

The published article is available on SpringerLink:

<http://link.springer.com/article/10.1007/s10865-009-9232-5>

The original publication is available at www.springerlink.com

**Taking into account the observers' uncertainty:
A graduated approach to the credibility of the patient's pain evaluation**

Patrice Rusconi, Paolo Riva, Paolo Cherubini, Lorenzo Montali

Department of Psychology, University of Milan-Bicocca, Italy

Author for correspondence is Patrice Rusconi:

Department of Psychology

Università Milano - Bicocca

1, Piazza Ateneo Nuovo

20126, Milano, Italy

e-mail: p.rusconi3@campus.unimib.it

fax: ++39 02 6448 3706

tel.: ++39 02 6448 3775

Acknowledgements

Authors wish to thank John Burns and three unknown referees for their pertinent comments and suggestions. We are also grateful to Marco Marelli for his suggestions on the analyses of the properties of the graduated approach used in Experiment 2.

Abstract

This article presents two experiments aiming to investigate the adoption of a graduated measure to describe credibility attribution by observers who evaluate patients' pain accounts. A total of 160 medical students were required to express a credibility judgment on the pain intensity level of hypothetical patients. We used 16 vignettes based on a factorial mixed-design. Within-participants factors were the reported pain, the presence of a physical sign, the patient's facial expression and the patient's gender, and between-groups factors were the patient's age and the geographical distribution of the patient's name. Results confirm the well-established tendency not to believe patients' self-reports and provide information regarding the evaluators' uncertainty. The findings suggest that a graduated measure is useful for assessing the degree of uncertainty of the observers and subtle effects of different factors upon the judgment of patient's pain.

Key words: Credibility; Pain Assessment; Graduated Measure; Miscalibration.

1. Introduction

There is a wealth of evidence indicating that patients' verbal reports regarding their pain are disbelieved, or put into question through the assessment of other clues, by observers, including healthcare professionals (e.g., Saxey, 1986; Scott, 1992; Thorn, 1997; Solomon, 2001; Clarke & Iphofen, 2005; Clarke & Iphofen, 2008). In particular, a vast range of studies point to the fact that there is a discrepancy between patients' self evaluations and judgments expressed by observers, often resulting in a phenomenon of "pain underestimation" (e.g., MacLeod *et al.*, 2001; Marquié *et al.*, 2003; Kappesser *et al.*, 2004; Kappesser *et al.*, 2006).

This well-established tendency might bear major consequences. On the one hand, it might affect the administration or the withholding of analgesics (Bell, 2000; Puntillo *et al.*, 2003). On the other hand, the disbelieved patient is more likely to experience an increased state of anxiety and negative emotions (Jacques, 1992; Clarke & Iphofen, 2008) which may undermine effective care actions. For this reason, it is worth clarifying the mechanisms underlying the phenomenon of disbelieving patients' accounts, as well as its boundaries, in order to improve pain control.

In the literature, the discrepancy between patients' and observers' judgments has often been investigated by using unidimensional pain scales – such as the numeric rating scale (e.g., Chibnall *et al.*, 1997) which measures the severity of the described pain by requiring a discrete response. It is made up of a fixed number of numerically tagged categories (usually eleven): the minimum value of the scale is labeled 'no pain', and the maximum value is labeled 'most intense pain imaginable'. Patients are instructed to indicate the scale number corresponding to the intensity of pain they actually feel.

In order to evaluate observers' judgments, Iafrati's criterion is commonly used: a score outside the range of ± 1 point compared to the patient's self-rating is considered either an overestimation or an underestimation (Iafrati, 1986; see also Kappesser *et al.*, 2006). The finding of the minimum clinically significant difference (MCSD) departs slightly from the Iafrati criterion assumption: Strout & Burton (2004) found a MCSD of 1.45 (SD \pm .84; 95% CI, 1.30-1.60), while

Kendrick and Strout (2005) reported a slightly different value equal to 1.39 (SD \pm 1.05; 95% CI, 1.27-1.51). Even though this method is well suited for establishing whether the pain intensity judged by the observer matches or not the level reported by the patient, it is not aimed at measuring the observer's degree of confidence, whereas this information might convey useful knowledge. For example, given a patient reporting a pain intensity of 8, and two observers, A and B, both reporting an estimated pain intensity of 6, the discrete measure classifies both observers as equally miscalibrated. However, if the judgment includes some measures of subjective confidence, it might result that observer A estimates the pain as ranging from 4 to 8, with an average of 6. By contrast, observer B might be less uncertain about her guess, estimating a pain intensity ranging from 5 to 7, with an average value of 6. This further piece of information helps understanding that A's and B's degrees of disbelief of the patient's rating are actually different: indeed A, but not B, includes the patient's rating within the interval that she deems plausible. Hence, our knowledge of the factors affecting the credibility of patients' self-reports might be fostered by considering it as a continuous dimension, graduated by the range of variability that the observer associates to her own judgment, instead of an all-or-nothing, discrete phenomenon. This approach seems suitable to judgments characterized by uncertainty such as pain estimates, considering the evidence showing that people can provide inconsistent estimates regarding the same pain stimulus, and that the same pain stimulus can be perceived very differently by different persons (Mader *et al.*, 2003).

In this study, we investigated whether using graduated estimates might actually improve our understanding of the observers' degree of belief in patients' self-reports. Observers were asked to indicate which pain intensity values they deemed credible, which were only partly credible, and which were not at all credible. As a consequence, the observers were classified as calibrated or miscalibrated, and a degree of confidence was associated to their calibration.

The aim was to devise a direct method for measuring the observer's confidence towards the patient's pain rating that also elicited some information about the window of variability associated to the judgment, and to compare the new method with the discrete one. As a secondary goal, we

investigated how different variables might affect the nature and quality of the confidence reported by the observers.

Our perspective is not to question the validity and usefulness of discrete methods of pain evaluation, but to seek for complementary measures that might enrich them, by providing supplementary information.

2. Method

2.1. Materials, design and procedure

We devised two tasks involving the same procedure, materials and experimental design, the only difference between them being the dependent variable and, consequently, part of the instructions. Each of a series of 16 paper-and-pencil vignettes featured a patient who came to the Emergency Department for a wart. During the visit, the patient reported having a headache. Participants were told the patient's name, gender, age, facial expression, the presence or absence of physical signs (sensitivity to light or noise), and the patient's pain self-rating. They were asked to evaluate the patient's pain.

For each of the 16 vignettes, participants were given a horizontal, 11-point, numeric rating scale on which the patient's rating was circled (see an example in the Appendix I). Then, they were asked to express their degree of belief in the values of the scale as estimates of the patient's pain level by blackening (Experiment 1) or ranking (Experiment 2) each of the boxes above the eleven units of the numeric rating scale. Booklets containing general instructions and the 16 vignettes were handed out individually to each participant.

The characteristics described in each vignette were orthogonally crossed in a factorial experimental mixed design (see Table I). Within-participants factors were: intensity of the reported pain (low, '3' on the 0-10 scale vs. high, '7' on the 0-10 scale), physical sign (sensitivity to light or noise vs. no sensitivity), facial expression (tense vs. relaxed), and patient's gender (female vs.

male). Between-groups factors were the patients' age (young vs. old) and the geographical distribution of the patients' names (Northern Italian vs. Southern Italian).¹

Insert Table I about here

We had two main reasons for choosing headache as a source of pain. Firstly, it is a “non-obvious” disease, which should induce observers to provide high ratings (Marquié *et al.*, 2003), and high ratings should highlight any miscalibration that low ratings might conceal (Chibnall & Tait, 2004). Yet, crucially, the non-obviousness of the cause of pain *per se* should not affect the extent of the miscalibration between patients' and observers' ratings (Marquié *et al.*, 2003). Secondly, headaches are a quite widespread type of pain whose main features are generally well-known not only to experts but also to novices, such as medical students.

3. Experiment 1

3.1. Participants

A total of 80 Italian medical students attending courses at university (52 female, 28 male; mean age 22.1 years, range 19-27 years) took part as volunteers in Experiment 1. The choice to use a sample of medical students was motivated by the preliminary nature of the study, whose generalizability may be tested in further researches. Moreover, some previous studies on similar topics recruited students as participants (e.g., undergraduate psychology students in Tait & Chibnall, 1994 and in Chibnall and Tait, 1995; first-year medical students in Chibnall *et al.*, 1997; undergraduate students in MacLeod *et al.*, 2001). Yet they provided an insightful basis for subsequent studies. Furthermore, Marquié *et al.* (2004) didn't find any difference in the degree of miscalibration between novices and experts.

¹ The latter factor was introduced on the basis of a preliminary qualitative study in which it emerged that healthcare professionals were influenced by the patients' place of origin (Montali *et al.*, 2009). Hence, in order to further test this aspect we decided to take into account the geographical distribution of the patients' names instead of tackling the better known ethnic disparities (Edwards *et al.*, 2001a; Edwards *et al.*, 2001b; Cintron & Morrison, 2006).

3.2. *Dependent variables*

In the first experiment we examined whether the patient's rating was included in a range of credible pain level values or not. Specifically, participants were instructed to indicate to what degree they believed each of the values of the scale; they did this by completely blackening the boxes above the points of the scale which they deemed credible as ratings of the patient's pain level, partially blackening the boxes above the points of the scale that they deemed only partly credible, and leaving the boxes blank for the values that they considered not credible.

3.3. *Results and analyses*

Inclusion of the patient's rating in the credibility range

Table II shows the mean percentages of inclusion of the patient's rating in each of the three intervals that participants could use to judge the credibility of the pain rating values.

Insert Table II about here

Overall, patient ratings fell within the credibility interval in 45.23% of the participants' judgments; they were within the partial credibility interval in 28.67% of cases and within the non-credibility interval in 26.09%. In other words, more than half of the credibility intervals did not include the patient's rating. In the following we will concentrate on the credibility interval as a measure of trust in the patients' pain accounts.

Insert Table III about here

Table III shows the results of non-parametric tests² on the rates of inclusion of the patient's rating in the credibility interval. Participants included the patient's rating in the range more frequently when the intensity of the reported pain was low than when it was high ($p < .00001$, $r = -.46^3$), and when the patient's face was tense more than when it was relaxed ($p < .005$, $r = -.23$). There was a trend to trust the patient more when he/she was described as having a physical sign ($p = .088$, $r = -.14$). No significant differences emerged regarding patient's gender, age and geographical distribution of the name.

Analysis of interactions

We performed an analysis of variance aimed to explore possible interactions. The robustness of this analysis even when analyzing dichotomous data has been shown by Lunney, 1970. As shown in Table III, the main effects were consistent with the results of the non-parametric analyses. The severity of the reported pain significantly interacted with the patient's facial expression ($F(1,75) = 78$; $p < .00001$; $\eta^2 = .112$): when the pain level was low, the respondents judged the patient's rating more credible if the facial expression was relaxed (mean .69) rather than if it was tense (mean .47); by contrast, when the pain level was high, credence was higher if the face was tense (.53) rather than if it was relaxed (.14). All these differences were significant at $p < .0001$. Even though the main effect of the physical sign was not significant, the interaction between the patient's rating and the presence of a physical sign was significant ($F(1,75) = 30.77$; $p < .00001$; $\eta^2 = .02$). This shows that when the physical sign was absent, participants believed more in a low patient's rating (.63) than in a high one (.25). The difference in credibility was smaller, and in the opposite direction, when a physical sign was present (.42 for high levels of pain vs. .53 for low levels of pain). All these differences were significant at least at $p < .01$. The intensity of the reported pain significantly interacted with the patient's age ($F(1,75) = 10.99$; $p < .01$; $\eta^2 = .014$): when the reported pain was

² In these analyses and in all the subsequent non-parametric analyses we performed a series of Wilcoxon tests when considering the within-participants variables, while we used the Mann-Whitney test when examining the between-groups factors.

³ The effect size r was computed, in these analyses and in all the subsequent non-parametric analyses, as follows: $r = Z / \sqrt{N}$, where N is the total number of observations (Field, 2005).

high participants believed more often to a young patient rather than to an old one (.39 vs. .28). The physical sign interacted with the gender ($F(1,75) = 8.64$; $p < .005$; $\eta^2 = .004$): when there was a physical sign, males were believed more (.51) than females (.44). The geographical distribution of the patients' name significantly interacted with the patient's age ($F(1,75) = 13.38$; $p < .001$; $\eta^2 = .029$). Old Northern patients were believed more than old Southern patients (.56 vs. .35).

Moreover, a three-way interaction was found to be significant between presence of physical sign, facial expression and gender ($F(1,75) = 5.93$; $p < .05$; $\eta^2 = .002$). When the face was tense and the physical sign was present males were believed more than females (.59 vs. .45) ($p < .001$).

The interval width

We then examined the interval width of the pain values deemed credible, corresponding to the number of scale points completely blackened. Table III shows the mean credibility interval widths and its non-parametric comparisons to the mean of inclusions of the patient's rating in the credibility interval. When the intensity of reported pain was high, the interval was wider than when it was low. The same occurred when a physical sign was present vs. absent, and when the face was tense vs. relaxed. Finally, there was a slight tendency to provide wider intervals when the patients were old rather than young.

The patient's rating within the credibility range

We assume that the distance between the midpoint of the credibility interval and patient's rating (operationalized as midpoint minus patient's rating, considering only the credibility intervals which included the patient's rating) can measure the participant's "mistrust" of the patient. If this distance is negative the observer underestimates the patient's rating, whereas if it is positive she overestimates it.

Insert Table IV about here

As shown in Table IV, this index was negative for each level of the variables we considered. A set of non-parametric tests showed that the participants underestimated the patient's rating more when the pain was high vs. low and when there was no physical sign. Also, there was a trend that old patients' ratings were underestimated more than young patients' ones.

3.4. Discussion

The aim of Experiment 1 was to test the plausibility of a new dependent variable – namely an interval estimate of the credibility of patient ratings.

Participants included the patient ratings in a range of credible values less than half of the time (45.23%), a result in keeping with previous studies on miscalibration. The inclusion of the patient ratings within the credibility range was influenced by the reported pain severity, the patient's facial expression and the presence of a physical sign as a clue. Both the p-value and the effect size provided converging evidence that the most important role was played by the patient's rating, followed by facial expression and by the physical sign. The interaction between physical sign and reported pain level shows that participants trusted more the high intensity pain reports when there was a physical sign, in line with the findings of previous studies (e.g., Loveman & Gale, 2000).

We then used the interval width as a measure of the uncertainty of the observer. This choice is formally justified by Shannon's (1948) information theory, that defines uncertainty (or entropy) as

$$E = - \sum_{i=1}^n p_i \log_2 p_i$$

where, given a set of n alternative and mutually exclusive hypotheses, p_i is the probability that each of them is the correct one.⁴ It follows that uncertainty is proportional to the number of plausible hypotheses, that, in our study, is expressed by the number of plausible pain levels indicated by the participants – that is, the interval width.⁵ The interval widths showed that participants were more uncertain when reported pain was high. Interestingly, they were more uncertain when there was a physical sign and a tensed face.

By integrating the two analyses, it emerges that, in line with the literature, participants were more prone to believe to low ratings (e.g., Prkachin *et al.*, 2007) and, at the same time, to exhibit more uncertainty about the high ratings. A reverse pattern was found concerning the physical sign and the facial expression. When the physical sign was present, the participants tended to believe more to high values of reported pain (again, consistently with previous studies, e.g., Chibnall *et al.*, 1997), but the interval width shows their cautiousness. The same occurred when the face was tense: high ratings of pain intensity were believed more (in line with previous researches, e.g., Igier *et al.*, 2007), but participants were more uncertain. This ambivalence is consistent with the literature, which reports that observers could be doubtful about the facial expression of pain, given the possible impression that patients may fake their expression (Craig *et al.*, 1991; Kappesser *et al.*, 2006). With a single, graduated measure we were able to capture the miscalibration phenomenon and the observer's cautiousness toward some "objective" clues.

Our third analyses – concerning the degree of mistrust toward the patient, and its direction – revealed that patients' pain was systematically underestimated, the more so when the reported pain was high, when there wasn't a physical sign and when the patients were old. These findings imply that, even when the patient's rating was considered credible, different degrees of trust could be detected – and they were affected by relevant variables.

⁴ When the n hypotheses are equiprobable the entropy reduces to the $\log_2(n)$.

⁵ In keeping with this reading, the literature on interval estimates indicates that a wide interval is less informative than a narrow interval (e.g., Yaniv & Foster, 1995; Yaniv & Foster, 1997; McKenzie *et al.*, 2008).

Overall the results of the Experiment 1 show that the interval measure proved capable of capturing the main known features of the miscalibration phenomenon, and, at the same time, provided some additional information concerning the observers' evaluation processes. A possible limitation of this measure is that it does not allow to rank each pain level according to its relative plausibility – thus, it does not allow to directly establish which level is the most plausible, as with the discrete measures. In the second experiment we addressed this issue and we compared the results of the use of a discrete vs. a graduated approach. We also explored further properties of a graduated measure which a discrete value cannot reveal.

4. Experiment 2

4.1. Participants

A total of 80 medical students attending courses at university (49 female, 31 male; mean age 22.4 years, range 19-35 years) volunteered to participate in Experiment 2.

4.2 Dependent variables

Participants were instructed to indicate their degree of belief regarding each of the values of the scale. They were asked to rank them by writing “1” in the box above the point of the scale they deemed most credible as a rating of the patient's pain level, “2” in the box above the next-most credible value of the scale, and so on. Participants were told that they couldn't assign the same rank to two or more values.

This new dependent variable provided a measure of the credibility of each scale point and thus allowed a comparison with the discrete measures. Specifically, we could pit the discrete perspective against the graduated approach both in the analysis of the credibility judgments given by the observers and in the description of pain miscalibration.

First, we analyzed the credibility judgments under the discrete vs. graduated perspectives. We considered the scale point which was rated “1” by a participant as equivalent to the discrete

value, whereas the graduated index was calculated as a weighted mean, specifically as the sum of the products of the scale points by the weights assigned to them divided by the sum of the weights (the formula and two examples of its application are provided in Appendix II).

We then used two different measures of pain miscalibration. First, we computed pain miscalibration as the value of the scale which was rated “1” by the participant, minus the patient’s rating. This miscalibration value is equivalent to the discrete measures used to highlight the discrepancy between patients and observers. The graduated measure that we devised encompassed – in a single value – the estimations made by participants on each point of the numeric rating scale as well as the uncertainty that was intrinsic to these estimations: it consisted in the weighted mean (which was used as a graduated measure in the previous analysis on the credibility judgments) minus the patient’s rating, divided by the standard deviation of the weighted mean itself (see Appendix III). Namely, it was a miscalibration measure standardized by the degree of uncertainty expressed by the participant: sign indicates underestimation vs. overestimation of the patient’s self-report, while the absolute value is proportional to the evaluator’s confidence that her own judgment is more appropriate than the patient’s self-report.

Finally, we analyzed some properties which are peculiar of the graduated measure and can reveal subtle information about the credibility attribution process.

4.3 Results and analyses

Credibility judgments

A series of non-parametric comparisons between the levels of each variable shows that the two measures of credibility agreed in highlighting the main effects of reported pain, physical sign and facial expression on observers’ judgments ($p < .00001$, $r \geq -.45$). However, the graduated index also brought out a statistically significant gender-based effect, namely the observers provided lower ratings when the patient was female than when the patient was male. This effect was not captured by the discrete measure (see Table V).

Insert Table V about here

Pain miscalibration

Table VI summarizes the non-parametric comparisons between the two pain miscalibration indices.

Insert Table VI about here

Both the discrete and the graduated measures indicated that the effects of reported pain intensity, physical sign and patient's facial expression were significant ($p < .00001$, $r = \geq -.45$), showing that when discrete miscalibration increased, the credibility of the patient's self-report (that is inversely proportional to the absolute value of the graduated measure) decreased. More interestingly, the two indices differed regarding the patient gender: while there were no significant differences in just *how much* the observer underestimated the patient's report (as shown by the discrete measure), the graduated measure showed a trend ($p = .053$, $r = -.16$) not to trust the patient's judgment more for female than for male patients.

An additional property of the graduated measure

By requiring the participants to rate each level of the pain scale we could also analyze a property that a discrete value does not allow to explore. Specifically, we summed on the one side the weights assigned to the values of the scale higher than the scale point ranked "1" and on the other side the weights attributed to the scale points lower than the one rated "1".⁶ We then

⁶ These two values are analog to the surface (i.e., the integral) to the right and left of the mode of a probability distribution; in this context, they corresponds to the overall amount of trust allocated to the right and to the left of the most trusted value.

computed the ratio between these two values, leaving at the numerator the one which included the weight assigned to the patient's rating. Accordingly, this measure yields values > 1 if the patient's rating is embedded in the most-trusted tail of judgments with respect to the observer's first-rank evaluation, and < 1 when the opposite is true. The mean ratio was equal to 2.19 and was ≥ 1.95 for each level of the variables included in our experimental design. This shows that, once they deemed a value as most credible, observers tended to attribute higher weights to the values on the side of the scale which included the patient's rating than to the pain levels on the opposite side, with respect to the value ranked "1". Psychologically, this means that evaluators adopted a mostly *inclusive* attitude, as opposed to an *exclusive* one – the latter possibly more dramatic than the former, in terms of pain underestimation (e.g., inclusive attitude: "the patient says 7; in my opinion she suffers 5, but not *less*"; exclusive attitude: "the patient says 7, in my opinion she suffers 5, and no *more*"). A series of Wilcoxon tests showed that the inclusive attitude was significantly more accentuated when the physical sign was absent than when it was present and when the patient's facial expression was relaxed rather than when it was tense (means: no physical sign 2.66, yes physical sign 2.32, $Z = -3.995$, $p < .0001$, $r = -.44$; relaxed face 2.51, tensed face 2.39, $Z = -4.348$, $p < .0001$, $r = -.48$).

4.4 Discussion

The aim of Experiment 2 was to compare the analyses of the credibility judgments provided by the participants and of pain miscalibration from a graduated vs. discrete perspectives. Furthermore, we analyzed some parameters that can be derived only by means of a graduated measure, and that can convey useful evaluations on the psychological properties of pain judgments. The discrete and graduated measures were in full agreement as far as the effects of pain intensity, physical sign and facial expression were concerned; yet, the graduated measures were more sensitive than the discrete one in detecting the effect of the patient gender, as follows:

- 1) the continuous measure of pain judgment (computed as the mean of pain levels weighted by their ranks) was significantly lower for female than for male patients (an effect that did not reach statistical significance for the discrete measure);
- 2) the standardized measure of miscalibration (weighted miscalibration / weighted standard deviation of pain judgments) showed that – even for similar miscalibration levels in absolute value – participants trusted female self-reports less than male self-reports.

Finally, the graduated measure caught the tendency of participants to be *inclusive*, as opposed to *exclusive*, with respect to the patients' ratings. They "softened" their judgments about the value that they deemed most credible by rating with (descending) high values the scale points going from their first-ranked judgment to the patient's rating. The pain levels opposite to the patients' ratings, with respect to the level ranked 1, were given very low-weight ranks. This inclusive attitude, even though it was always present, actually increased when there was no physical sign than when it was present and when the patient's facial expression was relaxed rather than when it was tense, possibly showing that – lacking strong clues of pain – the participants judgments were more anchored to the patients' ratings (even though they widely revised them).

5. Conclusions

Starting from the pioneer work of Marks and Sachar (1973), several studies over the last thirty-five years have shown that pain control is judged considerably inaccurate by the patients (e.g. Yates *et al.*, 1998; Visentin *et al.*, 2005; Breivik *et al.*, 2006). The ineffective management of pain has been ascribed to a range of factors, and among them a key role has been attributed to inaccurate assessment (e.g. McCaffery *et al.*, 2000; Bell, 2000; Puntillo *et al.*, 2003). In order to achieve an effective assessment process, the adoption of formal instruments (e.g. the numeric rating scale and the visual analogue scale) represents a necessary but insufficient step, since patients' self-reports, after being recorded, undergo an evaluation by healthcare professionals who may just as well accept

as dismiss the patient's rating (Jacques, 1992; Waterhouse, 1996; Thorn, 1997; Clarke & Iphofen, 2008).

Commonly, the judgment of credibility has been studied in terms of disbelief. This phenomenon has been pointed out by requiring both patients and practitioners to estimate pain intensity based on a scale. A discrepancy of ± 1 on the numeric rating scale has generally been considered the threshold beyond which one talks about miscalibration (Iafrazi, 1986; see also Kappesser *et al.*, 2006).

In the present study, we set out to critically analyze such an approach to the credibility process. The Iafrazi criterion assumption might fail to fit the process of credibility attribution if the latter is conceived not as an all-or-nothing phenomenon, but rather as a continuum. In this perspective, knowledge of the degrees of belief that the observers attribute to different pain levels might highlight differences in terms of cautiousness and uncertainty more than a discrete measure, considering that pain may be defined as a “continuous latent variable” (Pesudovs & Noble, 2005) .

In order to address these issues, we tested a new measure of credibility, by requiring participants to provide an interval of values of the numeric rating scale deemed credible (Experiment 1). We then performed a comparison between the analyses of credibility judgments and between the descriptions of pain miscalibration provided by this graduated measure and the discrete one (Experiment 2). Results of Experiment 1 confirmed the tendency not to believe patients' accounts. The graduated measure allowed estimating the observer's uncertainty – by means of the confidence interval width, that turned out to be affected by “objective” variables: namely, the intensity of reported pain, the facial expression and the presence of a physical sign.

The findings of Experiment 2 showed that a graduated measure might reveal subtle differences in both credibility judgments and pain miscalibration which would otherwise remain concealed when using a discrete measure. This specifically concerned the detection of a gender effect that was revealed only by the graduated measure. Finally, the graduated measure allowed to highlight a property of pain judgments which a discrete measure cannot detect, namely the *inclusive*

attitude of the observer's judgment: once they have selected their most plausible pain level, observers apportion more confidence to the part of the scale that includes the patient's rating, as opposed to the other one.

The present study has some limitations. First, participants evaluated vignettes describing fictitious patients, which cannot faithfully reproduce the complexity of real clinical settings. However, it should be noted that the vignettes are considered a valid means to assess the quality of care provided by physicians (Peabody *et al.*, 2000) and have been usefully adopted in several studies which focused on the impact of the patients' self-reports on pain judgments (e.g., Tait & Chibnall, 1994; Chibnall & Tait, 1995; Chibnall *et al.*, 1997; Chibnall *et al.*, 2000; MacLeod *et al.*, 2001; Elander *et al.*, 2006; Igier *et al.*, 2007; Marquié *et al.*, 2007; Kappesser & Williams, 2008). Our ability to replicate many previously known results concerning credibility judgments, pain miscalibration and the factors that may affect them supports the use of paper-and-pencil vignettes. Moreover, the limited ecological validity of the methodology we adopted did not prevent us from finding differences between a graduated measure and a discrete index both in the detection of the observers' credibility and in the description of pain miscalibration.

A further limitation of our study lies in the focus of our materials on patients coming to the Emergency Department. This means that our findings have yet to be generalized to inpatients and chronic pain. Furthermore, our measures should be tested on a sample of healthcare professionals in order to confirm their validity. Finally, the value of our graduated measures should be compared to other commonly used instruments whose outcome is discrete, such as the visual analogue scale.

Despite these limitations, this study should be considered as a preliminary attempt to provide evidence that the research on the credibility attribution to the patients' pain self-reports might take advantage of the adoption of a graduated measure, for it can provide information that other measures cannot bring out.

References

- Bell, F. (2000). A review of the literature on the attitudes of nurses to acute pain management. *Journal of Orthopaedic Nursing, 4*: 64-70.
- Breivik, H. H., Collett, B. B., Ventafridda, V. V., Cohen, R. R., & Gallacher, D. D. (2006). Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *European journal of pain, 10*(4): 287-333.
- Chibnall, J. T., and Tait, R. C (1995). Observer perceptions of low back pain: Effects of pain report and other contextual factors. *Journal of Applied Social Psychology, 25*(5): 418-439.
- Chibnall, J. T., Tait, R. C., & Ross, L. R. (1997). The Effects of Medical Evidence and Pain Intensity on Medical Student Judgments of Chronic Pain Patients. *Journal of Behavioral Medicine, 20*(3): 257-271.
- Chibnall, J. T., Dabney, A., & Tait, R. C. (2000). Internist judgments of chronic low back pain. *Pain Medicine, 1*(3): 231-237.
- Chibnall, J. T., & Tait, R. C. (2004). Comment on Marquie L. et al. Pain rating by patients and physicians: evidence of systematic pain miscalibration (Pain 2003;102:289-96). *Pain, 107*: 192-193.
- Cintron, A., & Morrison, R. S. (2006). Pain and ethnicity in the United States: a systematic review. *Journal of Palliative Medicine, 9*(6): 1454-1473.
- Clarke, K. A., & Iphofen, R. (2005). Believing the patient with chronic pain: a review of the literature. *British Journal of Nursing, 14*(9): 490-493.
- Clarke, K. A., & Iphofen, R. (2008). The effects of failing to believe patients' experience of chronic pain. *Nursing Times, 104*(8): 30-31.
- Craig, K. D., Hyde S. A., & Patrick C. J. (1991). Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain. *Pain, 46*: 161-171.

- Edwards R. R., Doleys D. M., Fillingim R. B. , & Lowery D. (2001a). Ethnic differences in pain tolerance: clinical implications in a chronic pain population. *Psychosomatic Medicine*, 63: 316–323.
- Edwards C. L., Fillingim R. B., & Keefe F. (2001b). Race, ethnicity and pain. *Pain*, 94: 133–137.
- Elander, J., Marczewska, M., Amos, R., Thomas, A., & Tangayi, S. (2006). Factors affecting hospital staff judgments about sickle cell disease pain. *Journal of Behavioral Medicine*, 29(2): 203-214.
- Field, A. (2005). *Discovering Statistics Using SPSS* (2nd ed.). London: Sage.
- Iafrati, N. S. (1986). Pain on the burn unit: Patient versus nurse perceptions. *Journal of Burn Care and Rehabilitation*, 7: 413-416.
- Igier, V., Mullet, E., & Sorum, P. C. (2007). How nursing personnel judge patients' pain. *European Journal of Pain*, 11(5): 542-550.
- Jacques, A. (1992). Do you believe I'm in pain? *Professional Nurse*, 7(4): 249-251.
- Loveman, E., & Gale, A. (2000). Factors influencing nurses' inferences about patient pain. *British Journal of Nursing*, 9(6): 334-337.
- Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: an empirical study. *Journal of Educational Measurement*, 7(4): 263-269.
- Kappesser, J., Williams, A. C. d. C., & Prkachin, K. (2004). What makes clinicians underestimate pain? *The Journal of Pain*, 5(3, Supplement): S128.
- Kappesser, J., Williams, A. C. d. C., & Prkachin, K. M. (2006). Testing two accounts of pain underestimation. *Pain*, 124: 109-116.
- Kappesser, J., & Williams, A. C. d. C. (2008). Pain judgements of patients' relatives: examining the use of social contract theory as theoretical framework. *Journal of Behavioral Medicine*, 31: 309-317.

- Kendrick, D. B., & Strout, T. D. (2005). The minimum clinically significant difference in patient-assigned numeric scores for pain. *The American Journal of Emergency Medicine, 23*(7): 828-832.
- MacLeod, F. K., LaChapelle, D. L., Hadjistavropoulos, T., & Pfeifer, J. E. (2001). The Effect of Disability Claimants' Coping Styles on Judgments of Pain, Disability, and Compensation: A Vignette Study. *Rehabilitation Psychology, 46*(4): 417-435.
- Mader, T. J., Blank, F. S. J., Smithline, H. A., & Wolfe, J. M. (2003). How Reliable Are Pain Scores? A Pilot Study of 20 Healthy Volunteers. *Journal of Emergency Nursing, 29*(4): 322-325.
- Marks, R.M. & Sachar, E.J. (1973). Undertreatment of medical inpatients with narcotic analgesics. *Annals of internal medicine, 78*, 173-181.
- Marquié, L., Raufaste, E., Lauque, D., Mariné, C., Ecoiffier, M., & Sorum, P. (2003). Pain rating by patients and physicians: evidence of systematic pain miscalibration. *Pain, 102*(3): 289-296.
- Marquié, L., Raufaste, E., Lauque, D., Mariné, C., Ecoiffier, M., & Sorum, P. (2004). Further results about pain rating by patients and physicians: reply to Chibnall and Tait. *Pain 107*: 194-195.
- Marquié, L., Sorum, P. C., & Mullet, E. (2007). Emergency physicians' pain judgments: cluster analyses on scenarios of acute abdominal pain. *Quality of Life Research, 16*(7): 1267-1273.
- McCaffery, M., Rolling Ferrell, B., & Pasero, C. (2000). Nurses' Personal Opinions About Patients' Pain and Their Effect on Recorded Assessments and Titration of Opioid Doses. *Pain Management Nursing, 1*(3): 79-87.
- McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes, 107*(2): 179-191.
- Montali, L., Colombo, M., & Riva, P. (2009). Theories And Practices In Pain Management: A Research On Doctors' Representations. *Psicologia della Salute, 1*, 33-56.

- Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of Vignettes, Standardized Patients, and Chart Abstraction. A Prospective Validation Study of 3 Methods for Measuring Quality. *The Journal of the American Medical Association*, 283(13): 1715-1722.
- Pesudovs, K., & Noble, B. A. (2005). Improving Subjective Scaling of Pain Using Rasch Analysis. *The Journal of Pain*, 6(9): 630-636.
- Prkachin, K. M., Solomon, P. E., & Ross, J. (2007). Underestimation of Pain by Health-Care Providers: Towards a Model of the Process of Inferring Pain in Others. *Canadian Journal of Nursing Research*, 39(2): 88-106.
- Puntillo, K., Neighbor, M., & Nixon, R. (2003). Accuracy of Emergency Nurses in Assessment of Patients' Pain. *Pain Management Nursing*, 4(4): 171-175.
- Saxey, S. (1986). The nurse's response to postoperative pain. *Nursing: The add on Journal of Clinical Nursing*, 3(10): 377-381.
- Scott, I. (1992). Nurses' attitudes to pain control and the use of pain assessment scales. *British Journal of Nursing*, 2(1): 11-14.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379-423, 623-656.
- Solomon, P. (2001). Congruence between health professionals' and patients' pain ratings: a review of the literature. *Scandinavian Journal of Caring Sciences*, 15(2): 174-180.
- Strout, T. D., & Burton, J. H. (2004). Clinically significant change in physician-assigned numeric pain rating scale scores. *American Journal of Emergency Medicine*, 22(3): 243-244.
- Tait, R. C., and Chibnall, J. T. (1994). Observer perceptions of chronic low back pain. *Journal of Applied Social Psychology*, 24(5): 415-431.
- Thorn, M. (1997). A survey of nurses' attitudes towards the assessment and control of postoperative pain. *Journal of Orthopaedic Nursing*, 1: 30-38.

- Visentin, M., Zanolin, E., Trentin, L., Sartori, S., & Marco, R. d. (2005). Prevalence and treatment of pain in adults admitted to Italian hospitals. *European Journal of Pain*, 9: 61-67.
- Waterhouse, M. (1996). Why pain assessment must start with believing the patient. *Nursing Times*, 92(38): 42-43.
- Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: an accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, 124(4): 424-432.
- Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10: 21-32.
- Yates, P., Dewar, A., Edwards, H., Fetiman, B., Najman, J., Nash, R., et al. (1998). The Prevalence and perception of pain amongst hospital in-patients. *Journal of Clinical Nursing*, 7: 521-530.

Appendices

Appendix I

A sample vignette.

You are going to be presented with some vignettes about patients who come into the Emergency Department for a problem on the back of the hand. Your diagnosis is that it is a wart.

During the visit, the patients also report having a headache (with this expression we do not mean a headache deriving from injuries, cancer, ictus, etc., but a primary headache).

At the end of each vignette a pain intensity scale is reported (0 = no pain, 10 = maximum pain) on which the value actually indicated by the patient is circled.

Experiment 1 version: Your task is to indicate, for each vignette, which pain intensity values you deem credible, which in your view are only partly credible and which are not all credible.

In detail:

- completely blacken the boxes above the pain intensity values that you deem *credible* → 
- partially blacken the boxes above the pain intensity values that you deem *only partly credible* → 
- leave blank the boxes above the pain intensity values that you deem *not credible* → 

Experiment 2 version: Your task is to numerically rate the values of the scale, by writing “1” beside the value that you deem most credible, “11” beside the value that you deem not credible at all, and using the intermediate numbers to indicate your credibility towards the remaining values.

In detail, for each vignette you must:

- write “1” in the box above the value which according to you most likely coincides with the patient’s actual pain intensity;
- write “2” in the box above the value that you judge to be second in terms of the patient’s actual pain intensity;
- write “3” in the box above the value that you judge to be third in terms of the patient’s actual pain intensity;

and so on until you get to number “11”.

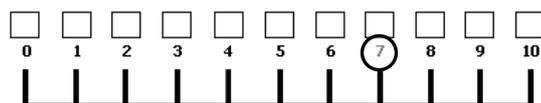
Remember that you can only assign each number to each value once.

Carmen Battaglia is a 66-year-old woman. When she tells you that she has a headache you notice that her face appears relaxed. The patient reports sensitivity to light.

When you ask her to indicate the intensity of her pain on a 0-10 scale, Carmen Battaglia rates her pain as a 7.

Experiment 1 version: Overall, which pain intensity levels would you deem credible (blacken corresponding boxes), which would you judge only partly credible (partially blacken corresponding boxes) and which would you judge not credible at all (leave corresponding boxes blank) for patient Carmen Battaglia?

Experiment 2 version: Numerically rate the values of the scale below, by writing “1” above the value that you deem most credible, “11” above the value that you deem not credible at all, and using intermediate numbers to indicate your credibility towards the remaining values. Remember that you can only assign each number to each value once.



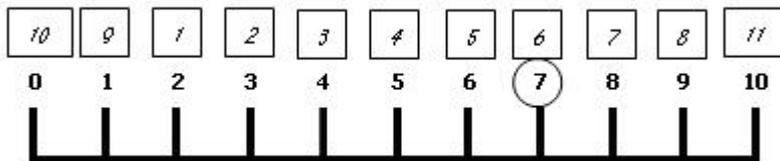
Appendix II

The formula and two examples of the calculation of the graduated index used in Experiment 2 to analyze the credibility judgments provided by the observers.

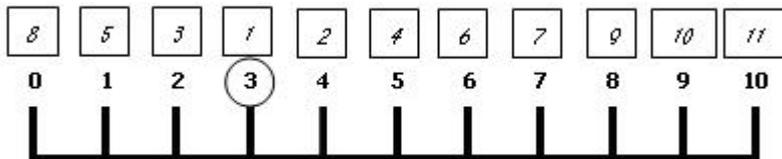
$$\sum n_i \times r_i / \sum r_i$$

where, n_i are the scale points and r_i are the weights assigned to them by the participants.

Example 1 – Index equal to 4.47



Example 2 – Index equal to 3.75



Appendix III

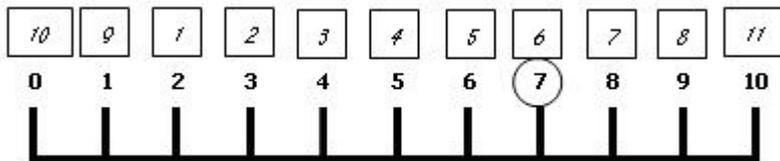
The formula of the standardized measure of pain miscalibration used in Experiment 2 and two examples of its calculation.

$$[(\sum n_i \times r_i / \sum r_i) - k] / \sigma$$

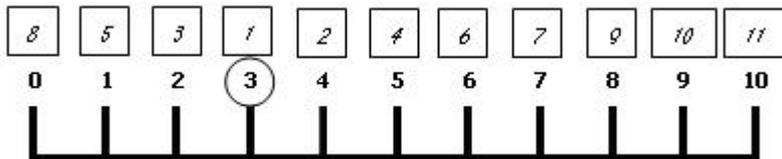
where, n_i are the scale points, r_i are the weights assigned to them by the participants, k is the patient's rating and σ is:

$$\sigma = \sqrt{\sum \{ [n_i - (\sum n_i \times r_i / \sum r_i)]^2 \times r_i \} / 54}$$

Example 1 – Index equal to -1.1



Example 2 – Index equal to .34



Tables

Table I. Outline of the experimental design.

	Variable	Level 1	Level 2
<i>Within factors</i>	patient's rating	3 (low)	7 (high)
	physical sign	absent	present
	facial expression	relaxed	tensed
	patient's gender	female	Male
<i>Between factors</i>	patient's age	25-40 (young)	65-80 (old)
	geographical distribution of	Southern	Northern
	patient's name		

Table II. Experiment 1. Mean percentages of inclusion of the patient’s rating in each of the three credibility intervals (in brackets the actual number of inclusions).

		Non-credibility Interval	Partial Credibility Interval	Credibility Interval
intensity of reported pain	low	16.09% (103)	26.56% (170)	57.34% (367)
	high	36.09% (231)	30.78% (197)	33.13% (212)
physical sign	no	28.13% (180)	28.28% (181)	43.59% (279)
	yes	24.06% (154)	29.06% (186)	46.88% (300)
patient facial expression	relaxed	31.25% (200)	27.81% (178)	40.94% (262)
	tensed	20.94% (134)	29.53% (189)	49.53% (317)
patient gender	female	25.31% (162)	30.16% (193)	44.53% (285)
	male	26.88% (172)	27.19% (174)	45.94% (294)
patient age	old	28.13% (180)	27.19% (174)	44.69% (286)
	young	24.06% (154)	30.16% (193)	45.78% (293)
geographical distribution of patient name	south	27.03% (173)	29.84% (191)	43.13% (276)
	north	25.16% (161)	27.50% (176)	47.34% (303)
Total		26.09% (2004)	28.67% (2202)	45.23% (3474)

Table III. Experiment 1. Mean of inclusions of the patient's rating in the credibility interval compared with mean credibility interval width. We report non-parametric and parametric (ANOVA) comparisons.

		Mean of inclusions in the credibility interval	Z (P-value)	r	F (P-value)	Eta-Square (η^2)	Interval width	Z (P-value)	r
intensity of reported pain	low	4.65*	-5.745	-.46	59.52 (<.00001)	.077	2.72	-3.636 (<.001)	-.29
	high	2.68*	(<.00001)				3.02		
physical sign	no	3.53*	-1.705	-.14	2.87 (= .095)	.001	2.73	-4.501 (<.00001)	-.36
	yes	3.80*	(= .088)				3.02		
patient facial expression	relaxed	3.32*	-2.840	-.23	8.88 (<.005)	.009	2.68	-4.314 (<.0001)	-.34
	tensed	4.01*	(<.005)				3.07		
patient gender	female	3.61*	-.318	-.03	.53 (= .469)	.000	2.89	-.473 (= .636)	-.04
	male	3.72*	(= .751)				2.86		
patient age	old	7.33**	-.306	-.03	.00 (= .964)	.000	3.10	-1.694 (= .090)	-.19
	young	7.33**	(= .760)				2.65		
geographical distribution of patient name	south	7.54**	-.764	-.09	.49 (= .485)	.001	2.97	-.621 (= .535)	-.07
	north	7.14**	(= .445)				2.77		

* range 0-8

** range 0-16

Table IV. Experiment 1. The results of non-parametric analyses on the position of the patient's rating within the credibility interval.

		Interval midpoint – patient's report	Z (<i>P</i> -value)	<i>r</i>
intensity of reported pain	low	-.02	-5.323 (< .00001)	-.46
	high	-.74		
physical sign	no	-.37	-3.733 (< .001)	-.31
	yes	-.13		
patient facial expression	relaxed	-.26	-.210 (= .834)	-.02
	tensed	-.25		
patient gender	female	-.25	-.476 (= .634)	-.04
	male	-.22		
patient age	old	-.34	-1.902 (= .057)	-.21
	young	-.14		
geographical distribution of patient name	south	-.30	-.698 (= .485)	-.08
	north	-.17		

Table V. Experiment 2. Comparison between the two indices of the credibility judgments according to the results of a series of non-parametric tests.

		discrete measure*	Z (P -value)	r	graduated measure**	Z (P - value)	r
intensity of reported pain	low	3.56	-6.479	-.51	4.32	-7.075	-.56
	high	4.74	(< .00001)		5.08	(< .00001)	
physical sign	no	3.7	-5.723	-.45	4.43	-6.708	-.53
	yes	4.6	(< .00001)		4.97	(< .00001)	
patient facial expression	relaxed	3.44	-6.817	-.54	4.25	-6.950	-.55
	tensed	4.87	(< .00001)		5.16	(< .00001)	
patient gender	female	4.1	-1.654	-.13	4.67	-2.161	-.17
	male	4.2	(= .098)		4.74	(< .05)	
patient age	old	4.1	-.525	-.06	4.67	-.019	.00
	young	4.2	(= .600)		4.73	(= .985)	
geographical distribution of patient name	south	4.15	-.742	-.08	4.71	-.390	-.04
	north	4.15	(= .458)		4.69	(= .697)	

* Value of the scale which was rated "1" by the participant.

** Sum of the products of the scale points by the weights assigned to them divided by the sum of the weights.

Table VI. Experiment 2. Comparison between the two pain miscalibration indices according to the results of a series of non-parametric tests.

		discrete measure*	Z (P -value)	<i>r</i>	graduated measure**	Z (P - value)	<i>r</i>
intensity of reported pain	low	.56	-7.704	-.61	.55	-7.574	-.61
	high	-2.26	(< .00001)		-.84	(< .00001)	
physical sign	no	-1.3	-5.723	-.45	-.26	-6.503	-.53
	yes	-.4	(< .00001)		-.02	(< .00001)	
patient facial expression	relaxed	-1.56	-6.817	-.54	-.33	-6.715	-.54
	tensed	-.13	(< .00001)		.05	(< .00001)	
patient gender	female	-.9	-1.654	-.13	-.16	-1.936	-.16
	male	-.8	(= .098)		-.13	(= .053)	
patient age	old	-.9	-.525	-.06	-.15	-.203	-.02
	young	-.8	(= .600)		-.13	(= .839)	
geographical distribution of patient name	south	-.85	-.742	-.08	-.14	-.847	-.1
	north	-.85	(= .458)		-.15	(= .397)	

* Value of the scale which was rated "1" by the participant, minus the patient's rating.

** The weighted mean minus the patient's rating, divided by the standard deviation of the weighted mean itself.