## Inter-Rater and Intra-Rater Agreement of the Rehabilitation Activities Profile

Jelles, F.; van Bennekom, C.A.M.; Lankhorst, G.J.; Sibbel, C.J.P.; Bouter, L.M.

***published in***
Journal of Clinical Epidemiology
1995

***DOI (link to publisher)***
10.1016/0895-4356(94)00152-G

***document version***
Publisher's PDF, also known as Version of record

Pergamon

0895-4356(94)00152-9

J Clin Epidemiol Vol. 48, No. 3, pp. 407–416, 1995

0895-4356/95 $9.50+0.00

# INTER- AND INTRA-RATER AGREEMENT OF THE REHABILITATION ACTIVITIES PROFILE

FRANK JELLES,* COEN A. M. VAN BENNEKOM,[1]
GUSTAAF J. LANKHORST,[1] CATHARINA J. P. SIBBEL[2] and
LEX M. BOUTER[3]

[1]Free University Hospital, Department of Rehabilitation Medicine, Amsterdam, The Netherlands, [2]Academic Medical Center, Department of Rehabilitation Medicine, Amsterdam, The Netherlands and [3]Vrije Universiteit, Department of Epidemiology and Biostatistics, Amsterdam, The Netherlands

*(Received 13 June 1994)*

**Abstract**—The objective of the study was to determine the inter- and intra-rater agreement of the Rehabilitation Activities Profile (RAP). The RAP is an assessment method that covers the domains of communication, mobility, personal care, occupation and relationships. Each domain consists of items which are further divided in sub-items for in-depth analysis. The RAP allows quantification of the severity of disabilities, handicaps and perceived problems of a patient with regard to the items and sub-items. For this purpose ordinal 4-point Likert scales were constructed. The RAP can be used for goal setting and evaluation of rehabilitation. Because of the broad intended use of the RAP and its construction, a special design for the reliability study was needed. The study was carried out in 5 rehabilitation facilities with the participation of various professions. The items and sub-items of the RAP were divided over these professions according to their expertise. Pairs of interviewers were formed that questioned a patient. For the determination of inter- and intra-rater agreement each pair of interviewers was allowed to question a patient only once. To establish the intra-rater agreement, video recordings were made during the interviews. The median (weighted) kappa value and percentage of agreement about the severity grading of a disability or handicap for all items and sub-items exceeded 0.84 and 81%, respectively, with regard to the inter- and intra-rater agreement. For the severity grading of perceived problems these values were 0.91 and 86%. The interpretation of kappa was hindered by two paradoxes recently described in the literature. The paradox "high agreement but low kappa" manifested itself in particular. It is concluded that inter- and intra-rater agreement of the RAP can be considered to be good to very good.

ICIDH Functional assessment Reliability Kappa Rehabilitation Rehabilitation team

## INTRODUCTION

Rehabilitation medicine is aimed at the prevention and reduction of disabilities and handicaps. Disabilities are defined as "any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being", whereas handicaps are described as "a disadvantage for a given individual, resulting from an impairment or a disability, that limits or prevents the fulfilment of a role that is normal (depending

*All correspondence should be addressed to: F. Jelles, Free University Hospital, Department of Rehabilitation Medicine, PO Box 7057, 1007 MB Amsterdam, The Netherlands.

on age, sex, and social and cultural factors) for that individual" [1]. To accomplish this aim a multidisciplinary team is often needed to cope with the complex and comprehensive consequences of diseases or affections of a patient [2]. Therefore, structured assessment methods are needed for setting rehabilitation goals and evaluating the rehabilitation process [3]. However, these assessment methods should keep a balance between conciseness and comprehensiveness, be aimed at disability and handicap level, and include patient's perception of his or her disability and handicap [3]. Concerning these three requirements existing assessment methods such as the Barthel Index [4], Katz Index [5], the Functional Independence Measure [6] and the Patient Evaluation and Conference System (PECS) [7] are deficient. The first three instruments are suitable for inpatients but do not include occupation, work, mobility outdoors and relationships. The PECS is discipline-oriented instead of disability- and handicap-oriented. Most important of all these instruments do not pay attention to a patient's own assessment about his or her disability and handicap.

Therefore, in 1991 the Rehabilitation Activities Profile (RAP) was constructed at our department [3]. This novel assessment method is intended for specific information gathering relevant to the rehabilitation process and fits easily in daily practice. The RAP has been developed to be used in all types of rehabilitation settings and by all the professions constituting the rehabilitation team. Both outpatients and inpatients with all kinds of diagnosis can be assessed with the RAP. The RAP covers disabilities in the domains communication, mobility, personal care, occupation and handicaps in the domain relationships. Severity of disabilities or handicaps and a patient's own assessment about his or her disabilities and handicaps are graded on ordinal 4-point Likert scales. The RAP is an assessment method specially designed for the rehabilitation team to structure their activities. The RAP is not meant to be a self-administered questionnaire.

Because the RAP is intended to be used for multifarious purposes, a high standard for its methodological properties is required. So, after the construction of the RAP [3], the first step was to study its reliability, which is reported in this article. Reliability refers to the reproducibility or precision of measurements with an instrument [8–13]. The method of estimation of the reliability depends on the use and the nature of the instrument [14]. Scoring of the RAP mainly takes place by means of a semi-structured interview. Because of the central role of the interviewer in the assignment of scores, we concentrated on the estimation of the reliability between and within interviewers [12, 15].

To determine the reliability of the RAP, the following questions will be answered:

1. To what extent do two interviewers agree when they independently score the same patient (inter-rater agreement)?
2. How reproducible are the scores given by one interviewer for the same patient (intra-rater agreement)?

After reliability has been determined the RAP should be validated and tested in the field of rehabilitation. For that purpose two studies are currently being performed. In one study the predictive validity of the RAP is being determined in a group of 120 patients with stroke. The recovery rate is analysed as well as discriminative power and responsiveness of the RAP. For that purpose comparisons will be made with the Barthel Index. In a field experiment the RAP is being used to structure multidisciplinary team conferences in rehabilitation. The effects for the rehabilitation team will be studied.

After completion of these two studies recommendations for use of the RAP in rehabilitation medicine can be formulated.

## METHODS

### *Instrument*

For the RAP* a two-level structure was chosen to accomplish a concise but comprehensive assessment method. Each of the five domains contain a number of items, representing activities and behaviors of daily living. A total of 21 items is identified at the first (global) level. Each item is further divided in sub-items which makes a more detailed assessment possible. For instance, the item "expressing" (domain communication) has 3 sub-items: "nonverbal", "talking" and "writing". In this way all 21 items are divided in a total of 71 sub-items. Table 1 presents the 5 domains and 21 items of the RAP.

Assessment of the RAP is related to common history-taking and is mainly based on verbal information from the patient. Through the answers of a patient two aspects have to be

*A copy of the RAP with its manual is available on request from the first author.

determined by the interviewer (e.g. physician, physical therapist, social worker). The first aspect is the professional judgement of the interviewer about the patient's dysfunction with respect to an item or sub-item. The description of the severity grading in the domains communication, mobility, personal care and occupation reads: 0 = performs activity without difficulty, 1 = performs activity with some difficulty, 2 = performs activity with much difficulty or with some help, 3 = does not perform activity. For the domain relationships a different description was needed. With respect to relationships, the amount of change is determined: 0 = no change, 1 = small change, 2 = large change, 3 = very large change.

The second aspect, which is always assessed simultaneously with disabilities and handicaps, is the perceived problem of a patient. Perceived problems represent the patient's own assessment about his or her disabilities and handicaps. Severity of perceived problems is also graded on an ordinal 4-point Likert scale: 0 = none, 1 = light, 2 = moderate, 3 = severe. This severity scale is similar for all 5 domains.

Disability or handicap and the perceived problem are assessed simultaneously. If required scores "not judgeable" or "not applicable" can also be assigned to an item or sub-item.

*Design*

*Inter-rater agreement*. A patient was questioned once about a specific part of the RAP by two interviewers at the same time. These interviewers had slightly different roles. A distinction was made in a primary and a secondary interviewer. The primary interviewer was leading the history-taking, while the second interviewer was present at that moment and was allowed to ask additional questions. Both interviewers were instructed to formulate their questions as neutrally as possible so that the one would not influence the other. Subsequently, scores were assigned independently by each interviewer. In this way each pair interviewed several patients. By comparing the scores of the interviewers for the same patients, the inter-rater agreement was obtained.

*Intra-rater agreement*. During the interview the patient was recorded by a video-camera with sound-recording. The video recordings were seen after a period of at least 3 weeks by the primary interviewer and on the basis of this information the scores were assigned again. Intra-rater agreement was determined by comparing the scores given at the actual interview with the scores given while watching the video recording.

Table 1. The 5 domains and 21 items of the RAP including the number of sub-items*

| | |
|---|---|
| *Communication* | *Occupation* |
| Expressing (3) | Providing for meals (5) |
| Comprehending (5) | Household activities (3) |
| *Mobility* | Professional activities (4) |
| Maintaining postures (3) | Leisure activities (2) |
| Changing posture (4) | *Relationships* |
| Walking (2) | Partner (4) |
| Using wheelchair (2) | Child(ren) (3) |
| Climbing stairs (2) | Friends/acquaintances (2) |
| Using transport (4) | |
| *Personal care* | |
| Sleeping (2) | |
| Eating and drinking (3) | |
| Washing and grooming (3) | |
| Dressing (5) | |
| Undressing (5) | |
| Maintaining continence (5) | |

*The number of sub-items is given between parentheses for every item.

*Data collection*

Scoring of the 21 items was done by physicians and the 71 sub-items by various members of the rehabilitation team. For the present reliability study, we allocated the sub-items to the various professions, according to their expertise.

Various institutions participated in the reliability study: two university hospitals, a general hospital, an outpatient rehabilitation clinic and a rehabilitation clinic.

In Table 2 the various professions are classified according to institution and the parts of the RAP which were allocated to them. Because we considered some sub-items the concern of social workers and psychologists, we tried to recruit representatives from both professions to score these sub-items.

In every institution and for each profession one pair of interviewers was formed. In the absence of a psychologist the pair consisted of two social workers. Only one social worker and no psychologist participated in university hospital B.

Before the study all interviewers were trained in the use of the RAP by means of assessing 5 patients on the (sub-)items that were allotted to them. This was the only experience all interviewers had with the RAP before the study.

Due to local circumstances each pair of interviewers questioned different numbers of patients.

Table 2. The number of interviewers and their profession classified with regard to institution and part of the RAP allocated to them

| Part of the RAP | Institution | | | | | Total number of interviews |
|---|---|---|---|---|---|---|
| | University hospital A | University hospital B | General hospital | Out-patient rehab. clinic | Rehabilitation clinic | |
| All domains:<br>21 items | 2 Physicians<br>(15 pat.)* | 2 Physicians<br>(4 pat.) | — | 2 Physicians<br>(15 pat.) | 2 Physicians<br>(15 pat.) | 49 |
| Domain communication:<br>sub-items of all items | 2 Speech therapists<br>(20 pat.)† | — | 2 Speech therapists<br>(11 pat.) | — | 2 Speech therapists<br>(20 pat.)† | 51 |
| Domain mobility:<br>sub-items of all items | 2 Physical therapists<br>(15 pat.) | 2 Physical therapists<br>(15 pat.) | — | 2 Physical therapists<br>(15 pat.) | 2 Physical therapists<br>(15 pat.) | 60 |
| Domain personal care:<br>sub-items of the items eating and drinking, washing and grooming, dressing, undressing, maintaining continence;<br>and<br>Domain occupation:<br>sub-items of the items providing for meals, household activities | 2 Occupational therapists<br>(13 pat.) | 2 Occupational therapists<br>(15 pat.) | — | 2 Occupational therapists<br>(15 pat.) | 2 Occupational therapists<br>(15 pat.) | 58 |
| Domain personal care:<br>sub-items of the item sleeping;<br>and<br>Domain occupation:<br>sub-items of the items professional activities, leisure activities;<br>and<br>Domain relationships:<br>sub-items of all items | 2 Social workers<br>(12 pat.) | 1 Social worker<br>(15 pat.) | — | 1 Social worker<br>1 Psychologist<br>(10 pat.) | 1 Social worker<br>1 Psychologist<br>(18 pat.) | 55 |
| Total number of interviews | 75 | 49 | 11 | 55 | 83 | 273 |

*The number of different patients (pat.) questioned by each pair of interviewers is given between parentheses.
†One patient was interviewed in University Hospital A and in the Rehabilitation clinic.

*Subjects*

We took care to include both patients with minor and those with major disabilities in order to cover the whole disability spectrum. The patients were under treatment in one of the participating institutions and were being treated by the profession to which the interviewers belong. Patients were assigned to 1 out of 8 diagnostic groups, commonly used in rehabilitation medicine. The Appendix shows the distribution of patients over diagnostic groups and professions of the interviewers. Patients gave informed consent for their participation in the study.

A total of 250 patients participated in 273 interviews. Nineteen patients were questioned twice and 2 patients took part in 3 interviews. However, no patient was questioned more than once by the same interviewer. Only one patient was interviewed twice by the same discipline, but in different institutions (Table 2).

*Data analysis*

Inter- and intra-rater agreement was calculated for each item and sub-item separately. For this, we aggregated the data of all patients interviewed about a particular item or sub-item over the participating institutions. For the calculation of inter-rater agreement the scores of the primary interviewers were compared with the scores of corresponding secondary interviewers. We considered this way of comparison legitimate despite the slightly different roles of the primary and secondary interviewers. For the intra-rater agreement the scores given during the interviews were compared to the scores which had been assigned during the viewing of the video recordings of the same patients.

The analysis of agreement reflected the order of the decisions which an interviewer takes concerning an item or a sub-item.

*Decision 1:* Should a score "not judgeable" or "not applicable" be given or should a judgement be made about the severity of the disability/handicap and the corresponding perceived problem? When one of the first two scores is given, both the severity of the disability/handicap as well as the severity of the corresponding perceived problem are not scored for that particular (sub-)item. For the calculations a $2 \times 2$ table with nominal categories was constructed (category 1: scoring "not judgeable" or "not applicable"; category 2: scoring a disability and perceived problem).

*Decision 2:* If an item or sub-item can be judged and is also applicable, what is the grading of the disability/handicap on the ordinal 4-point Likert scale?

*Decision 3:* If an item or sub-item can be judged and is also applicable, what is the grading of the perceived problem on the ordinal 4-point Likert scale?

For both decisions 2 and 3 $4 \times 4$ tables were constructed.

Agreement is quantified with Cohen's kappa and percentage of agreement. Kappa was developed as a coefficient of agreement for nominal scales and represents the proportion of agreement corrected for chance agreement [16]. The formula for kappa is: $(p_o - p_e)/(1 - p_e)$ in which $p_o$ is the proportion of observed agreement, while $p_e$ stands for the proportion of agreement expected by chance. The maximum value of kappa is 1 which signals optimal agreement. When kappa equals 0, only chance agreement is present. Negative values of kappa, with a minimum of $-1$ indicate less than chance agreement. We used this kappa formula to calculate the agreement on decision 1.

Weighted kappa was developed for use in studies in which the degree of agreement or disagreement between judges is to be taken into account [17]. As such, weighted kappa is applicable to ordinal scales [18–21], so it is a suitable measure for the calculation of agreement for the ordinal 4-point scales (decisions 2 and 3).

The choice of weights for the degree of agreement is to a certain extent arbitrary and has to be specified [18, 22–24]. We used quadratic agreement weights $w_{ij} = 1 - (i-j)^2/(k-1)^2$ (each cell is defined by row $i$ and column $j$; $k$ is the number of points on a scale) and assigned them to the $k^2$ cells [18, 25]. For a difference between raters of 0, 1, 2 and 3 points this resulted in agreement weights 1, 0.89, 0.56 and 0 respectively.

Use of weighted kappa requires a sufficient number of patients that must be judged per item or sub-item. As determined in a Monte Carlo study, weighted kappa may safely be used if the number of cases is about 30 for a 4-point scale [18–20]. Therefore, we took care to recruit at least 30 patients to be judged per item or sub-item.

## RESULTS

For both inter- and intra-rater agreement the results are given as ranges per domain for ease of survey. The lowest and highest value of (weighted) kappa and percentage of agreement are reported as well as the median (between brackets).

Due to technical failures during the video recordings the number of intra-rater assessments in all tables is somewhat smaller than the number of inter-rater assessments. Therefore, slightly different numbers of patients are found in the various groups.

Detailed results for the items are shown in Tables 3 (decision 1), 4 (decision 2) and 5 (decision 3). Detailed results for the sub-items are omitted because they show the same pattern as the items. A list of values for all items and sub-items can be obtained on request from the first author.

On a number of occasions only a few kappa values or even none at all could be calculated, because one or both interviewers always used the same score [26]. In these instances the results for kappa are accompanied by a note indicating the number of items or sub-items for which kappa could be determined.

## DISCUSSION

In the present study the reliability of the Rehabilitation Activities Profile (RAP) is investigated. The mean duration of all interviews was approx. 15 min. It must be noted, however, that the interview was mostly the first contact between patient and interviewers. This approach ensured that both the primary and secondary interviewer had the same amount of information. It did, however, lengthen the duration of the interview. In daily practice, where the patient is familiar to the interviewer, this period will be (much) shorter.

We considered it inappropriate to question a patient twice to quantify inter- and intra-rater agreement, because answers may change due to learning effects or a patient's decreasing motivation as a consequence of asking the same questions [27]. Thus, with respect to assessment of inter-rater agreement we decided to use a combined interview with independent scorings. This undoubtedly has led to some degree of standardization of the questions. Therefore, inter-rater agreement may decrease marginally when two interviewers use different questions.

Inter- and intra-rater agreement is expressed as values of (quadratic weighted) kappa and the percentage of agreement. The interpretation of (weighted) kappa is not straightforward [21, 28]. Fleiss [25] modified the arbitrary benchmarks given by Landis and Koch [15], for kappa, as well as for weighted kappa. For most purposes values $\leqslant 0.40$ represent poor agreement, values between 0.40 and 0.75 represent fair to good agreement and values $\geqslant 0.75$ indicate excellent agreement. This categorization is frequently used in the literature [29–31]. The artificiality of this classification can be partly overcome by using quadratic weights for weighted kappa. It was demonstrated that such a weighted kappa can be interpreted as an intraclass correlation coefficient (ICC) [21]. Bartko (p. 763) states [32]: "The 1-ICC for intraclass correlation $\geqslant 0$ is interpreted as the percentage of variance due to the disagreement among the raters" [33, 34]. Recently, it was demonstrated that quadratic weighted kappa in some respects should be looked upon as a measure of association rather than agreement [35]. In addition, the (unweighted) kappa for decision 1 can also be interpreted as an ICC, but only when the marginal distributions of the $2 \times 2$ table are equal [23, 36]. This is, of course, always the case when kappa equals 1.

It is known that kappa has some undesirable properties, all depending on the fact that kappa is affected by prevalence [26, 28, 36–39]. If both interviewers use only one score (100% agreement) or one of the interviewers constantly applies the same score, kappa cannot be calculated [26]. Moreover, Feinstein and Cicchetti discuss two paradoxes regarding kappa values obtained from $2 \times 2$ tables [39]. Sometimes a high percentage of agreement can be present while the corresponding value of kappa is very low (paradox 1). In connection with this first paradox kappa can be unexpectedly increased for the same percentage of agreement (paradox 2). Paradox 1 results from symmetrical imbalances in the horizontal and vertical marginal totals. This means that in a 4-fold table the totals of the first column and the first row are relatively high (low), e.g. symmetrical, whereas the totals of the second column and the second row are relatively low (high). The word imbalances point to the difference between the values of the first and second column and the first and second row. Paradox 2 is a consequence of asymmetrical imbalances in the horizontal and vertical marginal totals. These imbalances are not symmetrical: the marginal total of the first

Table 3. Inter- and intra-rater agreement for the identification of a disability or a handicap and perceived problem versus the application of the scores "not judgeable" or "not applicable" with regard to the items (decision 1)

| | | Inter-rater agreement | | | Intra-rater agreement | | |
|---|---|---|---|---|---|---|---|
| Domain | Number of items | Kappa range (median)* | Percentage range (median)* | Number of patients range (median)* | Kappa range (median)* | Percentage range (median)* | Number of patients range (median)* |
| Communication | 2 | 1.00 | 100 | 49 | 0.79 | 98 | 45–46 |
| Mobility | 6 | 0.64–1.00 (1.00)† | 96–100 (100) | 49 | 0.79–0.91 (0.84)† | 96–100 (98) | 46 |
| Personal care | 6 | 1.00‡ | 98–100 (100) | 49 | 0.79–1.00‡ | 98–100 (100) | 46 |
| Occupation | 4 | 0.48–0.95 (0.91) | 96–98 (96) | 49 | 0.79–1.00 (0.91) | 93–100 (98) | 45–46 (46) |
| Relationships | 3 | 0.90–1.00‡ | 96–100 (100) | 49 | 1.00‡ | 100 | 46 |

*The median is presented if more than 2 different values were found.
†Kappa could be determined for 4 items only.
‡Kappa could be determined for 2 items only.

Table 4. Inter- and intra-rater agreement for the grading of a disability on the 4-point scale with regard to the items (decision 2)

| | | Inter-rater agreement | | | Intra-rater agreement | | |
|---|---|---|---|---|---|---|---|
| Domain | Number of items | Weighted kappa range (median)* | Percentage range (median)* | Number of patients range (median)* | Weighted kappa range (median)* | Percentage range (median)* | Number of patients range (median)* |
| Communication | 2 | 0.54–0.93 | 94 | 47 | 0.94–1.00 | 95–100 | 42–43 |
| Mobility | 6 | 0.84–0.96 (0.88) | 80–91 (85) | 22–49 (46) | 0.72–0.94 (0.86) | 70–88 (83) | 21–46 (43) |
| Personal care | 6 | 0.77–0.94 (0.87) | 73–92 (83) | 44–49 (49) | 0.81–0.91 (0.87) | 78–91 (83) | 43–46 (46) |
| Occupation | 4 | 0.86–0.98 (0.94) | 75–93 (87) | 29–46 (33) | 0.78–0.93 (0.89) | 70–92 (77) | 25–42 (30) |
| Relationships | 3 | 0.80–0.88 (0.81) | 73–86 (84) | 33–49 (35) | 0.65–1.00 (0.91) | 74–100 (84) | 31–46 (32) |

*The median is presented if more than 2 different values were found.

Table 5. Inter- and intra-rater agreement for the grading of a perceived problem on the 4-point scale with regard to the items (decision 3)

| | | Inter-rater agreement | | | Intra-rater agreement | | |
|---|---|---|---|---|---|---|---|
| Domain | Number of items | Weighted kappa range (median)* | Percentage range (median)* | Number of patients range (median)* | Weighted kappa range (median)* | Percentage range (median)* | Number of patients range (median)* |
| Communication | 2 | 0.56–0.94 | 94–96 | 47 | 0.94–1.00 | 95–100 | 42–43 |
| Mobility | 6 | 0.83–0.94 (0.88) | 71–86 (83) | 22–49 (46) | 0.75–0.94 (0.82) | 78–88 (84) | 21–46 (43) |
| Personal care | 6 | 0.82–1.00 (0.90) | 85–100 (91) | 44–49 (49) | 0.81–0.98 (0.93) | 91–98 (95) | 43–46 (46) |
| Occupation | 4 | 0.88–0.97 (0.93) | 75–98 (86) | 29–46 (33) | 0.88–0.97 (0.93) | 76–90 (87) | 25–42 (30) |
| Relationships | 3 | 0.91–0.92 (0.91) | 85–90 (89) | 33–49 (35) | 0.83–1.00 (0.95) | 90–100 (91) | 31–46 (32) |

*The median is presented if more than 2 different values were found.

column is relatively high (low) while the marginal total of the first row is relatively low (high).

Especially the paradox "high agreement but low kappa" appeared a number of times in the present study. An extreme example for the intra-rater agreement on decision 1 was kappa for sub-item "activities" from the item "friends/acquaintances". Kappa was −0.03 indicating less than chance agreement although a corresponding agreement of 91% was established. Moreover, this paradox also manifested itself for decisions 2 and 3. The agreement on these two decisions was calculated in $4 \times 4$ tables. The inter-rater agreement for the sub-item "nonverbal" of the item "comprehending" had weighted kappa values of 0.18 (decision 2) and 0.24 (decision 3). However, the corresponding percentages of agreement were 89% for decision 2 and 93% for decision 3. To resolve these paradoxes additional measures (of agreement) were introduced, however, only for $2 \times 2$ tables [40–42]. Therefore, the combination (weighted) kappa and percentage of agreement at present are the best available parameters.

In the literature it has been argued that more emphasis should be laid on the raw data [28]. This would result in the representation of all the tables with which (weighted) kappa is calculated. This approach, however, is only suitable when few kappa values have to be determined. In the present study this way of representation would be highly impractical. A total of 184 (21 items + 71 sub-items for inter- and intra-rater agreement; decision 1) $2 \times 2$ tables and 368 (decisions 2 and 3) $4 \times 4$ tables would have to be depicted.

The agreement on decision 1 was usually somewhat higher than for decisions 2 and 3. This is probably the result of the smaller number of scoring categories for decision 1 (two scores vs four scores for decisions 2 and 3). In general kappa shows a tendency for higher values when the number of scoring categories is small [26, 28].

What is important to the external validity of our study is the variety of participating rehabilitation settings, interviewers and patients (inpatients and outpatients, different diagnoses). This allows the results to be generalized to other rehabilitation institutions, interviewers and patients.

Although unexpected, the intra-rater agreement tended to be smaller than the inter-rater agreement [27]. There are three possible explanations for this. The first regards the use of video. Having interviewers watch the video recordings introduced an additional source of variation. Some interviewers reported feeling the need to obtain more information than recorded, although the amount of information was the same as during the actual interviews. Others commented that watching the video recordings was rather tiring and they were easily distracted. Secondly, due to learning effects as a result of the actual interviews of the primary interviewers different scores could have been applied when assessing the video recordings. Before the reliability study started all interviewers had been trained by interviewing 5 patients. Maybe this number was too small. A third explanation might be one of the paradoxes mentioned (high agreement but low kappa) [28, 36, 39].

The number of patients that had to be judged per item or sub-item should at least have been 30 [18–20]. In relation to decisions 2 and 3, a total of 2 items and 14 sub-items did not reach the 30 number limit for the determination of the inter-rater agreement, whereas for the intra-rater agreement 2 items and 6 sub-items did not reach this limit. The agreement results for these items and sub-items consequently have to be interpreted with some caution.

With regard to the value of (weighted) kappa and the matching percentage of agreement for every item and sub-item, the results indicate that a very satisfactory level of inter- and intra-rater agreement for the RAP can be obtained for all three decisions. It is important to note that none of the items or sub-items had low values for both (weighted) kappa and the percentage of agreement. Because of this there is no need to reconsider certain items and sub-items in the RAP for reasons of poor reliability.

*Acknowledgements*—We thank the participating interviewers for their contribution to the study. In alphabetical order: J. E. Albrecht, I. C. Alderse Baes, J. Blees-Pelk, M. J. Boomars, A. H. M. Brink-Oppenkamp, E. Broere, M. Cardol, M. J. M. Cheriex, R. Daniëls, A. M. J. J. van Dooren, A. L. H. Eysackers, P. C. P. M. Elst, N. van den Eng, M. A. H. Engels, N. Groenendaal, W. J. van der Heide, B. A. de Jong, J. M. Jongejan, J. E. van der Kaaij, J. J. ten Kate, H. J. J. Kooijmans, J. Koomans, P. A. Koppe, A. E. Kruisselbrink, G. J. F. Kuipers, M. E. J. Lumens, L. A. A. Migchelsen, G. Neuhuys-Oltheten, P. J. Noordijk, C. E. M. Savonije, L. R. M. Spitz, A. M. V. Stoopendaal, H. I. M. Tonneijck, H. O. Wiggerts, M. Wolters-Vink, E. J. Wouda, R. C. J. Zondervan. This publication is part of a research project, which is supported by a grant from the Dutch Ministry of Welfare, Public Health and Cultural Affairs.

## REFERENCES

1. World Health Organization (WHO). **International Classification of Impairments, Disabilities, and Handicaps: a Manual of Classification Relating to the Consequences of Disease.** Geneva: WHO; 1980.

2. Halstead LS, Rintala DH, Kanellos M, Griffin B, Higgins L, Rheinecker S, Whiteside W, Healy JE. The innovative rehabilitation team: an experiment in teambuilding. **Arch Phys Med Rehabil** 1986; 67: 357–361.
3. Bennekom CAM van, Jelles F, Lankhorst GJ. Rehabilitation Activities Profile: the ICIDH as a framework for a problem-oriented assessment method in rehabilitation medicine. **Disabil Rehabil** In press.
4. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. **Md State Med J** 1965; 14: 61–65.
5. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardised measure of biological and psychosocial function. **JAMA** 1963; 185: 914–919.
6. The Center for Functional Assessment Research and the Uniform Data System for Medical Rehabilitation. **Guide for Use of the Uniform Data Set for Medical Rehabilitation. Version 3.1.** Buffalo: Research Foundation of the State University of New York; 1991.
7. Harvey RF, Jellinek HM. Functional performance assessment: a program approach. **Arch Phys Med Rehabil** 1981; 62: 456–461.
8. Currier DP. **Elements of Research in Physical Therapy.** Baltimore: Williams & Wilkins; 1984.
9. Fletcher RH, Fletcher SW, Wagner EH. **Clinical Epidemiology: the Essentials.** Baltimore: Williams & Wilkins; 1988.
10. Keith RA. Functional assessment measures in medical rehabilitation: current status. **Arch Phys Med Rehabil** 1984; 65: 74–78.
11. Payton OD. **Research: the Validation of Clinical Practice.** Philadelphia: Davis; 1988.
12. Sheikh K. Disability scales: assessment of reliability. **Arch Phys Med Rehabil** 1986; 67: 245–249.
13. Tugwell P, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. **J Rheumatol** 1982; 9: 758–762.
14. Liang MH, Jette AM. Measuring functional ability in chronic arthritis. **Arthritis Rheum** 1981; 24: 80–86.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. **Biometrics** 1977; 33: 159–174.
16. Cohen J. A coefficient of agreement for nominal scales. **Educ Psychol Meas** 1960; 20: 37–46.
17. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychol Bull** 1968; 70: 213–220.
18. Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. **Br J Psychiat** 1976; 129: 452–456.
19. Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. **Appl Psychol Meas** 1981; 5: 101–104.
20. Cicchetti DV, Fleiss JL. Comparison of the null distribution of weighted kappa and the C ordinal statistic. **Appl Psychol Meas** 1977; 1: 195–201.
21. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. **Educ Psychol Meas** 1973; 33: 613–619.
22. Tinsley HEA, Weiss DJ. Interrater reliability and agreement of subjective judgments. **J Counsel Psychol** 1975; 22: 358–376.
23. Bartko JJ, Carpenter WT. On the methods and theory of reliability. **J Nerv Ment Dis** 1976; 163: 307–317.
24. Soeken KL, Prescott PA. Issues in the use of kappa to estimate reliability. **Med Care** 1986; 24: 733–741.
25. Fleiss JL. **Statistical Methods for Rates and Proportions.** New York: Wiley; 1981.
26. Schouten HJA. **Statistical Measurement of Inter-observer Agreement: Analysis of Agreements and Disagreements Between Observers.** Ph.D. Thesis. Utrecht: Elinkwijk; 1985.
27. Department of Clinical Epidemiology and Biostatistics, McMaster University. Clinical disagreement I. How often it occurs and why. **Can Med Assoc J** 1980; 123: 499–504.
28. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. **Br Med J** 1992; 304: 1491–1494.
29. Gisbergen JWMM van, Dekker J. Reliability of the diagnosis of impairments and disabilities by exercise therapists. **J Rehabil Sciences** 1992; 5: 67–73.
30. Haley SM, Osberg JS. Kappa coefficient calculation using multiple ratings per subject: a special communication. **Phys Ther** 1989; 69: 970–974.
31. Triet EF van, Dekker J, Kerssens JJ, Curfs EChr. Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy. **Int Disabil Stud** 1990; 12: 61–65.
32. Bartko JJ. On various intraclass correlation reliability coefficients. **Psychol Bull** 1976; 83: 762–765.
33. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. **Psychol Rep** 1966; 19: 3–11.
34. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. **Psychol Bull** 1979; 86: 420–428.
35. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. **J Clin Epidemiol** 1993; 46: 1055–1062.
36. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. **J Clin Epidemiol** 1988; 41: 959–968.
37. Kraemer HC. Ramifications of a population model for $\kappa$ as a coefficient of reliability. **Psychometrika** 1979; 44: 461–472.
38. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. **J Clin Epidemiol** 1988; 41: 949–958.
39. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. **J Clin Epidemiol** 1990; 43: 543–549.
40. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. **J Clin Epidemiol** 1990; 43: 551–558.
41. Nice DA, Riddle DL, Lamb RL, Mayhew TP, Rucker K. Intertester reliability of judgments of the presence of trigger points in patients with low back pain. **Arch Phys Med Rehabil** 1992; 73: 893–898.
42. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. **J Clin Epidemiol** 1993; 46: 423–429.


*(Appendix overleaf)*

## APPENDIX

The number of interviewed patients classified into diagnostic category, gender and the professions by whom they were interviewed

| | | Profession | | | | | |
|---|---|---|---|---|---|---|---|
| Diagnostic category | Gender* | Physician | Speech therapist | Physical therapist | Occupational therapist | Social worker/ Psychologist | Total† |
| Rheumatoid arthritis and other rheumatic diseases | F | — (1)‡ | — (—) | — (12) | — (6) | — (5) | 24 |
| | M | — (2) | — (—) | — (1) | — (—) | — (1) | 4 |
| Osteoarthrosis, post-traumatic conditions and other orthopedic conditions | F | — (4) | — (—) | — (3) | 1 (4) | — (3) | 15 |
| | M | — (—) | — (—) | 1 (2) | — (1) | — (4) | 8 |
| Amputation | F | — (1) | — (—) | 2 (—) | 2 (—) | — (2) | 7 |
| | M | 3 (2) | — (—) | — (3) | 1 (—) | 2 (1) | 12 |
| Low back pain, neck and shoulder pain (soft tissue rheumatism) | F | — (6) | — (—) | — (2) | — (2) | — (1) | 11 |
| | M | — (2) | — (—) | — (—) | — (—) | — (—) | 2 |
| Stroke and other diseases of the central nervous system (excl. spinal cord injury) | F | 2 (5) | 24 (4)§ | 3 (10) | 4 (9) | 1 (10) | 72 |
| | M | 5 (5) | 21 (2) | 4 (7) | 1 (16) | 2 (9) | 72 |
| Diseases of the peripheral nervous system | F | — (2) | — (—) | — (1) | — (2) | — (3) | 8 |
| | M | — (2) | — (—) | — (—) | — (2) | — (2) | 6 |
| Spinal cord injury | F | 1 (1) | — (—) | 1 (—) | 2 (—) | — (—) | 5 |
| | M | 4 (—) | — (—) | 3 (—) | 3 (—) | 4 (—) | 14 |
| Other/unknown | F | — (1) | — (—) | — (1) | 1 (1) | — (1) | 5 |
| | M | — (—) | — (—) | 1 (3) | — (—) | — (4) | 8 |
| | Total† | 15 (34) | 45 (6) | 15 (45) | 15 (43) | 9 (46) | 273 |

*F, female and M, male.

†The given frequency of patients represents the cumulated number over all institutions. A total of 273 interviews were carried out.

‡A distinction is made between inpatients and outpatients (between parentheses).

§One female patient was questioned twice by speech therapists, however in different institutions.