THE RELIABILITY OF THREE ACTIVE MOTOR TESTS USED IN PAINFUL SHOULDER DISORDERS

PRESENTATION OF A METHOD OF GENERAL APPLICABILITY FOR THE ANALYSIS OF RELIABILITY IN THE PRESENCE OF PAIN

Carl-Einar Westerberg, Eva Solens-Bergstroem and Irene Lundh

From the Department of Neurology, University Hospital, Uppsala and Department of Physiotherapy, Central Hospital, Vastera, Sweden

ABSTRACT. This article deals with reliability aspects of standardized, active motor tests (“functional tests”) when applied to patients with painful shoulder disorders. Motor performance was rated independently by the same two examiners in a standardized way in three different manoeuvres: the Hand in Neck, Hand in Back, and Pour out of a Pot tests. Pain experienced during these tests was rated by the patients on a verbal scale. A method of general applicability is presented for the analysis of reliability of standardized, active motor tests when applied to painful shoulder joint disorders. The importance of differential motivation is stressed, as is the importance of using reliability measures that are adapted to the specific purpose of a particular clinical investigation.

Key words: shoulder, pain, impairment, performance scores, functional assessment, motor tests, reliability.

INTRODUCTION

In order to assess the results of treatment for painful shoulder disorders, several active standardized composite movements of the shoulder (“functional movements”) have been developed (4, 5, 7, 9). In a previous study we investigated the influence of pain on motor performance in three motor tests, the Hand in Neck (HIN), the Hand in Back (HIB), and the Pour out of a Pot (POP) manoeuvres (11). The amount of pain, provoked by the movement, is partly determined by the sensitivity of one or more pain generators to mechanical stimuli, and partly by the force acting on the pain-generating(s). Via CNS mechanisms the provoked pain and the anticipation of worsening pain will prematurely interrupt or modify a movement that causes pain (11).

In studies of the reliability of active motor tests it is important to remember that the examiner is also the instructor who verbally describes or physically demonstrates the functional movement that the patient is expected to perform. When giving instructions to one particular patient it is unlikely that two examiners will evoke exactly the same motivation to carry out a movement which may cause pain. It therefore makes sense to say that each instructor has an “encouragement function”, which in all probability differs between different instructors. Differential motivation might lead to a difference between the performance scores recorded by two examiners on the same occasion. Thus, from the psychological point of view it seems meaningful to describe the process of instruction as an interaction between a particular patient and a particular examiner on a particular occasion.

The aim of this study was to investigate the inter-rater reliability of three standardized motor tests and to analyse the causes of disagreement. It has been pointed out by Böckström & Haker (2) and Westerberg et al. (4) that total agreement between raters can hardly be expected in reliability studies in patients with pain. Here we present a new model with general applicability for analysis of interrater reliability in motor performance tests which takes the pain ratings of the patients into consideration. We have identified systematic differences in performance ratings between two physiotherapists which could be explained by differences in the motivation and voluntary effort of the patients during the painful motor tests.

MATERIALS AND METHODS

Two series of patients were examined by two different pairs of fixed examiners. Series 1 consisted of 5 patients with fracture of proximal humerus (FPH), 4 women and 1 man...
with a mean age of 67.8 years (range 56-73) examined 3, 8, 16 and 24 weeks after the injury. On each occasion they performed the HIN, HIB and POP manoeuvres and the performance in each test was assessed independently by two physiotherapists, A and B, who were both experienced in using the three tests and scoring systems (5)). The test order between the two examinations was randomised by the use of opaque envelopes, containing the names of the patients, sealed in the envelopes. The order of the tests was first HIN, then POP and finally HIB. The patients had a rest period of at least one hour between examinations. Each patient rated the pain during each of the three manoeuvres on a modified Borg scale (3).

Series 2 consisted of 15 patients (11 men and 4 women) with a mean age of 52.4 years (range 23-71), with pre-employment diagnosis of subacromial impingement syndrome (SIS). In each patient the motor performance in the HIN test was assessed independently on one single occasion by two physiotherapists, B and C. The test order between the two examinations was randomised and the patients were allowed a rest period of one hour between examinations. B had long experience with the scoring system, whereas C had not used it before.

Data presentation

Each 3 x 3 contingency table in Fig. 1 is based on paired observations of motor performance and pain obtained independently by two physiotherapists, A and B, on 5 x 4 = 20 occasions (series 1) or the number of occasions with differences in motor performance scores between examiners A and B, and the numbers with equal scores, are shown in the rows. A and B, and the middle row is used to express differences between scores as the score of examiner A minus that of examiner B. For each test, the upper rows are marked (+), indicating that the motor performance was rated higher by examiner A than B. The middle row is marked (0), indicating that the motor performance scores were rated equally by both examiners. The bottom rows are marked (-), indicating that examiner A rated the performance lower than B.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked with (+), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle column is marked with (0), indicating equal pain with both examiners. The right-hand columns, marked (-), indicate less pain with examiner A than B.

The columns are constructed analogously with regard to the patient’s ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked with (+), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle column is marked with (0), indicating equal pain with both examiners. The right-hand columns, marked (-), indicate less pain with examiner A than B.

Statistical methods

The contingency table is 3 x 3, with the lowest number of observations being 15. The calculations are based on the assumption that the data follow a binomial distribution with parameters p and (1-p). The significance of the differences between the observed and expected frequencies is tested by using the chi-square test. The significance level is 0.05.

The calculations are based on the assumption that the data follow a binomial distribution with parameters p and (1-p). The significance of the differences between the observed and expected frequencies is tested by using the chi-square test. The significance level is 0.05.

RESULTS

Estimates of agreement

The sum of the observations in each (0)-row, divided by N = 20 and multiplied by 100, gives the per cent agreement value for each test. This is 40% (p < 0.02) for the HIN-test, 85% (p < 0.0001) for the HIB-test and 79% (p < 0.0001) for the POP-test. The p values guard against chance agreement. Thus, it is concluded that there was a good agreement between the performance ratings of examiners A and B in the HIN and POP tests.

Systematic interrater disagreement

A comparison between the (+) and (-) rows gives an estimate of possible systematic differences in performance ratings between examiners A and B. In the HIB test, the relative frequencies were 11 +/− 1/− (p = 0.01), in the HIN test 0 +/− 3/− (p = 0.50) and in the POP-test 6 +/− 0/− (p = 0.04). The p values (two-sided) guard against skew distributions arising.
with a mean age of 67.8 years (range 65–73) examined 3, 8, 16 and 24 weeks after the injury. On each occasion they performed the HIN, HIB, POP and POP measures and the performance in each was assessed independently by two physiotherapists, A and B, who were both experienced in using these tests. The scoring system (9) for the test procedure was similar to the HIN test, the scoring procedure (9) for the test procedure was similar to the HIN test, and the scoring system (9) for the test procedure was similar to the HIN test.

The test order was randomized for each patient, and the patients were allowed a rest period of one hour between examinations. B had no previous experience with the scoring system, whereas C had used it before.

Data presentation
Each 3 x 3 contingency table in Fig. 1 is based on paired observations of motor performance and pain obtained independently by two physiotherapists, A and B, on 5 x 4 = 20 occasions (series 1). The numbers of occasions with differences in motor performance scores between examiners A and B, and the numbers with equal scores, are shown in the rows. A, B, and equal scores are used to express differences between the scores of the examiner A minus that of examiner B. For each test, the upper rows are marked (+), indicating that the motor performance was rated higher by examiner A than by B. The middle row is marked (0), indicating that the performance scores were rated equally by both examiners. The bottom row are marked (-), indicating that examiner A rated the performance lower than B.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

Statistical methods

The three columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.

The columns are constructed analogously with regard to the patients’ ratings of pain during performance of the three motor tests in the presence of examiners A and B, respectively. For each test the left-hand column is marked (±), indicating that the patient reported more pain during motor performance with examiner A than with examiner B. The middle columns are marked with (0), indicating equal pain levels for both examiners. The right-hand columns, marked (−), indicate less pain during motor performance with A as compared with B.

Each contingency table thus contains nine cells. The individual cells are given a combination of signs, i.e., (+), (−) or (0), where the first sign always refers to the A minus B difference in performance scores, and the second sign to the A minus B difference in pain ratings.
by chance. To summarize, examiner A tended to assign higher performance scores than examiner B in the HIN and POP tests, but not in the HIB test.

The comparisons of the numbers of performance ratings that differed between examiners A and B in the (9) columns (equal pain ratings (3+0/0) for HIN, 0+/-2/0 for HIB and 3+0/0 for POP) all gave clearly non-significant results ($p = 0.25$) even in one-sided tests. This meant that examiners A and B did not tend systematically to give different performance scores when the patients rated the provoked pain equally during the two examinations.

Discordant pain ratings

In the HIB test there is a conspicuously skew distribution, 12+7, within the (9) row (equal performance ratings), which suggests that the patients tended to report more pain for a certain performance level in interaction with A than in interaction with B. This skew distribution is unlikely to be a random occurrence ($p = 2 \times 10^{-13}$).

Diagonal comparisons with regard to disagreement in performance and pain ratings

The HIN test data are used for illustration. Under the null hypothesis (see Fig. 2) the sum of the observations in the (+) and (--) cells should equal the sum of the observations in the (++) and (-) cells, but the observed between diagonal distribution was 8.1. Furthermore, the within diagonal distribution (++)/(--) should be equal, but the observed distribution was 7/1. The statistical analysis was performed in two steps. First, the two-sided binomial probability $p_B$ was calculated for the between diagonal distribution 8/1, using $p = 0.5$ ($p_B = 0.0195 = 0.039$). i.e. the discordant observations tended to aggregate within the (+)/(--) diagonal. This $p_B$-value is the same for all possible (+)/(--) distributions. Secondly, the one-sided binomial probability $p_B$ was calculated for the within diagonal distribution 7/1, using $p = 0.5$ ($p_B = 0.0332$). The reason for using a one-sided $p$ value is that a significant overall A minus B difference has already been established in a two-sided test.

To summarize, the systematic A/B difference in performance ratings in the HIN test is largely explained by these paired observations where better performance was achieved at the cost of higher pain.

Fig. 2. Graphic representation of regression of ranks of performance scores on ranks of time for the HIB and HIN tests, using data of examiners A and B, respectively (5 patients with FPH, series 1). A linear model is used for the HIB data, a quadratic regression model for the HIN data.

The same explanation is not valid for the POP test data, another explanation being more likely for this test (see Discussion).

Interexaminer reliability—implications for the description of a healing course

Fig. 3 describes graphically the regression of performance on time, expressed as rank of time, for the HIN and HIB tests, using the data of examiners A and B, respectively (5 patients with FPH, series 1). It has been previously shown (11) that function as measured by the HIB test increases linearly over time, whereas the time-performance relationship is better described by a quadratic regression model for the HIN test.

As seen, the two regression lines describe the healing course as measured by these functional movements much in the same way, i.e. the two examiners made almost identical judgements of changes in performance over time.

DISCUSSION

Statistical considerations

The data in Fig. 1 originate from a study of 5 patients examined on four different occasions. We have regarded the resulting 20 occasions as independent "units of measurement". This means that we regard the data obtained on every occasion as the result of unique circumstances—the sensitivity of the pain provocation subsiding and the interaction between the patient and each of the two raters on a particular occasion.

However, inevitably an element of within patient dependency is introduced into the three $3 \times 3$ tables. This has been taken into account in the analysis. For instance, differences in the "encouragement function", which can be regarded as an advantage. Using data from the same patients several times makes it possible to detect individual peculiarities. Furthermore, as the data were obtained during six months of a healing course after FPH, a wide variety of pain levels was guaranteed, i.e. a great dispersion of performance score/pain score combinations. On the other hand, the conclusions are generalized from only five patients, which increases the risk of drawing erroneous inferences from a sample that is not representative of the population. Of course, the same method of analysis could have been used if we had examined 20 different patients, each on one single occasion, instead of 5 on four occasions. Such a design would have the advantage of complete independence between the individual observations, but the disadvantage of making conclusions regarding individual peculiarities impossible.

A new method for analysis of reliability of active motor tests in painful shoulder disorders

Our method of presenting performance and pain rating data simultaneously in $3 \times 3$ contingency tables allows a quick visual analysis of interexaminer agreement/disagreement and, in addition, offers the possibility of drawing inferences regarding causes of disagreement. To our knowledge this kind of analysis has not been utilized before in reliability analysis of active motor tests. It has been used here for the purpose of analysing three particular motor tests, but we suggest that the method has also applicability to other functional tests, at least in the shoulder.

Agreement in relation to the specific purpose of a clinical study

The agreement between examiners A and B was as high as 70% for the POP test and 85% for the HIB test. For the HIN test the agreement was 40%. This may seem low and thus raise doubts about conclusions based on HIN performance ratings. Would different examiners obtain results leading to different conclusions based on HIN test ratings? Two practical examples are given to elucidate this question.

Example 1: In a previous study (10) the main reason for using the HIN ratings was to describe the recovery of function over time after FPH. In the present study, as seen in Fig. 3, the two examiners described the
by chance. To summarize, examinees A and B tended to assign higher performance scores than examiner C, in the HIN and POP tests, but not in the HIB test.

The comparisons of the numbers of performance ratings that differed between examiners A and B in the 0 (equal performance ratings, 3 +/−/−) and 1 (higher performance ratings for examinee B in the HIN, 0 +/−/− for HIB and 3 +/−/− for POP) all gave clearly non-significant results (p > 0.25) even in one-sided tests. This meant that examiners A and B did not tend systematically to give different performance scores when the patients rated the provoked pain equally during the two examinations.

Discordant pain ratings

In the HIB test there is a conspicuously skew distribution, 12 +/−/−, within the 0 row (equal performance ratings), which suggests that the patients tended to report more pain for a certain performance level in interaction with A than in interaction with B. This skew distribution is not likely to be a random occurrence (p = 2 × 0.0017 = 0.0034).

Diagonal comparisons with regard to disagreement in performance and pain ratings

The HIN test data are used for illustration. Under the null hypothesis (see Fig. 2) the sum of the observations in the cells (+) and (−) should equal the sum of the observations in the (+) and (−) cells, but the observed between diagonal distribution was 8/1. Furthermore, the within diagonal distribution (+) and (−) should be equal, but the observed distribution was 7/1. The statistical analysis was performed in two steps. First, the two-sided binomial probability p1 was calculated for the between diagonal distribution 8/1, using p = 0.5

p1 = 2 × 0.0017 = 0.0034

The discordant observations tend to aggregate within the (+) and (−) diagonal. This p1-value is the same for all possible (+) and (−) diagonal distributions. Secondly, the one-sided binomial probability p2 was calculated for the within diagonal distribution 7/1, using p = 0.5

p2 = 0.0332

The reason for using a one-sided p value is that a significant overall difference between A and B has already been established in a two-sided test.

To summarize, the systematic A/B difference in performance ratings in the HIN test is largely explained by these paired observations where better performance was achieved at the cost of higher pain.

**Intercor reliability—implications for the description of a healing course**

Fig. 3 describes graphically the regression of performance scores on rank of time for the HIN and HIB tests, using the data of examiners A and B, respectively (5 patients with FPH, series 1). A linear model is used for the HIB data, a quadratic model for the HIN test data. The same explanation is not valid for the POP test data, another explanation being more likely for this test (see Discussion).

**Discussion**

The data and Fig. 1 originate from a study of 5 patients examined on four different occasions. We have regarded the resulting 20 occasions as independent "units of measurement". This means that we regard the data obtained on each occasion as the result of unique circumstances—the sensitivity of the pain generators and the interaction between the patient and each of the two raters on a particular occasion. However, it is not possible to test for the clinically important distinctions in HIN and POP performance, the interrater reliability was very good.

**Statistical considerations**

The data and Fig. 1 originate from a study of 5 patients examined on four different occasions. We have regarded the resulting 20 occasions as independent "units of measurement". This means that we regard the data obtained on each occasion as the result of unique circumstances—the sensitivity of the pain generators and the interaction between the patient and each of the two raters on a particular occasion. However, it is not possible to test for the clinically important distinctions in HIN and POP performance, the interrater reliability was very good.

Our method of presenting performance and pain rating data simultaneously in 3 × 3 contingency tables allows a quick visual analysis of interrater agreement/disagreement and, in addition, offers the possibility of drawing inferences regarding causes of disagreement. To our knowledge this kind of analysis has not been utilized before in reliability analyses and has been used here for the purpose of analyzing three particular motor tests, but we suggest that the method also has applicability to other functional tests, at least in the shoulder.

**Agreement in relation to the specific purpose of a clinical study**

The agreement between examiners A and B was as high as 70% for the POP test and 85% for the HIB test. For the HIN test the agreement was 40%. This may seem low and thus raise doubts about conclusions based on HIN performance ratings. Would different examiners obtain results leading to different conclusions based on HIN test ratings? Two practical examples are given to check this question.

**Example 1:** In a previous study (10) the main reason for using the HIN ratings was to describe the recovery of function over time after FPH. In the present study, as seen in Fig. 3, the two examiners described the
healing course in a way that was very similar. This means that examiners A and B made parallel estimates of change of function over time, even though the absolute values of the performance scores were consistently lower for examiner B than for A. Thus, from this aspect the HIN test had good repeatability.

Example 2: As shown in another study (8), a normal HIN test (together with a clearly abnormal POP test) was of positive predictive value for the outcome of surgery in patients with a presumed diagnosis of SIS. Since there was 94% agreement in the estimates of a normal versus an abnormal HIN test, it appears that this test has sufficient reliability in this particular context. This holds true even if the ratings are made by an examiner inexperienced with the use of the rating scale.

Causes of disagreement

The scale factor. The (0) columns contain all occasions when the patients had given identical pain scores during the motor test, i.e. when the A minus B difference was zero. Consequently, for these occasions, differences in performance cannot be attributed to differences in experienced pain. Therefore, the (0) columns offer the best means of analysing whether the two examiners use the same scale in the same way or not. The number of discrepant performance ratings within the (0) column was reasonably small in all three tests and there was no significant evidence of systematically differing interpretations of the scale steps between examiners who were familiar with the standardized scoring system.

The motivation factor; effects of encouragement. A study of the number of observations in the (+++) and (---) cells provides information on a particular aspect of the patient/physiotherapist interaction. The (+++) notation means that the patients attained a higher performance score when examined by A than when examined by B at the expense of a greater pain. The (---) notations means the same thing but with the better performance and higher pain on examination by B.

Given the frequency distribution in the (0) column, the skew distribution, 7/1, observed in the HIN test between the (+++) and (---) cells is significant (p < 0.04). All patients except one contributed to the contents of the (+++) cell. This means that better performance at the cost of higher pain was more common in interaction with A than in interaction with B. The logical interpretation of this is that A more often than B evoked relatively higher motivation in the patients to carry out the movements in spite of the anticipation of worsening pain. Generally speaking, a difference in the "encouragement function" might easily arise if one examiner uses expressions such as "come on, you can do better" and the other does not (this was not the case in the present comparison). However, more subtle differences in the patient/examiner interaction might very well be sufficient.

In the HIB test there was no overall difference in performance ratings (A/B = 0/3, p = 0.5). There was one conspicuous difference in the HIB test, however, namely the I/21 distribution between the (0+) and the (0-) cells (p < 0.02). The most likely explanation for this is the above-mentioned difference in the "encouragement function" between examiners A and B. The patients seem to have used more voluntary effort with examiner A at the expense of greater pain. However, this voluntary effort did not result in higher performance scores. Decreased motivation due to memory of pain. Hypothetically, the observed 7/1 (+++)/(--+) distribution in the HIN test could have arisen for another reason, quite unrelated to an A/B difference in "encouragement function": Assuming that the examination (comprising HIN, HIB and POP tests) performed by A preceded the examination by B on the 7 (+++) occasions and that the reverse was true for the one (---) occasion, the recent memory of pain experienced at the first examination might have had an inhibitory influence on the voluntary effort at the second examination.

However, this alternative explanation should reasonably have had roughly the same consequences for all three tests, i.e. the same individual 8 (7 + 1) paired observations would tend to occupy the same positions in the 3 x 3 tables for the HIB and POP tests also. In fact none of these 8 observations were to be found within the (+++) and (---) cells in the HIB and POP tables. This alternative explanation therefore seems very unlikely.

Decreased performance due to ongoing pain. This situation could occur if all patients were examined in the same order (the examinations by A always preceding those by B) and the time intervals between the examinations were very short. In this case, residual pain in the shoulder caused by the first examination could lead to systematically higher pain ratings during the second examination, thus leading to systematically lower performance with examiner B. This is not likely to have been the case in the present study, since the time interval between examinations with the different examiners was at least one hour. Furthermore, such a mechanism would lead to a skew distribution between the (+++) and (---) cells, with the highest frequency in cell (---), which was not the case.

Fractional factors. In the POP data in Fig. 1 there is a significant overall difference in performance, 6.5 / 7/0 (-/+)(p < 0.04). This difference is probably not explained by any of the mechanisms mentioned above. However, an examination of individual data revealed that as many as three of the six occasions tabulated in the (++) row referred to one single patient. This patient displayed an obvious dislike for rater B for reasons that are known, but are completely irrelevant for this analysis. This very special kind of differential motivation gives a satisfactory explanation for the otherwise not easily understood fact that this patient on one occasion had a score of 4 in the POP performance with rater A but of 1 with rater B in spite of the fact that she declared herself to be free of pain on both examinations.

How robust are the HIN, HIB and POP tests against interrater differences in encouragement?

In the HIN and POP tests, there was no difference between examiners A and B in the (0) column. This fact indicates that different interpretations of the scale steps are not a great problem with these scales. In the HIN test there is a significant difference in the "encouragement function" between raters A and B. In the HIB test, the patients seem to have used more voluntary effort with rater A at the expense of higher pain. In fact, this is again most probably a reflection of the difference in the "encouragement function".

Therefore it seems that the HIB test is robust against differences in "encouragement", whereas the HIN test is not. One reason may be that in the HIB test the scale steps 3-5 are measured in 5 cm intervals, and hence may be insensitive to minor changes in voluntary effort. An alternative explanation is that during the HIN test it is possible for the patient to have visual control over the progression of the movement, that is, the patient can get visual reward for increasing effort. In the HIB test she is completely deprived of this visual reward because the movement takes place outside the visual field.

Both the HIN and HIB tests involve motor achievements within the patient's voluntary control. The POP test, however, does not have the character of a motor achievement and the patients are probably not at all aware of the pain-induced modification of their motor strategy. This test is therefore probably robust against differences in the "encouragement function" which is supported by the non-significant (p = 0.5) (+++)/(--+) distribution in the POP data.

CONCLUSIONS

It is concluded that the HIB and POP tests have good reliability as functional tests in painful shoulder disorders. The HIN test is less robust than the others against motivational factors. The new method presented here seems to be useful in the analysis and interpretation of interrater differences in performance assessments in relation to pain in the shoulder. Overall indices, such as per cent agreement, may not always be the best measures of reliability. Specific measures, adapted to the specific purpose of a particular clinical investigation, should also be used.

ACKNOWLEDGEMENTS

We are grateful to Bengt Renström, Professor of mathematical statistics, for statistical advice. Lenni Ludvigsson, R.T.P., for examining patients in series 2 and Roland Schmidt, M.D., for help in preparing the illustrations.

REFERENCES

healing course in a way that was very similar. This means that examiners A and B made parallel estimates of change of function over time, even though the absolute values of the performance scores were consistently lower for examiner B than for A. Thus, from this aspect the HIN test had good repeatability.

Example 2: As shown in another study (8), a normal HIN test (together with a clearly abnormal POP test) was of positive predictive value for the outcome of surgery in patients with a presumed diagnosis of SIS. Since there was 94% agreement in the estimates of a normal versus an abnormal HIN test, it appears that this test has sufficient reliability in this particular context. This holds true even if the ratings are made by an examiner inexperienced with the use of the rating scale.

Causes of disagreement

The scale factor. The (0) columns contain all occasions when the patients had given identical pain scores during the motor test, i.e. when the A minus B difference was zero. Consequently, for these occasions, differences in performance cannot be attributed to differences in experienced pain. Therefore, the (0) columns offer the best means of analysing whether the two examiners arrived at the same conclusion in the same way or not. The number of discordant performance ratings within the (0) column was reasonably small in all three tests and there was no significant evidence of systematically differing interpretations of the scale steps between examiners who were familiar with the standardized scoring system.

The motivation factor; effects of encouragement. A study of the number of observations in the (++) and (−−) cells provides information on a particular aspect of the patient/physiotherapist interaction. The (+++) notation means that the patients attained a higher performance score when examined by A than when examined by B at the expense of a greater pain. The (−−) notation means the same thing but with the better performance and higher pain on examination by B.

Given the frequency distribution in the (0) column, the skew distribution, 7/1, observed in the HIN test between the (+++) and (−−) cells is significant (p < 0.04). All patients except one contributed to the contents of the (+++) cell. This means that better performance at the cost of higher pain was more common in interaction with A than in interaction with B. The logical interpretation of this is that A more often than B evoked relatively higher motivation in the patients to carry out the movements in spite of the anticipation of worsening pain. Generally speaking, differences in the "encouragement function" might easily arise if one examiner uses expressions such as "come on, you can do better" and the other does not (this was not the case in the present comparison). However, more subtle differences in the patient/examiner interaction might very well be sufficient.

In the HIB test there was no overall difference in performance ratings (A/B = 0.73; p = 0.5). There was only one significant difference in the HIB test, however, namely the 12/1 distribution between the (0+) and the (−−) cells (p < 0.02). The most likely explanation for this is the above-mentioned difference in the "encouragement function" between examiners A and B. The patients seem to have used more voluntary effort with examiner A at the expense of greater pain. However, this voluntary effort did not result in higher performance scores.

Decreased motivation due to memory of pain. Hypothetically, the observed 7/1 (++)/(−−) distribution in the HIN test could have arisen for another reason, quite un-related to an A/B difference in "encouragement function": Assuming that the examination (comprising HIN, HIB and POP tests) performed by A preceded the examination by B on the 7 (+++) occasions and that the reverse was true for the one (−−) occasion, the recent memory of pain experienced at the first examination might have had an inhibitory influence on the voluntary effort at the second examination.

However, this alternative explanation should reasonably have had roughly the same consequences for all three tests, i.e. the same individual 8 (7 ++ + 1) paired observations would tend to occupy the same positions in the 3 x 3 tables for the HIB and POP tests also. In fact none of these 8 observations were to be found within the (++) and (−−) cells in the HIB and POP tables. This alternative explanation therefore seems very unlikely.

Decreased performance due to ongoing pain. This situation could occur if all patients were examined in the same order (the examinations by A always preceding those by B) and the time intervals between the examinations were very short. In this case, residual pain in the shoulder caused by the first examination could lead to systematically higher pain ratings during the second examination, thus leading to systematically lower performance with examiner B. This is not likely to happen in the present study, since the time interval between examinations with the different examiners was at least one hour. Furthermore, such a mechanism would lead to a skew distribution between the (++) and (−−) cells, with the highest frequency in cell (−−), which was not the case.

Statistical factors. In the POP test in Fig. 1, there is a significant overall difference in performance, 6.5/(0) (p < 0.04). This difference is probably not explained by any of the mechanisms mentioned above. However, an examination of individual data revealed that as many as three of the six occasions tabulated in the (+) row referred to one single patient. This patient displayed an obvious dislike for rater B for reasons that are known, but are completely irrelevant for this analysis. This very special kind of differential motivation gives a satisfactory explanation for the otherwise not easily understood fact that this patient on one occasion had a score of 4 in the POP performance with rater A but of 1 with rater B in spite of the fact that she declared herself to be free of pain on both examinations.

How robust are the HIN, HIB and POP tests against interrater differences in encouragement?

In the HIN and POP tests, there was no difference between examiners A and B in the (0) column. This fact indicates that different interpretations of the scale steps are not a great problem with these scales. However, in the HIN test there is a significant difference in the "encouragement function" between rater A and B. In the HIB test, the patients seem to have used more voluntary effort with rater A at the expense of higher pain that fact is again most probably a reflection of the difference in the "encouragement function".

It seems that the HIB test is robust against differences in "encouragement", whereas the HIN test is not. One reason may be that in the HIB test the scales are 3.5 measured in 5 cm intervals, and hence may be insensitive to minor changes in voluntary effort. An alternative explanation is that during the HIN test it is possible for the patient to have visual control over the progression of the movement, that is, she can get visual reward for increasing effort. In the HIB test she is completely deprived of this visual reward because the movement takes place outside the visual field.

Both the HIN and HIB tests involve motor achievements within the patient's voluntary control. The POP test, however, does not have the character of a motor achievement and the patients are probably not at all aware of the pain-induced modification of their motor strategy. This test is therefore probably robust against differences in the "encouragement function" which is supported by the non-significant (p = 0.5) (+++)/(−−) distribution in the POP data.
PAIN IS A MAJOR DETERMINANT OF IMPAIRED PERFORMANCE IN STANDARDIZED ACTIVE MOTOR TESTS

A STUDY IN PATIENTS WITH FRACTURE OF THE PROXIMAL HUMERUS

Eva Solem-Berthoff,¹ ² Irène Lundh¹ and Carl-Einar Westerberg¹

From the Departments of 1Orthopedic Surgery and 2Physiotherapy, Central Hospital, Västerås and 3Department of
Neurosurgery, Uppsala University Hospital, Uppsala, Sweden

ABSTRACT. We have followed the recovery of motor performance and the subsidence of pain for one year in 12 patients after fracture of the proximal humerus. Performance scores during standardized active motor tests were recorded 3, 8, 16, 24, and 52 weeks after injury, and the pain ratings during each of three manoeuvres were assessed on a modified Borg verbal scale. The manoeuvres were: Hand in Neck, Hand in Back, and Pour out of a Pot. In a cross-sectional analysis of data obtained 3 weeks after injury, significant correlations were found between movement-induced pain and impairment of performance in all three tests. A multivariate analysis indicated a strong association between decreasing pain and increasing performance and this was significant after elimination of the influence of healing as measured by time. In contrast, the association between time and increasing performance, after eliminating the influence of decreasing pain, was weak and non-significant. It is concluded that pain is a major determinant of impaired performance after fracture of the proximal humerus, and that performance scores in standardized active motor tests are inversely correlated with the amount of movement-induced pain.

Key words: shoulder, pain, motor tests, performance scores, impairment, functional assessment.

INTRODUCTION

In clinical physiotherapeutic practice and research, standardized and reliable methods for evaluation of motor performance are of great importance. Such methods should be suitable for analysing motor behaviour in relation to pain and pain eliciting mechanisms. They should also be of assistance in planning the treatment strategy besides being suitable tools for evaluation of treatment results.