SPINE Volume 30, Number 16, pp E455–E458
©2005, Lippincott Williams & Wilkins, Inc.

Interrater Reliability of Scoring of Pain Drawings in a Self-Report Health Survey

Rosie J. Lacey, PhD, Martyn Lewis, PhD, Kelvin Jordan, PhD, Clare Jinks, PhD, and Julius Sim, PhD

Study Design. Study of interrater reliability.

Objective. To assess the interrater reliability of data from pain drawings scored by multiple raters and the consistency of the subsequent classification of cases of widespread pain.

Summary of Background Data. In large health surveys, pain drawings used to capture self-reported pain, and to classify cases of widespread pain, are often scored by several raters. The reliability of multiple rater scoring of pain drawings has not been investigated.

Methods. As part of a postal survey sent to adults 50 years and older, subjects were asked to shade their pain on a blank body manikin. The first 50 pain drawings in which respondents had shaded pain were selected for this study. Eight nonclinical staff were trained to score pain drawings using transparent templates divided into 50 body areas. Interrater reliability was assessed by comparing the scoring of "pain" or "no pain" for all 50 areas of each pain drawing.

Results. Complete scoring agreement among all raters was observed for at least 78% of pain drawings across all body areas (kappa $>$ 0.60). The raters had complete agreement in 42 of 50 areas in 90% or more of pain drawings. From the raters' scoring of pain areas, there was complete agreement on the presence or absence of widespread pain for 49 of 50 pain drawings (98% agreement, Kappa = 0.98).

Conclusions. This study shows that multiple raters, with training and guidelines, can reliably score pain drawings, and high consistency in the subsequent classification of cases of widespread pain can be obtained from such data.

Key words: health survey, interrater, manikin, pain drawing, pain measurement, rater, reliability, scoring. **Spine 2005;30:E455–E458**

Pain drawings (or body manikins) are frequently used as a method of pain assessment in musculoskeletal research. In self-report health surveys, they are often included as a simple screening instrument to assess the location of pain or to estimate the prevalence of pain in certain body areas.[1–3] Subjects are usually asked to shade their pain within the outlines of front and back views of a blank body manikin. The presence or absence of pain is assessed by placing a transparent template, divided into a number of defined body areas, over the completed pain drawing, and observing any shading within these areas. Previous studies have found good test-retest reliability of completion of pain drawings by subjects in different settings.[2–6] Completed pain drawings can also be used to determine the presence of widespread pain according to a predetermined definition.[7]

In population surveys, scoring of completed pain drawings is often performed manually by a data entry clerk, although for a large survey, it would not be cost-effective (either in time or burden on staff) to use a single rater. However, the use of multiple raters raises the issue of reliability of scoring because the quality of data is dependent on how consistently raters score the shading on pain drawings (*i.e.*, interrater reliability). In studies of self-report surveys, there is little evidence regarding the reliability of scoring of pain drawings by multiple raters. Margolis *et al*[1] reported good agreement between 2 raters assessing pain drawings from a postal survey, although the subjects had existing low back pain, and information on the reliability of scoring for individual pain areas was not included. Furthermore, although many studies use pain drawings to distinguish cases of widespread pain from those with regional pain, to our knowledge, the reliability of scoring of widespread pain from completed pain drawings has not been researched. Therefore, the aims of the current study were to: (1) evaluate the interrater reliability of scoring of pain drawing data by 8 different raters from a postal survey of the general population, (2) report any variation in scoring across different areas of the pain drawing, and (3) assess the consistency of the subsequent classification of widespread pain from the pain drawing data.

Materials and Methods

Study Design. A postal survey of knee pain was sent to all adults 50 years and older (n = 8995), registered at 3 general practices in North Staffordshire, United Kingdom.[8] The questionnaire contained a blank body manikin asking respondents to "shade in the diagram any pain that has lasted for 1 day or longer in the past 4 weeks." To assess interrater reliability of scoring of completed pain drawings, we selected the first 50 questionnaires that had at least one area of the manikin shaded.

From the Primary Care Sciences Research Centre, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom.

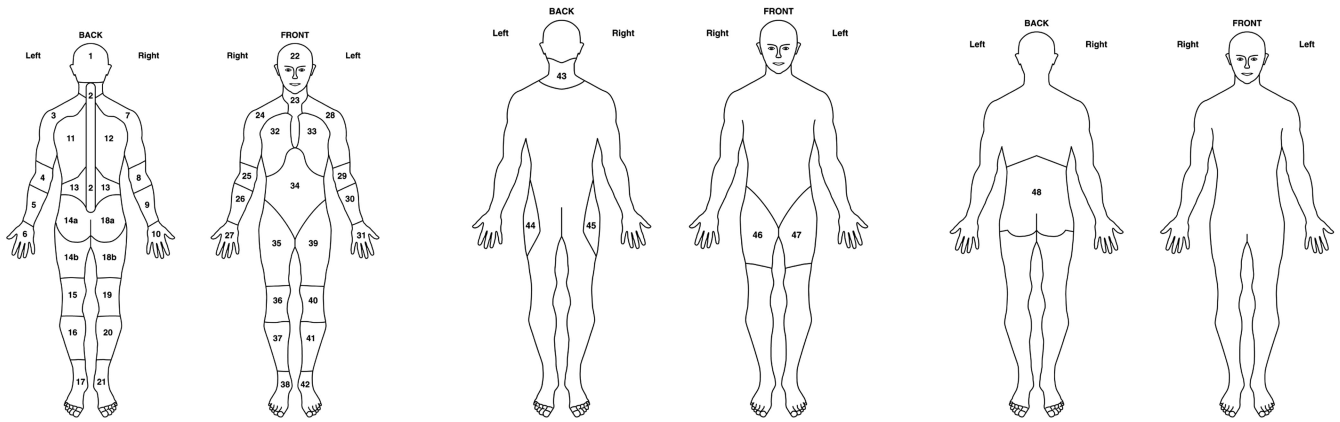
Acknowledgment date: April 20, 2004. First revision date: September 23, 2004. Acceptance date: February 15, 2005.

Supported by The West Midlands NHSE New Blood Research Fellowship Committee and the Haywood Rheumatism Research and Development Foundation, North Staffordshire.

The manuscript submitted does not contain information about medical device(s)/drug(s).

Institutional and Foundation funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Address correspondence and reprint requests to Rosie J. Lacey, PhD, Primary Care Sciences Research Centre, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom; E-mail: r.lacey@cphc.keele.ac.uk



Widespread Pain

Neck and Hip Pain

Lower Back Pain

Figure 1. Templates used to score pain drawings.

Pain drawings were scored using 3 transparent templates divided into a total of 50 body areas (Figure 1). The body areas were based on those used in the Manchester definition of widespread pain.[7] Eight nonclinical staff (*i.e.,* the raters) scored all 50 areas of the 50 pain drawings. One rater (rater 1) had previous experience; the other 7 (raters 2−8) were trained on how to score pain drawings using a standardized protocol currently operative within the research center. Guidelines were: (1) any mark (*e.g.*, shading, scribble, cross or line) within a template area to be scored as "pain" present; (2) any arrow touching a template area to be scored as "pain" present; and (3) any marks or arrows outside (*i.e.*, not touching) the pain drawing to be ignored. Raters were talked through 2 examples; these were not included in the 50 pain drawings selected for the study. The raters scored all pain drawings independently and were blinded to each other's responses. Three sets of templates were used for the scoring.

Statistical Analysis. The sum of areas scored positively by each rater for the presence of shading in each selected pain drawing was used in an analysis measuring the global interrater agreement among the 8 raters using an intraclass correlation coefficient (2,1) (2-way random effects model).[9] This procedure was performed to establish a crude measure of overall consistency in the pattern of scoring among the raters, which was not area specific. In addition, to investigate any inconsistencies in scoring across specific areas of the pain drawings, we performed analyses of interrater agreement for each area. This procedure was performed in 2 ways. First, we calculated the complete agreement (defined as the percentage of pain drawings for which all 8 raters agreed on the scoring) for each area. Second, we calculated the kappa statistic (κ) for multiple ratings to evaluate the chance-corrected agreement.[10] The lower limit of the 99% 1-sided confidence interval for κ was also derived. Furthermore, according to Shrout's[11] 1998 classification of κ values, any areas with less than substantial agreement (*i.e.*, $\kappa < 0.81$) between the 8 raters were rescreened by one of the researchers (R.J.L.) to assess whether the disagreement was related to: (1) obvious error (*i.e.*, incorrect scoring); or (2) interpretation error (*i.e.*, interpretation of shadings just touching or crossing lines that divided areas).

Finally, we assessed the reliability of classifying widespread pain, as derived from the raters' scoring of individual body areas of the pain drawings. This procedure was performed by applying the criteria used in the Manchester definition of widespread pain[7] to the scoring of the pain drawings. The complete agreement and κ were calculated. Statistical analyses were performed using SPSS (version 11.0, SPSS Inc., Chicago, IL, 2001) and PEPI software (version 4.0x).[12]

■ Results

The mean number of positive pain areas recorded by the 8 raters was: 10.2 (raters 3, 4, and 6), 10.7 (rater 5), 10.9 (raters 1, 2, and 8), and 11.3 (rater 7). The value for the global interrater agreement, based on the number of positive pain areas recorded, was intraclass correlation coefficient (2,1) = 0.99. A summary of pain prevalence for individual areas of the pain drawings is presented in Table 1. The areas with the least shading were the back of the head (area 1) and left forearm (area 30) regions, while the most common areas of pain recording were the spine (area 2) and lower back (area 48) regions. As an illustration of the variation in scoring, Table 1 also shows the range of prevalence recorded by the raters for each body area. The range of prevalence of pain exceeded 10% for only 2 areas: back of the head (area 1, range 22%, from 4% to 26%) and back of left hip (area 44, range 12%, from 24% to 36%).

Complete scoring agreement among the 8 raters across the body areas ranged from 78% (39 of 50) of pain drawings to 100% (50 of 50); the range of κ values was from 0.61 to 1.00. In a total of 42 areas, there was complete agreement in at least 90% of pain drawings. The 8 areas with complete agreement in less than 90% of pain drawings were areas 1, 2, 11, 13, 43, 44, 45, and 46. The area of most disagreement was the back of the head (area 1), which negatively skewed the range of agreement figures. After excluding area 1, there was complete agreement among the 8 raters in at least 82% (41 of 50) of the 50 selected pain drawings for the remaining 49 areas ($\kappa \geq 0.81$).

Because area 1 was the only area with less than substantial agreement (*i.e.*, $\kappa < 0.81$)[11] among the 8

Table 1. Prevalence of Pain by Body Area Scored by 8 Raters Across 50 Pain Drawings

Area*	Median % Prevalence of Pain Reported by 8 Raters (range)	No. of Pain Drawings Agreed on by 8 Raters (%)	κ (99% confidence limit)†
1	6 (4–26)	39 (78)	0.61 (0.55)
2	58 (52–60)	44 (88)	0.92 (0.86)
3	32 (30–32)	49 (98)	0.99 (0.93)
4	16 (16–16)	48 (96)	0.96 (0.90)
5	8 (8–10)	49 (98)	0.97 (0.91)
6	10 (8–12)	48 (96)	0.94 (0.88)
7	36 (34–38)	45 (90)	0.92 (0.86)
8	16 (14–16)	49 (98)	0.98 (0.92)
9	16 (14–18)	45 (90)	0.81 (0.75)
10	16 (12–16)	48 (96)	0.93 (0.87)
11	24 (20–28)	42 (84)	0.83 (0.77)
12	24 (20–26)	45 (90)	0.91 (0.85)
13	22 (18–28)	42 (84)	0.84 (0.78)
14a	14 (14–16)	48 (96)	0.97 (0.91)
14	14 (14–16)	49 (98)	0.98 (0.92)
15	22 (20–22)	49 (98)	0.99 (0.93)
16	14 (14–16)	49 (98)	0.98 (0.92)
17	14 (14–18)	47 (94)	0.89 (0.83)
18a	40 (40–42)	49 (98)	0.98 (0.92)
18	14 (14–14)	50 (100)	1.00 (0.94)
19	18 (18–18)	50 (100)	1.00 (0.94)
20	10 (8–10)	49 (98)	0.97 (0.91)
21	10 (8–10)	49 (98)	0.97 (0.91)
22	10 (10–12)	49 (98)	0.97 (0.91)
23	10 (8–12)	48 (96)	0.93 (0.87)
24	24 (24–26)	49 (98)	0.99 (0.93)
25	14 (14–16)	49 (98)	0.98 (0.92)
26	16 (16–18)	49 (98)	0.98 (0.92)
27	21 (20–22)	49 (98)	0.97 (0.91)
28	15 (12–16)	46 (92)	0.88 (0.82)
29	12 (10–12)	49 (98)	0.96 (0.90)
30	7 (6–8)	48 (96)	0.88 (0.82)
31	14 (14–16)	49 (98)	0.98 (0.92)
32	12 (12–14)	48 (96)	0.93 (0.87)
33	12 (12–12)	50 (100)	1.00 (0.94)
34	20 (20–22)	49 (98)	0.98 (0.92)
35	26 (24–30)	47 (94)	0.95 (0.91)
36	34 (32–36)	47 (94)	0.96 (0.90)
37	16 (14–20)	46 (92)	0.88 (0.82)
38	14 (14–16)	49 (98)	0.98 (0.92)
39	24 (20–26)	46 (92)	0.93 (0.87)
40	40 (38–40)	49 (98)	0.99 (0.93)
41	17 (14–20)	45 (90)	0.89 (0.83)
42	16 (16–16)	50 (100)	1.00 (0.94)
43	36 (30–36)	43 (86)	0.91 (0.85)
44	33 (24–36)	41 (82)	0.86 (0.80)
45	30 (28–34)	44 (88)	0.90 (0.84)
46	28 (22–30)	44 (88)	0.91 (0.85)
47	25 (20–26)	46 (92)	0.93 (0.87)
48	58 (52–58)	46 (92)	0.96 (0.90)

*See Figure 1 for body areas.
†Lower limit of the 99% 1-sided confidence interval for kappa.

raters, all pain drawings in which at least 1 rater had scored area 1 as positive for the presence of pain were selected for rescreening. This process showed that rater 7 had scored area 1 as positive for pain in 13 pain drawings, compared to 2, 3, or 5 positives by the other raters (complete agreement for area 1 was 94% [47 of 50] [$\kappa = 0.85$] when omitting rater 7 from the analysis). This disagreement was related to interpretation error. Rater 7 misinterpreted area 1 by including shading below the line defining area 1 (base of the skull). On further investigation, we observed that the line defining area 1 on one of the transparent templates was faint, and this may have been the template used by rater 7. From the raters' scoring of pain areas, there was complete agreement on the presence or absence of widespread pain in 49 of the 50 pain drawings (98% agreement, $\kappa = 0.98$). Of the pain drawings, 7 were classified as being cases of widespread pain according to the scoring by 7 of the 8 raters; the other rater had one less positive case in comparison.

■ Discussion

Pain drawings used in population surveys are usually completed and scored differently from those used in the clinical setting, although, to our knowledge, nearly all previous studies on the interrater reliability of pain drawings as a tool in musculoskeletal research have focused on the clinical assessment of shading or symbols on such drawings.[13–16] In a large survey, it is often more practical and cost-effective to use several raters to score completed pain drawings; therefore, the quality of the data is dependent on interrater reliability. Our study confirms that the scoring of pain drawings by multiple raters is highly reliable. Good interrater reliability of scoring of pain drawings has also been shown in a select population of patients with chronic pain, although only 2 raters assessed the drawings.[1] Furthermore, we provide evidence for good consistency in the classification of widespread pain, as defined by Macfarlane *et al.*[7]

In our study, complete agreement among all raters in the 50 pain drawings for all body areas, excluding the back of the head (area 1), ranged from 82% to 100%. Area 1 appears to be an "outlier" from the other 49 areas, with a markedly lower agreement among the raters. However, reinspection of the pain drawings in which rater 7 scored area 1 positively for pain suggests that the explanation for this disagreement relates to interpretation error. This result is supported by the fact that the majority of disagreements were for shadings just touching or crossing lines that formed the boundaries of the areas. This finding also highlights the importance of training and guidelines for dealing with ambiguities in scoring pain drawings to help keep disagreements to a minimum. Finally, it is essential to inspect regularly all copies of scoring templates, to check that all line definitions are clear.

■ Conclusions

This study addresses an issue that is important in most studies: how reliable is the scoring method for the tool being used? Specifically, it shows that: (1) several raters, with the type of training and guidelines given in this study, can reliably score pain drawings; and (2) good consistency in the subsequent classification of cases of widespread pain can be obtained from such data. However, we would encourage training and interrater reliability analysis of new teams of multiple raters at the outset of studies using this type of pain drawing and scoring system. Equally, researchers with a single rater can use training and intrarater reliability analysis to ensure reliable data from pain drawings. Overall, our findings provide further evidence that the pain drawing can be a reliable tool for use in self-report health survey research.

■ Key Points

- Multiple raters can reliably score pain drawings.
- High consistency in the subsequent classification of cases of widespread pain can be obtained from pain drawings scored by multiple raters.
- Teams of multiple raters should be trained and assessed for interrater reliability at the outset of studies using pain drawings.

Acknowledgments

The authors thank Ms. Jackie Gray, BA, Ms. Natalie Smith, Ms. Emma Cohen, BSc, Ms. Meg Walmsley-Pitts, Mr. Tom Shave, BA, Mr. Julian Parker, BA, Mr. Ian Watson, and Mr. Martin Neilson. The authors also thank Mr. Andy Lawrence, Keele University Digital Imaging and Illustration Services (KUDIS), for creating the electronic version of the manikin figure.

References

1. Margolis RB, Tait RC, Krause SJ. A rating system for use with patient pain drawings. *Pain* 1986;24:57–65.
2. Jinks C, Lewis M, Ong BN, et al. A brief screening tool for knee pain in primary care. 1. Validity and reliability. *Rheumatology* 2001;40:528–36.
3. Lacey RJ, Lewis M, Sim J. Validity and reliability of a questionnaire for upper quadrant pain and occupational risk factors. *Rheumatology* 2002;41:45.
4. Margolis RB, Chibnall JT, Tait RC. Test-retest reliability of the pain drawing instrument. *Pain* 1988;33:49–51.
5. Roach KE, Brown MD, Dunigan KM, et al. Test-retest reliability of patient reports of low back pain. *J Orthop Sports Phys Ther* 1997;26:253–8.
6. Weiner D, Peterson B, Keefe F. Evaluating persistent pain in long term care residents: What role for pain maps? *Pain* 1998;76:249–57.
7. Macfarlane GJ, Croft PR, Schollum J, et al. Widespread pain: Is an improved classification possible? *J Rheumatol* 1996;23:1628–32.
8. Jinks C, Jordan K, Ong BN, et al. A brief screening tool for knee pain in primary care (KNEST). 2. Results from a survey in the general population aged 50 and over. *Rheumatology* 2004;43:55–61.
9. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
10. Fleiss JL. In: *Statistical Methods for Rates and Proportions*. 2nd ed. New York, NY: John Wiley & Sons; 1981:225–32.
11. Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998;7:301–17.
12. Abramson JH, Gahlinger PM. *Computer Programs for Epidemiologists: PEPI V.4.0*. Salt Lake City, UT: Sagebrush Press; 2001.
13. Chan CW, Goldman S, Ilstrup DM, et al. The pain drawing and Waddell's nonorganic physical signs in chronic low-back pain. *Spine* 1993;18:1717–22.
14. Parker H, Wood PL, Main CJ. The use of the pain drawing as a screening measure to predict psychological distress in chronic low back pain. *Spine* 1995;20:236–43.
15. Reigo T, Tropp H, Timpka T. Pain drawing evaluation—The problem with the clinically biased surgeon: Intra- and interobserver agreement in 50 cases related to clinical bias. *Acta Orthop Scand* 1998;69:408–11.
16. Udén A, Åström M, Bergenudd H. Pain drawings in chronic back pain. *Spine* 1988;13:389–92.