Article:
Morley, S orcid.org/0000-0003-2972-7396 (2016) Bias and reliability in pain ratings. Pain, 157 (5). pp. 993-994. ISSN 0304-3959

https://doi.org/10.1097/j.pain.0000000000000448

Bias and reliability in pain ratings

Stephen Morley

University of Leeds

Correspondence

Stephen Morley PhD, Leeds Institute of Health Sciences, University of Leeds, 101 Clarendon Road, Leeds, LS2 9LJ, United Kingdom.

email: s.j.morley@leeds.ac.uk

tel; +44 113 343 2733

It is sometimes forgotten that measurement of psychological variables is an interaction between the individual, the test material and the context in which the measure is taken. In the case of clinically administered cognitive tests, such as the WAIS intelligence test, standardization of test administration, scoring and interpretation is explicit and testers require training to ensure that they do not introduce error into the measurement procedure and auditing is needed to ensure that tester performance does not drift over time. Standardization is essential for the calibration of the test. The benefits of standardization are that it eliminates systematic errors and facilitates the valid use of norms. While the importance of standardization is self-evidence for complex cognitive tests it is no less important for the many questionnaire measures and rating scales in common use. The British experimental psychologist E.C. Poulton concluded that 'quantitative subjective assessments are almost always biased, sometimes completely misleading' [5; 6] and making a simple rating can be a non-trivial event. The list of features known to affect ratings of experimentally controlled sensory and social stimuli is extensive and attempts have been made to integrate both response, stimulus and contextual characteristics into predictive models c.f. [4; 11]. Much is understood about factors contributing to subjective ratings of sensory intensity and the affective component of pain when the values of the pain stimuli are known, controlled and experimentally manipulated[1]. Determining sources of bias and unreliability in ratings of clinical pain is more problematic and less understood but understanding them has practical implications. Shannon Smith and her co-authors, in an article published in this issue of PAIN [9], argue that increasing the reliability of ratings should improve the precision of measurement and hence the discriminability between treatments with differential effectiveness.

Smith and her colleagues took a pragmatic approach to improving the reliability of daily self-reported clinical pain. They reasoned that provision of training and the addition of daily prompts i.e., enhancing the standardization protocol, should improve the reliability of pain ratings. Chronic pain patients with various diagnoses returned daily ratings of least, average and worst pain via the telephone. A control group (C) received no training, a second group received training (T) and the third group had both training and a daily reminder of the basic training instruction (T+). The expectation was that training would increase the reliability coefficients and decrease the number of inconsistent rankings of least, average and worst pain experienced each day. In the latter case the number of inconsistent rankings in the enhanced training group was 0.9% compared with 7% in the control group. Contrary to expectation there was no evidence that training had any impact on the reliability of ratings assessed using test-retest method. Indeed during the first post-training week it appeared that those in the T+ group were *less* reliable. Why might this be so?

Amongst the many possible factors there are two features worth considering. First is the application of the concept of reliability in this particular case. In classical test theory reliability is the ratio of the true score variance to the sum of true and error score variance, where the errors are only random. The presence of systematic errors will further contribute a mis-estimation of the true score. Furthermore it is assumed that the true score is fixed for any one estimation of reliability. As neither the true nor error score is directly observable various practical methods (test-retest, internal consistency and parallel forms) that rely on forms of the correlation are required to estimate it. If the true score fluctuates for valid reasons then the reliability estimated by the test-retest method will be compromised. Thus one would not consider estimating the reliability of an outcome measure by correlating pre-treatment and post-treatment test scores.

The second issue concerns how people might make pain judgments. Smith et al. rightly note that there are likely to be two salient sources of bias (systematic errors) affecting the estimation of pain; 1) scale anchoring effects and, 2) daily variations in context e.g. changes in mood, current pain, and physical setting. They sought to counter these by training patients to scale between the least and worst *imaginable pain*. They argued that two imagined pain anchors would be fixed relative to fluctuating personal experience (an untested but testable assumption). We do not know what these personal anchors were, how they were used, or whether the imagined experience is constant over time. How do people scale and remember imagined pain? More importantly what aspect of the imagine pain is reimagined and used as an anchor at the point of rating. We know that re-experiencing the somatosensory-intensity aspect of pain is rare, although people are able to give a summative evaluative assessment of remembered episodic pain and recall contextual information around the experience [2; 3]. Smith et al. attempted to minimize contextual effects on judgment by instructing patients to 'reflect again on the intensity and duration … (and) … focus only on pain intensity rather than other physical and emotional experiences like fatigue, stress and pains other than the targeted pain condition'. Even if it were possible to exclude these aspects of context from judgment reflection on intensity and duration for single pain episodes is known to be subject to systematic bias by the peak-end phenomenon [8]. The extent to which the peak-end phenomenon applies to recurrent episodic or variation in persistent pain is unknown, but other studies of recall of chronic pain experience show that expectation of the time-course of pain, experiential mismatch [7] and stable individual differences [10] are potential sources of systematic bias. In summary selecting and constructing a model for the recall pain gives several options. Which factors are appropriate in particular contexts is an issue for debate based on the likelihood of being able to generalise from different studies.

The paradox of the Smith et al study is that as a consequence of training, by removing a relatively large element of systematic bias, patients will show *greater* variability in their day-to-day ratings and thus reduce the estimate of reliability assessed by the conventional test-retest method. This may be what was observed in the early stages of the study in the T+ group. The article by Smith and colleagues raises important issues and highlights the tension between the need to answer pragmatic questions of relevance to clinical research and the need for better theoretical understanding of very basic measurement problems. This commentary has highlighted the fragmented research base of memory for pain and its implications; a topic which should arguably receive greater systematic investigation.

Conflict of interest

The author declares no conflict of interest.

References

[1] Gracely RH, Eliav E. Psychophysics of Pain. In: AI Basbaum, A Kaneko, GM Shepherd, G Westheimer, TD Albright, RH Masland, P Dallos, D Oertel, S Firestein, GK Beaucham, MC Bushnell, JH Kaas, E Gardner, editors. The Senses: A Comprehensive Reference, Vol. Vol. 5. New York: Academic Press, 2008. pp. 927-959.

[2] Katz J, Melzack R. Pain 'memories' in phantom limbs: review and clinical observations. Pain 1990;43(3):319-336.

[3] Morley S. Vivid memory for 'everyday' pains. Pain 1993;55(1):55-62.

[4] Parducci A, Wedell DH. The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. J Exper Psychol Hum Percept Perform 1986; 12(4):496-516.

[5] Poulton EC. Quantitative subjective assessments are almost always biased, sometimes completely misleading. Br J Psychol 1977;68:409-425.

[6] Poulton EC. Bias in quantifying judgments. Hove Erlbaum, 1989.

[7] Rachman S, Arntz A. The overprediction and underprediction of pain. Clin Psychol Rev 1991;11(4):339-355.

[8] Redelmeier DA, Katz J, Kahneman D. Memories of colonoscopy: a randomized trial. Pain 2003;104(1-2):187-194.

[9] Smith SM, Amtmann D, Askew RL, Gewandter JS, Hunsinger M, Jensen MP, McDermott MP, Williams M, Bacci ED, Burke LB, Chambers CT, Cooper SA, Cowan P, Desjardins P, Etropolski M, Farrar JT, Gilron I, Huang I-z, Katz M, Kerns RD, Kopecky EA, Rappaport BA, Resnick M, Strand V, Vanhove GF, Veasley C, Versavel M, Wasan AD, Turk DC, Dworkin RH. Pain intensity rating: results from an exploratory study of the ACTTION PROJECT system(c). Pain 2016.

[10] Walentynowicz M, Bogaerts K, Van Diest I, Raes F, Van den Bergh O. Was it so bad? The role of retrospective memory in symptom reporting. Health Psychol (in press).

[11] Watkinson P, Wood AM, Lloyd DM, Brown GDA. Pain ratings reflect cognitive context: A range frequency model of pain perception. Pain 2013;154 (5):743-749. doi: 10.1016/j.pain.2013.01.016.