of possible ratings for both painful and nonpainful events, whereas the conventional 0 to 100 NRS provides a range for only painful events. Alternatively, it could be based on the fact that we did not use the "relative ratings" approach of studies that found a power relationship. Finally, perhaps different instructions on how to use the scale yield the different curves. We deliberately avoided specifying the nature of pain to participants in order to optimize ecological validity. The original reports validating the NRS and the VAS[27–29] did not provide their instruction scripts. However, it is interesting that the highly systematic German Research Network on Neuropathic Pain Quantitative Sensory Testing procedure specifies the quality of sensation that should be rated >0 on a 0 to 100 scale.[30] For example, the script describes the heat pain threshold as the point at which "the impression of 'warmth' or 'heat' [changes] its quality towards an additional impression of a 'burning', 'stinging', 'drilling' or 'aching' sensation." We speculate that this type of explicit instruction as to the quality of sensation that should be labeled "pain" could introduce bias and result in a different stimulus–response curve from that when such explicit instruction is avoided. Indeed, visual inspection of Fig 11 suggests a flatter stimulus–response relationship at the lower end of the NRS than that reported in the original validation studies.[28,29]

Regardless of the explanation for the stimulus–response curve of the SPARS, its curvilinear characteristics may be advantageous, because the shape will lend itself to more straightforward analysis of data than the power profile of the conventional NRS and VAS. Experiment 2 showed a quadratic (rather than cubic) stimulus–response curve, which is unsurprising, considering the sample size for Experiment 2 (n = 7) and the findings of the stability analysis for Experiment 1, in which a small sample size (n = 6) yielded the highest variability in the best-fit model. The SPARS's lack of hysteresis and relatively low variance in ratings suggest that it performs well.

Experiment 2 showed limited agreement between the SPARS and the other 2 scales, an unsurprising but interesting finding. Specifically, agreement between the SPARS and the NRS was poor at the low end of the intensity range, whereas agreement between the SPARS and the SRS was poor at the upper end of the intensity range. This pattern likely reflects the extra dimensionality of the SPARS over the other 2 scales. In other words, the SPARS allows the ratings of stimulus intensities from no sensation to most intense pain you can imagine. Therefore, in the lower intensities, the SPARS continues, whereas the NRS ends (with some floor effect) at "no pain"; at the upper intensities, the SPARS continues, whereas the SRS ends (with some ceiling effect) at "pain." The finding that the ratings on the SPARS were generally lower than those on the NRS (even at higher stimulus intensities) and greater than those on the SRS (even at lower stimulus intensities) is likely also owing to the extra dimensionality of the SPARS—it captures a wider range of percept, likely leading to compression of the ratings into a narrower band on the SPARS (around 0: pain threshold) than on the 2 polar scales (which measure either painful or nonpainful ranges).

Strengths and Limitations

The current analytical approach is an important development on the original studies validating the widely used 0 to 100 NRS and VAS for pain.[27–29] Those studies described stimulus–response relationships by using group means, thus artificially obscuring both intra- and interindividual variation. In contrast, contemporary analytical methods such as those used here account for such important individual differences. The current analyses of hysteresis and rating variance also provide important information on the stability of the SPARS. This study also has limitations. Owing to safety considerations, we obtained limited data in the extreme upper range of the SPARS, perhaps affecting the shape of the curve. Careful briefing of participants on the SPARS's structure was likely important in facilitating participants' understanding of the scale, particularly for its verbal format. We tested the SPARS with healthy volunteers, using experimental laser stimulation, which excludes generalization to clinical pain and to other modalities. Future work could explore the clinical utility of the SPARS;

some of the limitations of conventional scales are not specific to the experimental context. The stability of the stimulus–response characteristics shown here suggests that participants did not have difficulty in using the scale, but whether that would also be the case in clinical practice remains to be determined.

Conclusions

The SPARS overcomes some important limitations of conventional ratings scales for pain and, by virtue of its curvilinear relationship to stimulus intensity, may lend itself to more straightforward data analysis techniques than the conventional ratings scales. It may be particularly well suited for determining pain threshold, clearly differentiating between nonpainful and painful stimuli, or for quantifying percepts beyond the painful range. Interestingly, its agreement with the conventional 0 to 100 NRS and a 0 to 100 (nonpainful) sensation scale is not strong, likely owing to the extra dimensionality of the SPARS. Examination of the SPARS's sensitivity to change in the experimental context and its utility for the clinical context seems warranted.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jpain.2018.10.006.

References

1. Bache SM, Wickham H: magrittr: A forward-pipe operator for R. Boston, MA: RStudio, 2014

2. Bates D, Maechler M, Bolker B, Walker S: Fitting linear mixed-effects models using lme4. J Stat Software 67:1-48, 2015

3. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307-310, 1986

4. Blyth FM, March LM, Brnabic AJM, Jorm LR, Williamson M, Cousins MJ: Chronic pain in Australia: A prevalence study. Pain 89:127-134, 2001

5. Canty A, Ripley B: boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20. https://cran.r-project.org/web/packages/boot/, 2017.

6. Davison AC, Hinkley DV: Bootstrap Methods and Their Applications. Cambridge, United Kingdom, Cambridge University Press, 1997

7. Derbyshire SWG, Nichols TE, Firestone L, Townsend DW, Jones AKP: Gender differences in patterns of cerebral activation during equal experience of painful laser stimulation. J Pain 3:401-411, 2002

8. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J: Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113:9-19, 2005

9. Fox J, Weisberg S: An {R} Companion to Applied Regression, 2nd ed. Thousand Oaks, CA, Sage, 2011.

10. Geraci M: Linear quantile mixed models: The lqmm package for laplace quantile regression. J Stat Software 57:1-29, 2014

11. Geraci M, Bottai M: Linear quantile mixed models. Stat Comput 24:461-479, 2014

12. Iannetti GD, Truini A, Romaniello A, Galeotti F, Rizzo C, Manfredi M, Cruccu G: Evidence of a specific spinal pathway for the sense of warmth in humans. J Neurophysiol 89:562-570, 2003

13. Iannetti GD, Zambreanu L, Cruccu G, Tracey I: Operculoinsular cortex encodes pain intensity at the earliest stages of cortical processing as indicated by amplitude of laser-evoked potentials in humans. Neuroscience 131:199-208, 2005

14. Iannetti GD, Zambreanu L, Tracey I: Similar nociceptive afferents mediate psychophysical and electrophysiological responses to heat stimulation of glabrous and hairy skin in humans. J Physiol 577:235-248, 2006

15. Kamerman P, Madden VJ, Bellan V, Catley M, Russek L, Camfferman D, Moseley GL: Analysis scripts and supplementary files: Was that painful or non-painful? The Sensation and Pain Rating Scale (SPARS) performs well in the experimental context. https://doi.org/10.6084/m9.figshare.6561743.v3

16. Koller M: robustlmm: An R package for robust estimation of linear mixed-effects models. J Stat Software 75:1-24, 2016

17. Kuznetsova A, Brockhoff PB, Christensen RHB: lmerTest package: Tests in linear mixed effects models. J Stat Software 82:1-26, 2017

18. Legrain V, Mouraux A: Activating selectively and reliably nociceptive afferents with concentric electrode stimulation: Yes we can! Provided that low stimulus intensities are used!. Clin Neurophysiol 124:424, 2013

19. Littman GS, Walker BR, Schneider BE: Reassessment of verbal and visual analog ratings in analgesic studies. Clin Pharmacol Ther 38:16-23, 1985

20. Loy A, Hofmann H: HLMdiag: A suite of diagnostics for hierarchical linear models in R. J Stat Software 56:1-28, 2014

21. Lüdecke D (2018). sjPlot: Data Visualization for Statistics in Social Science. https://doi.org/10.5281/zenodo.1308157, R package version 2.6.1, https://CRAN.R-project.org/package=sjPlot. Daniel Lüdecke and Carsten Schwemmer developed the software.

22. Madden VJ, Catley MJ, Bellan V, Russek LN, Camfferman D, Moseley GL: Was that painful? The FESTNRS—a numerical rating scale to avoid common pitfalls of experimental pain studies. In: Proceedings of the 16th World Congress on Pain. Yokohama, Japan, 2016

23. Madden VJ, Catley MJ, Grabherr L, Mazzola F, Shohag M, Moseley GL: The effect of repeated laser stimuli to ink-marked skin on skin temperature—recommendations for a safe experimental protocol in humans. PeerJ 4:e1577, 2016

24. Nieuwenhuis R, te Grotenhuis M, Pelzer B: influence. ME: Tools for detecting influential data in mixed effects models. R Journal 4:38-47, 2012

25. Pedersen TL: patchwork: The composer of ggplots, 2017, URL: https://github.com/thomasp85/patchwork

26. Plaghki L, Mouraux A: How do we selectively activate skin nociceptors with a high power infrared laser? Physiology and biophysics of laser stimulation. Neurophysiol Clin 33:269-277, 2003

27. Price DD, Bush FM, Long S, Harkins SW: A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. Pain 56:217-226, 1994

28. Price DD, Harkins SW: Combined use of experimental pain and visual analogue scales in providing standardized measurement of clinical pain. Clin J Pain 3:1-8, 1987

29. Price DD, McGrath PA, Rafii A, Buckingham B: The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 17:45-56, 1983

30. Rolke R, Andrews K, Magerl W, Treede RD: Quantitative Sensory Testing Investigator's Brochure, Version 2.1. Mannheim, Germany, University Medicine Mannheim, 2010

31. Seers T, Derry S, Seers K, Moore RA: Professionals underestimate patients' pain: A comprehensive review. Pain 159:811-818, 2018

32. Spruyt A, Clarysse J, Vansteenwegen D, Baeyens F, Hermans D: Affect 4.0. Experiment Psychol 57:36-45, 2010

33. Stevens SS: Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects. New York, NY, Routledge, 1975

34. Timmermann L, Ploner M, Haucke K, Schmitz F, Baltissen R, Schnitzler A: Differential coding of pain intensity in the human primary and secondary somatosensory cortex. J Neurophysiol 86:1499-1503, 2001

35. Wickham H: tidyverse: Easily install and load the 'Tidyverse', 2017, URL: http://tidyverse.tidyverse.org.

36. Wiech K, Lin C, Brodersen KH, Bingel U, Ploner M, Tracey I: Anterior insula integrates information about salience into perceptual decisions about pain. J Neurosci 30:16324-16331, 2010

37. Wilke CO: ggridges: Ridgeline plots in 'ggplot2', 2017. Bootstrap R (S-Plus) Functions. R package version 1.3-20. https://cran.r-project.org/web/packages/boot/, 2017.