

# Rasch analysis of the Dutch version of the Oxford elbow score

Jeroen de Haan<sup>1</sup>  
Niels Schep<sup>2</sup>  
Wim Tuinebreijer<sup>2</sup>  
Peter Patka<sup>2</sup>  
Dennis den Hartog<sup>2</sup>

<sup>1</sup>Department of Surgery and Traumatology, Westfriesgasthuis, Hoorn, the Netherlands;

<sup>2</sup>Department of Surgery and Traumatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

**Background:** The Oxford elbow score (OES) is a patient-rated, 12-item questionnaire that measures quality of life in relation to elbow disorders. This English questionnaire has been proven to be a reliable and valid instrument. Recently, the OES has been translated into Dutch and examined for its reliability, validity, and responsiveness in a group of Dutch patients with elbow pathology. The aim of this study was to analyze the Dutch version of the OES (OES-DV) in combination with Rasch analysis or the one-parameter item response theory to examine the structure of the questionnaire.

**Methods:** The OES-DV was administered to 103 patients (68 female, 35 male). The mean age of the patients was  $44.3 \pm 14.7$  (range 15–75) years. Rasch analysis was performed using the Winsteps® Rasch Measurement Version 3.70.1.1 and a rating scale parameterization.

**Results:** The person separation index, which is a measure of person reliability, was excellent (2.30). All the items of the OES had a reasonable mean square infit or outfit value between 0.6 and 1.7. The threshold of items were ordered, so the categories can function as intended. Principal component analysis of the residuals partly confirmed the multidimensionality of the English version of the OES. The OES distinguished 3.4 strata, which indicates that about three ranges can be differentiated.

**Conclusion:** Rasch analysis of the OES-DV showed that the data fit to the stringent Rasch model. The multidimensionality of the English version of the OES was partly confirmed, and the four items of the function and three items of the pain domain were recognized as separate domains. The category rating scale of the OES-DV works well. The OES can distinguish 3.4 strata. This conclusion can only be applied to elbow dislocations, which were the largest group of patients studied.

**Keywords:** elbow, questionnaire, quality of life, traumatology, modern test theory, reliability, outcomes

## Introduction

The Oxford elbow score (OES) is a patient-rated 12-item questionnaire (Table 1) that measures quality of life in relation to elbow disorders.<sup>1</sup> The development of this English questionnaire has involved Rasch analysis as well as analysis with classical test theory.<sup>1</sup> This English version comprises three domains, ie, elbow function, pain, and social-psychological factors. The answers are recorded on a five-point Likert scale. Every domain score is calculated for a final score that ranges from 0 (worst) to 100 (best). Another study reported responsiveness and minimal change for the OES following elbow surgery.<sup>2</sup> These studies showed that the OES questionnaire is a reliable and valid instrument. Recently, the 12 items of the OES were translated into Dutch

Correspondence: Dennis den Hartog  
Department of Surgery and Traumatology, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, the Netherlands  
Tel +31 10 7032395  
Fax +31 10 7032396  
Email d.denhartog@erasmusmc.nl

**Table 1** The 12 items of the Oxford elbow score**“During the past 4 weeks ...”**

1. Have you had difficulty lifting things in your home, such as putting out the rubbish, because of your elbow problem?
2. Have you had difficulty carrying bags of shopping, because of your elbow problem?
3. Have you had any difficulty washing yourself all over, because of your elbow problem?
4. Have you had any difficulty dressing yourself, because of your elbow problem?
5. Have you felt that your elbow problem is “controlling your life”?
6. How much has your elbow problem been “on your mind”?
7. Have you been troubled by pain from your elbow in bed at night?
8. How often has your elbow pain interfered with your sleeping?
9. How much has your elbow problem interfered with your usual work or everyday activities?
10. Has your elbow problem limited your ability to take part in leisure activities that you enjoy doing?
11. How would you describe the worst pain you had from your elbow?
12. How would you describe the pain you usually had from your elbow?

according to the generally accepted rules for translation of non-Dutch questionnaires and examined for their reliability, validity, and responsiveness among a group of Dutch patients who had experienced elbow trauma.<sup>3-5</sup> However, that study was performed with classical test theory (unpublished data). Modern test theory has many advantages, such as a thorough examination of dimensionality, analysis of the fit of the data to the Rasch model, and category function analysis.<sup>6,7</sup> Therefore, the objective of this study was to analyze the Dutch version of the OES (OES-DV) with Rasch analysis or the one-parameter item response theory.

## Methods

The OES was translated into Dutch by four trauma surgeons. One clinician was also an epidemiologist with experience in clinimetrics. The four translated versions were compared, and the differences were resolved by discussion. The Dutch version of the OES was then back-translated to English by a certified English translator (and native English speaker). The four clinicians compared this back-translation with the original English version of the OES, and edited the Dutch translation to make it more accurate. After the translation process, mistakes were encountered in the tense of the Dutch version of questions seven and eight, which referred to pain during the past 4 weeks. These mistakes were found after the back-translation and were corrected.

The sample population consisted of 103 patients (68 female, 35 male). The mean age of the patients was  $44.3 \pm 14.7$  (range 15–75) years. This group of patients consisted of 67 patients with elbow luxation, 24 patients with a recent fracture of the

elbow region, seven patients with active epicondylitis, two patients who were undergoing arthrolysis of the elbow, and three patients with other elbow conditions (eg, bursitis).

Forty-three patients were randomly selected to complete a second OES following treatment for their elbow disorder after a mean follow-up of  $52 \pm 24.1$  days; thus, 146 questionnaires were available for analysis. This group of 43 patients consisted of 19 patients with elbow luxation, 14 patients with a recent fracture of the elbow region, five patients with active epicondylitis, two patients who were undergoing arthrolysis of the elbow, and three patients with other elbow conditions (eg, bursitis).

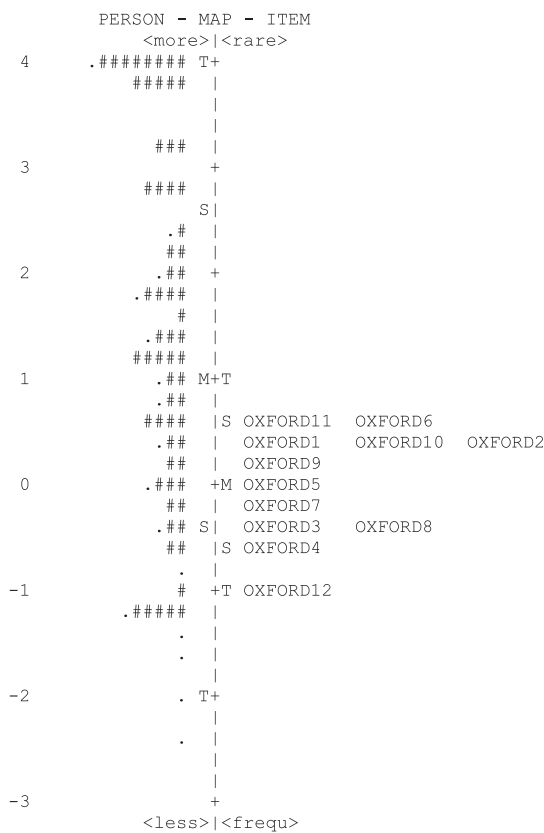
Rasch analysis was performed with Winsteps measurement software (Winsteps® Rasch Measurement Version 3.70.1.1). The following analyses were performed: construction of the person and item or Wright map, testing of the (mis)fit between the data and the model, estimation of the person and item reliability and separation coefficient, testing of the ordering of the categories, and analysis of the dimensionality.

## Results

Of the 146 observations collected, all were available for analysis, including 17 extreme scores. A rating scale parameterization was used because all the items had the same numbers of categories. The person and item map is shown in Figure 1. The items on the right side are in order based on the logit scale. Natural logarithms of the odds are called logits and range from minus infinity to plus infinity. The default mean difficulty was set to zero. The OES covered about seven logits (range 5.08 to  $-1.60$ ). The OES was coded with the highest values for the better patients and the lowest values for the bad cases. The item map also shows the hierarchy of the item difficulties on the right side. The items at the top are those items the patients easily endorsed. For example, item 1 (difficulty in lifting objects) was easier to confirm than item 12 (describing the pain usually experienced).

On the left of the Wright map the patient performances are represented by numbers. Most patients were located opposite and above the items (mean person estimate  $1.04 \pm 1.52$ ), which was above the default mean item difficulty of zero; therefore, the items were moderately well targeted for this person group, meaning that the items represent the patients' level of quality of life. Overall, the patient group had good OES scores, indicating few complaints concerning the affected elbow.

The item reliability coefficient was 0.95, and the item separation coefficient was 4.40. The person reliability was 0.84,



**Figure 1** Person (n = 146) and item (12 items) or Wright map for the Oxford elbow score scale. Positive scores indicate better quality of life, whereas negative scores indicate poorer quality of life. Items from the scale are shown on the right side of the figure, and person measures are highlighted by “#” or “.”  
**Notes:** Each “#” represents two subjects, and each “.” represents one subject.  
**Abbreviations:** M, mean; S, one standard deviation from the mean; T, two standard deviations from the mean.

and the person separation index was 2.30, so the reliability was good. Reliability coefficients have a ceiling of 1.0, but separation coefficients and indexes have no ceiling. The person separation was used to calculate the number of distinct levels of quality of life (strata) that the items could distinguish ( $Strata = [4 \times \text{person separation index} + 1]/3 = 3.4$ ).<sup>8,9</sup> The strata that the OES distinguished was 3.4, which indicates that about three ranges could be differentiated.

The items are placed according to the hierarchy of the item difficulties in Table 2. The measures are the item difficulty estimates. Items 2, 6, 11, items 1, 2, 10, items 1, 8, 9, and items 3 and 8 had inter-item separations less than 0.15 logits, indicating overlap between these item difficulties.

The individual item fit statistics are presented in Table 2. To determine how well the empirical data fit the Rasch model, Chi-square fit statistics were calculated. These fit statistics are the infit mean square (Infit MNSQ) and the outfit mean square (Outfit MNSQ). The infit MNSQ represents the information-weighted mean square residual difference between observed

**Table 2** Item statistics, Oxford elbow score, Dutch version

Item	Count	Measure	Infit MNSQ	Outfit MNSQ
6	146	0.57	1.02	1.26
11	146	0.57	0.68	0.65
2	146	0.47	1.08	0.96
10	146	0.41	1.19	1.05
1	146	0.35	0.99	0.91
9	146	0.28	0.69	0.61
5	146	-0.01	0.89	0.89
7	146	-0.17	1.19	0.96
8	146	-0.36	1.11	0.99
3	146	-0.44	1.03	0.97
4	146	-0.66	0.81	0.85
12	144	-1.02	1.70	1.53

and expected responses. The infit statistics are sensitive to unexpected responses near the person’s ability level. The outfit statistic is the usual unweighted mean square residual and is more sensitive to outliers. High infit and outfit reflect underfit, which means lack of predictability of an item. Low mean square infit and outfit reflect overfit, which means overpredictability of an item. Mean square infit or outfit values should range between 0.6 and 1.4 for rating scales or 0.5 to 1.7 for clinical observations.<sup>7</sup> All the items in the OES had reasonable mean square infit or outfit values between 0.6 and 1.7. Item 12 had the highest infit and outfit values of all the items, 1.70 and 1.53, respectively, indicating some underfit, which means an unpredictable interpretation (ie, erratic response or noise).

Table 3 presents the functioning of the five categories of the OES Dutch version. All categories were well represented except for the zero category, which had a low frequency of 109 observations. The zero category included the patients with the worst quality of life; therefore, the low frequency is consistent with the few cases with a bad or very bad outcome. The observed average measures increased in a smooth distribution from -0.99 to 2.39. The threshold of the categories increased monotonically (so were never decreasing). None of the categories showed a misfit. Figure 2 shows the category probability curves of the categories with a smooth distribution. The thresholds were ordered. In this Rasch-Andrich model (one of the polytomous models), the rating scale structure was defined to be equal for all items. The category rating scale worked well.

According to Rasch theory, when the data fit the Rasch model, the Rasch dimension is the only dimension in the data. Rasch factor analysis is a factor analysis of the residuals that remain after the linear Rasch measure has been extracted from the data set. A secondary dimension in the data must explain at least two items (also called 2.0 Eigenvalue units) worth of

**Table 3** Summary of the category structure of the Oxford elbow score, Dutch version

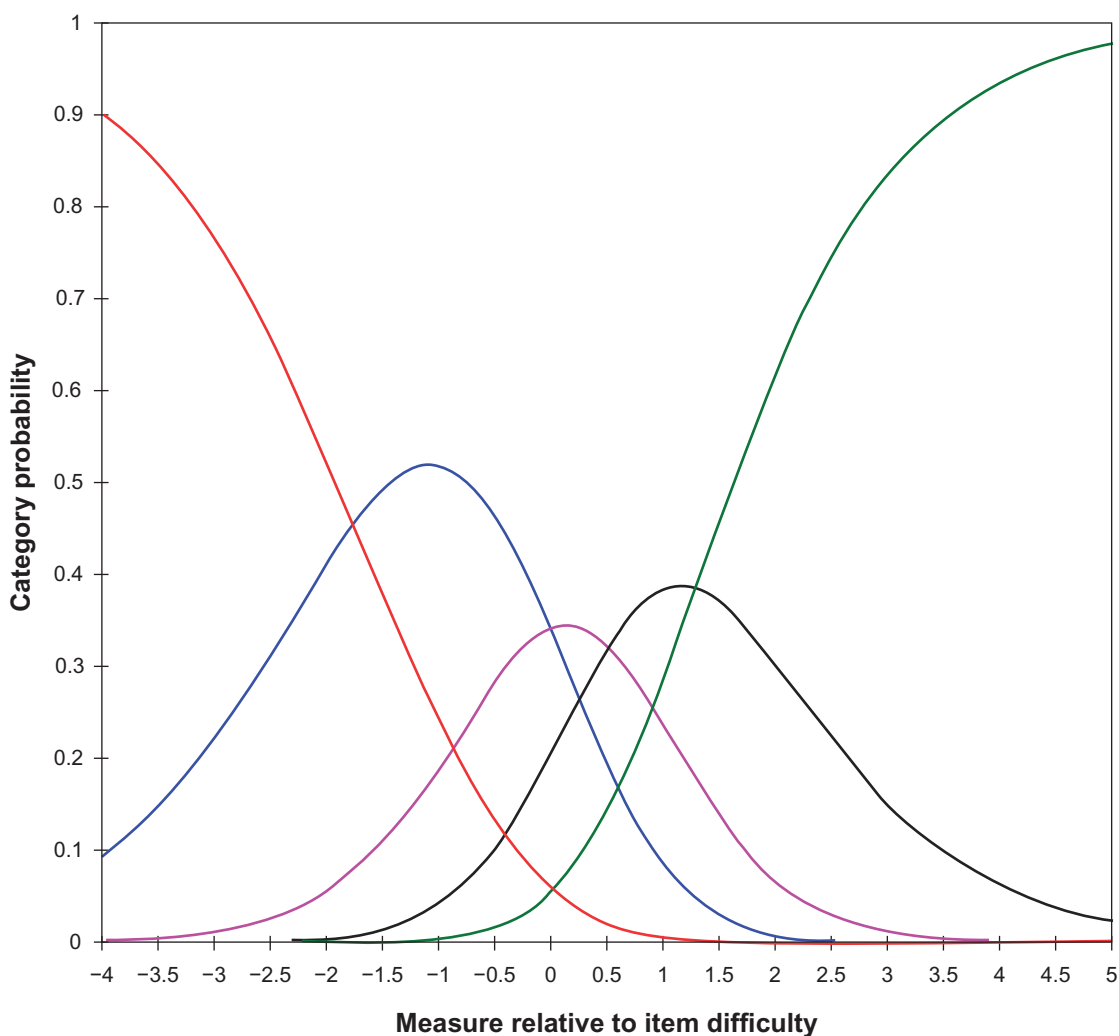
Category label/score	Observed count	Observed count %	Observed average	Outfit mean square	Threshold
0	109	6	-0.99	1.14	none
1	276	16	-0.63	0.95	-1.77
2	258	15	0.38	0.67	-0.02
3	335	19	1.31	0.89	0.51
4	772	44	2.39	1.22	1.28

variance, ie, unless a component has the strength of at least two items it may merely be due to an idiosyncratic item. A Rasch principal component analysis (PCA) of the residuals of the OES-DV was performed. The raw variance of the OES explained by the Rasch measure was 36.2% (expected by the model 35.7%). The unexplained variance in the first contrast was 7.4% (2.4 Eigenvalue units), and in the second contrast was 6.0% (2.0 Eigenvalue units). The first contrast consisted

of three of the four pain items. The second contrast consisted of the function items 1 to 4.

## Discussion

Rasch analysis of the OES-DV showed that the data fit to the stringent Rasch model. The person separation index as a measure of reliability was high. Three statistically distinct levels of quality of life, ie, good, intermediate, and poor, could be differentiated by the



**Figure 2** Category probability curve of the Dutch version of the Oxford elbow score scale showing the probability of assigning to any particular category (y axis) given the difference in estimates between any patient quality of life measurement and any item difficulty. The threshold estimates correspond to the intersection of rating scale categories.

OES-DV. The category rating scale of the OES-DV worked well. The patients could discriminate the five levels of the items.

Our factor analysis of the OES-DV with the classical test theory showed only one factor, which is in contrast with the original English version (unpublished data). Factor analysis of the English OES showed three domains, ie, function, pain, and socio-psychological.<sup>1</sup> In the Rasch analysis, the PCA was performed on the differences between the model and observed data, called residuals. The multidimensionality was partly confirmed by the PCA of the residuals of the OES-DV. The four items of the function and three items of the pain domain were recognized as separate domains. Two contrasts had a strength of two or more Eigenvalue units. This supports the idea that the OES-DV is a multidimensional instrument. The difference in dimensionality between the OES-DV and the original OES can be explained by differences in the composition of the study population, context, intervention, and timing of assessments. Our study population, which consisted mainly of elbow dislocations, was very different from the original developmental study.

This study had several limitations, including a small sample size and a homogeneous patient population (ie, patients with elbow trauma). Our conclusions can only be applied to elbow dislocations, which were the largest group of patients studied. A flaw in our study is that we did not ask the patients to predict the hierarchy of the Wright map and use these predictions to study the predictive validity of the OES-DV. We could have asked the patients to order the items according to how difficult they were to perform and endorse them positively or negatively. If the order of difficulty of the items was correctly predicted by the patients, this would have enhanced the validity of the OES. Another limitation was the poorly targeted population. Patients on the poor side of the quality of life spectrum were missing. Future studies with the OES-DV should examine patients with other types of elbow disorders in a larger population, because dimensionality examinations of questionnaires are influenced by the study population, and the greater the sample size, the greater the power in detecting misfit.

## Conclusion

Rasch analysis of the OES-DV showed that the data fit the stringent Rasch model. The multidimensionality of the

English version of the OES was partially confirmed. The four items of the function and three items of the pain domain were recognized as separate domains. The category rating scale of the OES-DV worked well. The OES distinguished 3.4 strata. This conclusion can only be applied to elbow dislocations, which were the largest group of patients studied.

## Acknowledgment

Oxford and Isis Outcomes, a part of Isis Innovation Limited, are acknowledged for their kind support. Oxford Elbow Score © Isis Innovation Limited, 2008. All rights reserved. The authors, Professor Ray Fitzpatrick and Dr Jill Dawson, have asserted their moral rights.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Dawson J, Doll H, Boller I, et al. The development and validation of a patient-reported questionnaire to assess outcomes of elbow surgery. *J Bone Joint Surg Br.* 2008;90:466–473.
2. Dawson J, Doll H, Boller I, et al. Comparative responsiveness and minimal change for the Oxford Elbow Score following surgery. *Qual Life Res.* 2008;17:1257–1267.
3. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* 2000;25:3186–3191.
4. Floor S, Overbeke AJ. Questionnaires on the quality of life in other than the Dutch language used in the Nederlands Tijdschrift voor Geneeskunde (Dutch Journal of Medicine): the translation procedure and arguments for the choice of the questionnaire. *Ned Tijdschr Geneesk.* 2006;150:1724–1727. Dutch.
5. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol.* 1993;46:1417–1432.
6. Bond TG, Fox CM. *Applying the rasch model. Fundamental measurement in the human sciences.* 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.
7. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? when should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum.* 2007; 57:1358–1362.
8. Fisher WP Jr. Reliability statistics. Rasch measurement transactions. 1992. Available from: <http://www.rasch.org/rmt/rmt63i.htm>. Accessed May 3, 2011.
9. Wright BD, Masters G. Number of person or item strata. Rasch measurement transactions. 2002. Available from: <http://www.rasch.org/rmt/rmt163f.htm>. Accessed May 3, 2011.

### Patient Related Outcome Measures

#### Publish your work in this journal

Patient Related Outcome Measures is an international, peer-reviewed, open access journal focusing on treatment outcomes specifically relevant to patients. All aspects of patient care are addressed within the journal and practitioners from all disciplines are invited to submit their work as well as healthcare researchers and patient support groups. Areas covered will

Submit your manuscript here: <http://www.dovepress.com/patient-related-outcome-measures-journal>

Dovepress

include: Quality of life scores; Patient satisfaction audits; Treatment outcomes that focus on the patient; Research into improving patient outcomes; Hypotheses of interventions to improve outcomes; Short communications that illustrate improved outcomes; Case reports or series that show an improved patient experience; Patient journey descriptions or research.