WHICH QUESTIONNAIRE IS BEST? THE RELIABILITY, VALIDITY AND EASE OF USE OF THE PATIENT EVALUATION MEASURE, THE DISABILITIES OF THE ARM, SHOULDER AND HAND AND THE MICHIGAN HAND OUTCOME MEASURE

J. J. DIAS, R. A. RAJAN and J. R. THOMPSON

J Hand Surg Eur Vol 2008 33: 9
DOI: 10.1177/1753193407087121

WHICH QUESTIONNAIRE IS BEST? THE RELIABILITY, VALIDITY AND EASE OF USE OF THE PATIENT EVALUATION MEASURE, THE DISABILITIES OF THE ARM, SHOULDER AND HAND AND THE MICHIGAN HAND OUTCOME MEASURE

J. J. DIAS, R. A. RAJAN and J. R. THOMPSON

From the University Hospitals of Leicester, Glenfield Hospital, Leicester, UK and the University of Leicester, Princess Road West, Leicester, UK

The Patient Evaluation Measure (PEM), The Michigan Hand Outcome Questionnaire (MHQ) and the Disabilities of the Arm, Shoulder and Hand (DASH) score were assessed independent of their originators for reliability, construct and criterion validity and acceptability, using an ease of use questionnaire. These were administered in random order to 100 patients with different hand and wrist disorders and with different impairments of movement, pain, sensation and strength. The internal consistency of all three questionnaires was very high suggesting redundancy in the questions. All questionnaires were reproducible and valid for finger and wrist disorders, but less for nerve disorders. All had poor construct validity. The PEM was the easiest to understand and complete, taking the least time. Correlation between the scales is high and conversion equations were calculated. All three are reliable and reproducible patient completed questionnaires, but the PEM is the easiest to use. The validity of all is suspected for nerve disorders.

The Journal of Hand Surgery (European Volume, 2008) 33E: 1: 9–17

Keywords: hand surgery, outcome measures, assessment

Hand outcome questionnaires allow for meaningful comparison of the results of therapy, aid in research and are clearly important for medicolegal purposes.

The Patient Evaluation Measure (PEM) was described in 1995 by Macey and Burke (1995) following an international multidisciplinary meeting in the UK. It is a questionnaire that evaluates the process of treatment, the current state of the hand and provides an overall assessment. The latter two sections have 14 questions, posed simply and with seven possible answers, presented as a categorised visual analogue scale. Six questions relate to symptoms, three to the impact of the disorder on the patient, two to satisfaction and three to general disability and handicap. The answers are expressed as a percentage disability ranging from zero to 100. This questionnaire is reliable, valid and responsive for assessing wrist disorders (Dias et al., 2001).

The Upper Extremity Collaborative Group developed the Disabilities of the Arm, Shoulder and Hand (DASH) score in 1996 (Hudak et al., 1996). It has 34 questions with five interval answers each. Sixteen of these relate to specific functions, five to symptoms, two to impact on the patient and 11 to general disability and handicap. There are no questions investigating satisfaction. This questionnaire has been demonstrated to be valid and reliable by the authors. It is widely translated and used for research reports.

The MHQ was devised from the University of Michigan in 1998, also using psychometric principles (Chung et al., 1998). It has 63 questions and measures six domains: overall hand function, activities of daily living, work performance, pain, aesthetics and patient satisfaction (12 questions) with hand function. Of these, the domains of function and pain refer to symptoms (15 items), those of work and ADL to disability and handicap (22 questions). The scoring system is complex, but clearly defined. The right and left hand can be individually assessed. The originators found the MHQ to be valid and reliable.

These three questionnaires have not been compared previously, although, more recently, the use of the PEM and DASH has been reported in patients with carpal tunnel syndrome (Hobby et al., 2005). In the study reported in this paper, the three questionnaires were investigated for their individual reliability, validity and ease of use and the overall scores given by the scales were compared, one against the other. This study was carried out independently of the originators of these questionnaires.

PATIENTS AND METHODS

One hundred patients with a range of wrist and hand disorders were prospectively recruited over a 6-month period from a single Hand Surgical Unit which serves a population of almost one million people. All patients were of consenting age and ability. They were required to have unrestrained movement of the hand and wrist without any form of splintage or cast. Patients had to understand and be able to complete all three questionnaires.

Sixty men and 40 women were recruited with a mean age of 49.8 (range 19–85) years. Forty patients had a disorder of their right hand, 40 patients had a disorder of their left hand and 20 had a bilateral disorder. The range of different disorders (Table 1) presented by the 100 patients is a reflection of the general caseload for a typical Hand Surgical Unit. It allowed for an assessment of the robustness of the three questionnaires for a typical caseload, with a wide range of symptoms and impairment of objective parameters. Since different disorders have different characteristics, the disorders were divided into three clinical groups for analysis: nerve ($n = 26$), wrist ($n = 27$) and finger ($n = 47$) disorders.

Table 1—Spread of diagnoses

Wrist ($n = 27$)	
Trapeziectomy	1
Stiffness	3
Fusion	2
Kienböck's disease	2
Pain	5
Fractures	3
Ganglion	1
Instability	1
Osteoarthritis	2
Flexor tendon calcification	1
Scaphoid fracture – conservative management	5
Scaphoid fracture – operative management	1
Nerve ($n = 26$)	
Ulnar nerve disorders	1
Median nerve disorders	25
Fingers ($n = 47$)	
Stiffness	7
Trigger finger	5
Cysts/ganglia	5
Tendon injury	5
Joint replacement	1
Fractures	5
Mallet injury	2
De Quervains	2
Dupuytren's disease	15

At the first interview, each patient was given a set of the three questionnaires, the PEM, DASH and MHQ. In this study, we investigated both Section 2 of the PEM alone on hand health with 11 questions and Sections 2 and 3 with 14 questions. Results were similar. Therefore, only the 14 question assessments are presented.

The order of the three questionnaires within the pack was randomised, so that different patients completed the questionnaires in different orders. Each questionnaire was set out in a similar manner, precisely as described by the questionnaire originators. In order to assess the ease of use of each questionnaire, a short standardised questionnaire (Table 2) was devised and included at the end of the PEM, DASH and MHQ. Patients were required to document the start and completion times for each questionnaire. An informal interview was conducted with every patient after completion of the questionnaires to assess their views.

Patients with carpal tunnel syndrome were, in addition, required to complete the Levine Symptom Score (Levine et al., 1993). In patients with wrist disorders, the Gartland and Werley (1951) score was calculated.

At the first interview, objective measures of key pinch and grip strength were measured using a calibrated Jamar pinch meter and a calibrated Jamar dynamometer with a constant setting at the second position (Betchtol, 1954). A mean of three readings was taken in kilograms for both hands. The average strength of both hands was recorded for patients with bilateral disorders (Table 3).

The range of wrist and finger movements was measured with a goniometer. The sum of flexion, extension, radial deviation and ulnar deviation was recorded for those with wrist disorders. Total active motion was recorded for the finger disorders. The average value was recorded in patients with bilateral disorders.

Table 2—Ease of use of the questionnaires

1. *This questionnaire was*:									
Very easy to understand							Very difficult to understand		
1	2	3	4	5	6	7	8	9	10
2. *This questionnaire was*:									
Very easy to complete							Very difficult to complete		
1	2	3	4	5	6	7	8	9	10
3. *I found the questions*:									
Very useful							Completely useless		
1	2	3	4	5	6	7	8	9	10
4. *I found the questions*:									
Very relevant to my disorder							Completely irrelevant		
1	2	3	4	5	6	7	8	9	10
5. *List the two things you like least about this questionnaire*:									
6. *List the two things you like most about this questionnaire*:									

Table 3

		Affected hands ($n = 100$)			*Unaffected hands* ($n = 80$)		
		Median	*Quartiles*		*Median*	*Quartiles*	
(a) *Objective measures on affected and unaffected hands*							
Pinch	Fingers ($n = 26$)	7.5	5.0	9.5	8.0	7.0	10.9
	Nerve ($n = 27$)	6.5	4.0	7.8	7.0	5.0	8.3
	Wrist ($n = 47$)	5.5	4.5	7.4	7.5	5.5	9.8
Grip		22.0	13.3	34.8	30.0	22.5	45.0
		25.5	12.0	30.5	28.0	23.0	36.0
		19.0	12.3	32.5	30.0	21.5	42.0

	PEM			*DASH*			*MHQ*		
	Median	*Quartiles*		*Median*	*Quartiles*		*Median*	*Quartiles*	
(b) *Outcome measures in affected hands*									
Fingers ($n = 47$)	60.7	42.9	82.1	75.8	58.1	93.4	66.7	52.8	87.2
Nerve ($n = 26$)	55.4	38.1	72.6	61.8	35.6	78.8	59.9	46.1	70.5
Wrist ($n = 27$)	61.9	44.0	84.5	75.7	61.8	89.7	83.4	59.4	93.9
All ($n = 100$)	60.7	42.0	77.2	72.8	57.4	89.7	69.3	52.7	85.4

Swelling, tenderness and sensory disturbance were recorded using an ordered categorical scale, as this is a common clinical practice. To reduce interobserver variability, all of the objective and subjective measures were carried out by a single investigator.

To assess reproducibility, 26 patients were randomly selected to complete a second pack of three questionnaires with the ease of use questionnaire for each. The order of the questionnaires was, again, randomised. The interval between the first and second administrations of the questionnaires was varied with an average of 1 day (45 minutes to 11 days).

The distributions of the objective measures and the questionnaires scores were skewed and so were summarised using their medians and quartiles. Cronbach's (1951) alpha was used to measure how well the battery of questions within a single questionnaire measure a single underlying construct. Test–retest differences were assessed using t-tests. Pearson correlations were used to measure the association between outcome measures and one-way analysis of variance was used to compare the ease of use measures in the three questionnaires. Regressions between scores of the different questionnaires were fitted using least squares. Where the relationships were non-linear, two joined linear regressions were use to summarise the relationship.

RESULTS

The mean DASH score was 69.1% (SD 22.8%), the mean MHQ score was 67.0% (SD 22.2%) and the mean PEM score for 14 questions was 60.2% (SD 23.5%). There was strong correlation between the total scores of the three questionnaires. This would suggest that these questionnaires were measuring similar aspects of hand function. The correlation coefficient of the DASH with the MHQ and PEM was 0.82 while that between the MHQ and PEM was 0.76. These are demonstrated in Fig 1. The scores for the three scales were lower for nerve disorders than for finger or wrist disorders (Table 3b). The correlations between scales were much higher than between these scales and the objective measures noted in Table 4.

Table 4—Cronbach's alpha for PEM, DASH and the components of MHQ ($n = 100$)

	Cronbach's alpha
DASH	0.98
PEM	0.94
MHQ	
Overall function	0.93
ADL	0.96
Work	0.94
Pain	0.82
Aesthetics	0.87
Satisfaction	0.93
Between domains	0.90

Reliability

To assess reliability, internal consistency and reproducibility were investigated. Internal consistency was

Table 5—Test–retest differences ($n = 26$)

	Test–retest differences		*Correlations between test–retest differences*	
	Mean	*95% CI**	*DASH*	*PEM*
DASH	0.1	−4.7 to 4.9		
PEM	−3.5	−9.3 to 2.3	0.2	
MHQ	−1	−4.3 to 2.2	0.3	0.5**

*CI – confidence interval.
**$p = 0.02$.

Table 6—Validity

	Objective measures			*Subjective measures*			**Scores*
	Pinch	*Grip*	*TAM*	*Swelling*	*Tenderness*	*Sensation*	L, G/W
Correlations between questionnaire scores, objective and subjective measures							
Fingers							
DASH	**0.47**	**0.61**	**0.30**	−0.01	**−0.47**	−0.04	
PEM	**0.44**	**0.53**	**0.41**	−0.01	**−0.37**	−0.16	
MHQ	**0.55**	**0.64**	**0.41**	−0.02	**−0.41**	−0.15	
Nerves							
DASH	**0.69**	**0.70**		−0.27	**−0.63**	0.15	−0.33
PEM	**0.57**	**0.52**		−0.12	**−0.66**	−0.07	−0.37
MHQ	**0.57**	**0.60**		−0.17	**−0.64**	−0.05	−0.31
Wrists							
DASH	**0.63**	**0.60**	0.25	0.05	−0.15	−0.37	−0.17
PEM	**0.51**	**0.58**	**0.57**	0.05	−0.29	−0.35	−0.14
MHQ	**0.52**	**0.61**	0.36	0.10	−0.19	−0.34	−0.03

*Correlations are individually significant (*bold*) at the 5% level if , for fingers ($n = 47$) the absolute values are >0.29, for nerves ($n = 26$) they are >0.38 and for the wrist ($n = 27$) they are >0.37. *Levine Scores (1993) for nerve disorders and Gartland and Worley (1951) scores for wrist disorders.

expressed by looking at the correlation of each item with each of the others to generate an unstandardised Cronbach's (1951) alpha. Each was considered to be internally consistent if Cronbach's alpha was between 0.7 and 0.9 (Shrout and Fleiss, 1979). Cronbach's alpha in excess of 0.9 suggests possible redundancy in the questionnaire.

Table 4 shows the values of Cronbach's alpha for the full PEM and DASH questionnaires and the domains of the MHQ. The value for the DASH scale is very high suggesting that, if all that is required is a single score, then it might be possible to ask fewer than 34 questions. The other values are closer to the target range. All three questionnaires were so highly internally consistent, that there was the possibility that certain items in each questionnaire are redundant.

Reproducibility was assessed by checking that instruments yield stable scores over time among respondents whose disorders are assumed not to have changed in the interval between the first and second completion of each questionnaire. Table 5 shows the mean differences between the first and second administration of each questionnaire with the 95% confidence intervals. None of the changes is significant and the magnitudes of the average changes are very small compared to the standard deviation of the scores in the whole sample. There was a significant correlation of test–retest differences between the PEM and MHQ, which both showed a decrease over time. The difference between the DASH scores did not correlate with those of the PEM or MHQ.

Validity

Validity is the ability of an instrument to measure what it is intended to measure (Guyatt et al., 1987, 1989). Construct validity is established when the measure relates to objective clinical measurements of hand function in an expected way. The validity of each of the three questionnaires was assessed separately in the three pathological groups: finger disorders, nerve disorders and wrist disease. How well the scores of the three questionnaires in the three pathological groups correlated with clinical measures of pinch, grip, total active motion (TAM), swelling, tenderness and sensation was determined. Table 6 shows that, in the finger group, all three questionnaires correlated significantly with the objective measures of pinch, grip and TAM. There was also a negative correlation with tenderness, but no noticeable association with swelling or sensation. Similar findings were found in the nerve group, in which TAM was not measured. Surprisingly, there was little correlation between the questionnaire scores and sensation. In the wrist group, there was good correlation with pinch and grip for all three questionnaires. PEM correlated best with wrist movement.

Criterion validity is present when the scores correlate with an accepted measure or "gold standard" of the condition being evaluated. Since there are no "gold standard" outcome measures in hand surgery (Deyo, 1984; Engelberg et al., 1996), the responses of the three questionnaires were compared against the Levine symptom score for the nerve group and the Gartland and Werley score for the wrist group. All three questionnaires in the nerve group did not correlate well with the Levine symptom score. In the wrist group also, none correlated well with the Gartland and Werley score. For none of the three questionnaires were we able to establish criterion validity (Table 6).

Prediction

The relationships needed for conversion between the measured values on the three scales were investigated by plotting (Figs 1–3) and linear regression over ranges

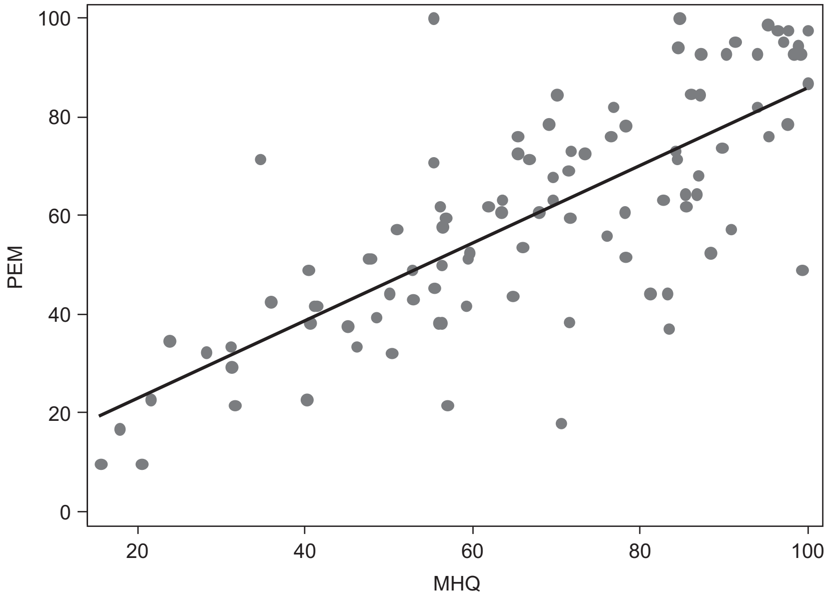

Fig 1 The relationship between PEM and MHQ is very close to linear with a slope of 0.80 (se 0.07, $p<0.0001$, $R^2 = 58\%$). The formula for predicting PEM from MHQ is PEM = 6.70 + 0.80 MHQ.

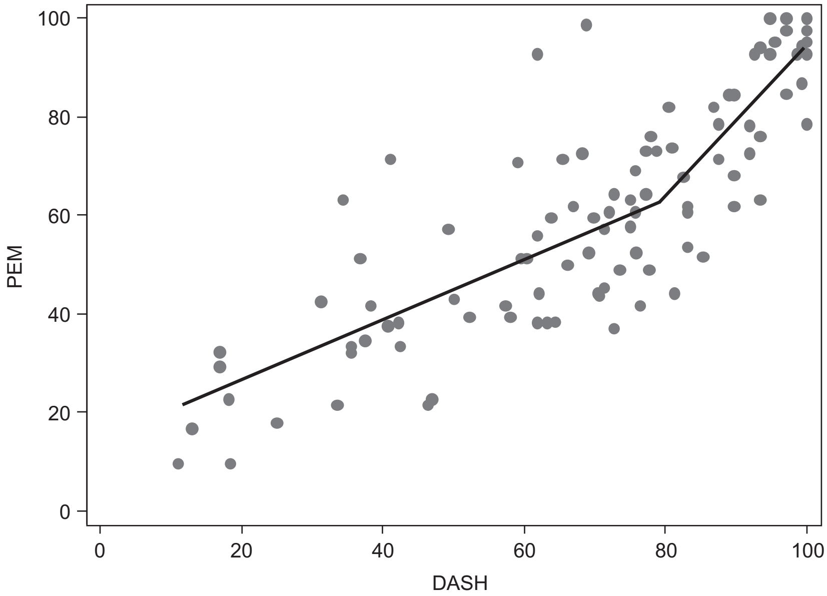

Fig 2 The relationship between PEM and DASH is non-linear. The formulae for predicting PEM from DASH are PEM = 12.31 + 0.65 DASH, DASH <80 and PEM = −86.97 + 1.89 DASH, DASH $\geqslant 80$. $R^2 = 69\%$, p-values for the slopes $p<0.001$ and $p<0.001$.

where the trend appeared linear. MHQ and PEM are linearly related over the whole range, although PEM scores tend to be lower (Table 3b). DASH is not linearly related to either PEM or MHQ, as DASH gives a greater proportion of the available range for differentiating between subjects who score low on PEM or MHQ. It is correspondingly less sensitive to changes in subjects with high scores. To represent this non-linearity, separate linear regression lines were fitted above and below a cut-off selected by eye. These regressions should be treated as a useful, but approximate, description of the relationship because a far larger data set would be required to investigate the non-linearity more fully. The plots show that the variation of individuals about the regression line is large; for MHQ against PEM the mean square error was 15.0. Consequently, the relationship will not be useful for converting the scores of individuals. However, the relationship will be accurate enough to convert the average scores of large groups of subjects.

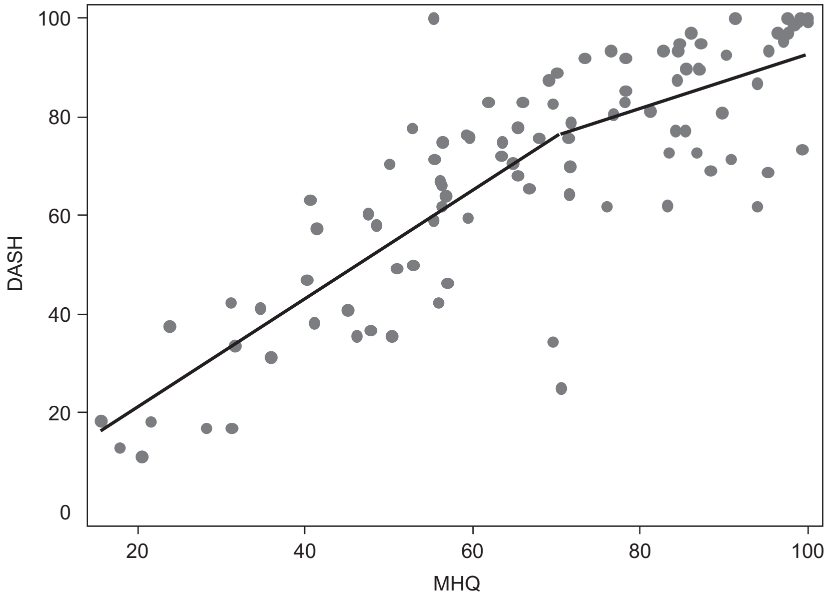

Fig 3 The relationship between DASH and MHQ is non-linear. The formulae for predicting DASH from MHQ are DASH = 0.88 + 1.10 MHQ, MHQ <70 and DASH = 39.24 + 0.53 MHQ, MHQ $\geqslant 70$. $R^2 = 70\%$, p-values for the slopes $p<0.001$ and $p = 0.001$.

Table 7—Ease of use

	Time (*m*)	*Ease of understanding*	*Ease of completion*	*Usefulness*	*Relevance*
Mean (*standard deviation*) *of various measures*					
PEM	**4.0** (2.7)	**2.1** (1.8)	**2.2** (1.7)	**3.9** (2.6)	**3.8** (2.6)
DASH	**6.2** (4.3)	**2.5** (2.1)	**2.6** (2.3)	**4.0** (2.7)	**3.9** (2.7)
MHQ	**8.7** (3.8)	**2.9** (2.2)	**2.9** (2.2)	**4.4** (2.5)	**3.9** (2.4)
**p*-value	<0.0001	0.0003	0.005	0.19	0.93
Test–retest differences (*n* = 26) – *Mean*, (*standard deviation*) *and p value*					
PEM	0.0 (2.2)	−0.2 (2.3)	0.1 (1.7)	0.2 (1.3)	0.1 (1.8)
	0.93	0.74	0.82	0.76	0.83
DASH	−0.7 (2.5)	0.2 (1.2)	−0.1 (1.4)	0.3 (1.5)	0.1 (1.4)
	0.20	0.50	0.27	0.45	0.78
MHQ	−2.2 (2.8)	−0.6 (1.2)	−0.8 (2.3)	−0.5 (2.1)	−0.7 (2.7)
	0.001	0.02	0.11	0.22	0.18

**p* values relate to a test of difference between all three questionnaires.

Ease of Use

The ease of use of each of these questionnaires was assessed using a simple questionnaire (Table 2) which was devised to determine the ease of understanding, ease of completion, usefulness and relevance. Patients also documented the time taken to complete the questionnaires. Table 7 shows the means for these measures, with their respective standard deviations. The PEM took the least time (mean 4 minutes) to finish and was found to be the easiest to understand and complete. The converse was true for the MHQ, which took 8.7 minutes to complete. These differences between questionnaires were significant for the time taken to complete the questionnaire, ease of understanding and completion. Patients thought that all three were equally useful and relevant to their disorder, although the scores obtained for all three suggest that patients were not fully convinced about these attributes for all three questionnaires. Test–retest reliability assessment of the ease of use did not demonstrate any difference between questionnaires for the ease of completion, usefulness or relevance, but did demonstrate a significant improvement ($p = 0.02$) in the ease of understanding of the longer questionnaires on subsequent administration. The times taken for completion of the longer questionnaires improved on the retest by 0.7 minutes for the DASH and 2.2 minutes for the MHQ, these differences being significant ($p = 0.001$) and listed in Table 7.

DISCUSSION

The last decade has seen the evolution of different questionnaires to assess outcome in hand disorders. The PEM, DASH and MHQ are completed by the patient. These questionnaires are more efficient than those that require clinical assessment of objective measures, as there is no need for an out-patient attendance. This clearly can save money and time. These questionnaires give an overall view of hand function and disability. They provide information additional to that obtained by the more specific single disorder questionnaires, such as the Levine Score for carpal tunnel syndrome or the Gartland and Werley score for wrist disorders. The patient completed questionnaires can identify patients with poorer outcome who may require a more formal review by the doctor.

Most questionnaires are evaluated by the originators. The MHQ has been validated by its development team and has been found to be reliable. The DASH, likewise, has been found to be reliable and valid by its originators. The PEM was shown to be valid and reliable independently of its originators (Dias et al., 2001). To our knowledge, there has not been an independent study designed to look at all three questionnaires, comparing one to the other and assessing their reliability, validity and ease of use.

This study has established that all three questionnaires are reliable. Each has a very high internal consistency, suggesting that there is some redundancy in the questions and provides the possibility of further reduction of items. Such item reduction may compromise the richness of the data and the ability to analyse each item separately when investigating causes of unsatisfactory outcome. There must, therefore, be a threshold below which any score can only be used for monitoring rather than scrutiny.

Reproducibility evaluates whether an instrument yields the same results on repeated applications, assuming that no true change had occurred. This is done by test–retest to look at the degree of agreement between scores at the first and subsequent assessment. Although there is no agreement about the length of time which should elapse between measurements, it has been suggested by Deyo et al. (1991) that it should be up to 2 weeks to avoid recall. In this study, all three questionnaires were found to be reproducible.

There was correlation between the changes in the PEM and MHQ between completions suggesting that factors influencing these differences were similar for these questionnaires. We were unable to explain why these changes did not correlate with those of the DASH, particularly as the scores correlate highly.

Validity is the extent to which the instrument measures what it purports to measure. All three questionnaires have face validity as they were constructed by multidisciplinary groups of experts. Investigating construct validity revealed a very good correlation between scores and objective clinical measures of pinch and grip strength for all three questionnaires. However, their correlation with movement was not as good. The three questionnaires correlated with finger movement, but only the PEM correlated with wrist movement. There was no correlation in the finger, nerve and wrist cases between the outcome scores and swelling and sensation, which are subjectively assessed. Tenderness correlated well for the finger and nerve groups with the three outcome measures, but there was no correlation for patients with wrist disorders, apart from the PEM. There appears to be good, but incomplete, construct validity established for these questionnaires.

Since there is no “gold standard” outcome measure in health status measures to assess criterion validity, two disease specific measures in common use (Deyo, 1984; Engelberg et al., 1996) were used. The three questionnaire scores correlated poorly with both the Levine Score, specific for carpal tunnel syndrome, and the Gartland and Werley Score, specific for wrist disorders. The lack of correlation between the Gartland and Werley wrist score and the three outcome questionnaires probably reflects the construct of the wrist score, which includes symptoms, measurements and radiographic findings. Nevertheless, this lack of correlation is troubling and demands further scrutiny. What was surprising was the lack of correlation between all three questionnaires and objective assessment for nerve disorders of hypoaesthesia and the widely regarded, and used, carpal tunnel syndrome symptom score (Levine et al., 1993). The validity of all three questionnaires for nerve disorders is suspect and their use for assessing outcome in such disorders should be questioned until such validity is established.

Instruments should be acceptable to patients in order to minimise boredom and distress and ensure a good response rate, so that results are easier to interpret and less prone to the problems of non-response (Fitzpatrick et al., 1998). The accuracy of the score is in question if the patient, because of boredom, chooses to complete the questionnaire anyhow, rather than making an effort to provide as accurate a response as possible. Non-completion of questionnaires can arise as a result of difficulty in understanding of a distressing, or unacceptable, questionnaire, for whatever reason. Missing responses may compromise the validity. Other factors, such as the health and literacy of the respondent, can also influence the response rate. The way each questionnaire deals with missing responses is critical and must be clearly understood by the users of these measures. The ease of understanding, ease of completion, usefulness, relevance and time taken to complete the three questionnaires was assessed. The three questionnaires were found to be equally useful and relevant with no real difference between the three questionnaires. The PEM took the least time to complete and was found to be the easiest to understand and complete. We think that this may be due to more general features of the layout of the questionnaire, with the least number of items and an uncluttered layout compared to the other two questionnaires. The PEM phrased the questions in a simple style and a short, clear manner. Since this study was completed, the DASH is available in a shortened version. This will require further investigation of its ease of use.

Users of these questionnaires must be aware of the many extrinsic factors that may affect the outcome. These may be beyond the control of the investigator. Timing of intervention in the natural history of the disease and cultural differences in population groups studied, with different expectations, may influence the outcome. Cosmesis, any residual deformity or scarring, stiffness, reduced coordination and residual joint imbalance also impact on the final outcome. Psychological reactions such as depression and anxiety may follow and may impinge on the result of any treatment. Lack of motivation and the lack of desire to return to work can also contribute to a poorer outcome. Outstanding litigation and the expectation of compensation following injury can affect the impact of treatment.

Our study had limitations. It was carried out in a single academic centre. A set of three different questionnaires was assessed containing a large number of items. Patients may have experienced ‘‘tick box fatigue’’ towards the completion of the questionnaires. To reduce this possibility, the order of the questionnaires was randomised and an informal interview was carried out at the completion of the questionnaires. The time taken to complete the set of three questionnaires could be considerable. The ease of use questionnaire, which we devised, has not been independently validated. Finally, this study did not address the ability of each questionnaire to measure change (precision and responsiveness), which is a critical attribute, as all these outcome measures are designed to monitor change.

Direct comparison of questionnaires is a time consuming and difficult, but important, exercise. Collection of data must: (a) cover an adequate sample with a range of disorders, impairments and disabilities, (b) counter bias of boredom by randomising the order of questionnaire completion and (c) have robust and meticulous assessment of impairment. The investigator must have a clear understanding of scoring mechanisms, especially how each questionnaire handles missing responses. The analysis presents several challenges.

The methodology for the development of health questionnaires is now well developed and relatively straightforward (Streiner and Norman, 1995). For many medical conditions, this has lead to the construction of a vast array of alternative outcome measures which generate a score. Such a scored outcome questionnaire is also referred to as a scale. It is increasingly important to be able to make comparisons between these scales, both so that the best may be selected for a particular research project and to judge the results of trials that have used different scales to assess outcome. In this study, the non-linear relationship between questionnaires may suggest a ceiling effect.

Unfortunately, scale comparison is not as simple as scale construction because, even when scales relate to the same condition, the developers may have had slightly different objectives. The best scale will often be the one for which the developers’ aims most closely match those of the intended user. Some scales attempt, through psychometric techniques, to measure a single underlying construct, while others use clinimetric methods to include a range of questions important to patients or clinicians (Wright and Feinstein, 1992). Also, some scales are designed to give a single overall score, while others measure separate domains. If the objectives of the developers differed, then it is very difficult to say that one scale is preferable to another. What can be said is that any scale should be reproducible and as simple as possible to use. For this reason, any comparison needs to investigate the relative test–retest performance of the scales, the redundancy of questions and ease of use. Test–retest variation ought to be assessed using the differences, and not just the correlations (Bland and Altman, 1986).

Association between an objective measure and the scales is helpful in establishing their validity but the relative extent and form of the association is only a measure of comparative performance if the intention is to use the scales as surrogate measures for that particular objective measurement. In some circumstances, relative sensitivity to change might be important in choosing a scale. Yet, even this is complicated by the fact that one scale may be more sensitive to change within one particular range of severity, while another might be better elsewhere in the range.

Although the concept of the best scale is difficult, or even impossible, to define, conversion between scales is still a valid and sensible aim. By plotting the scores on one scale against those on another, it will be possible to assess whether the scales are linearly related and to develop conversion formulae for translating from one scale to another. Simple regression, with the usual model checking, will enable conversion of one measured score to another, while, if interest is in relating the ‘‘true’’ scores for an individual, free of measurement errors, this relationship can be estimated by using structural equations, provided that there are repeat measurements on the same individual (Kendall and Stuart, 1973). The resulting equations are likely to serve as useful approximations, especially when converting the average scores of groups of subjects, but it is unlikely that they will be useful for converting the scores of individuals, because of the large variation that can be expected about the lines.

This study concludes that the PEM, DASH and MHQ are good patient-completed questionnaires, which are reliable and reproducible. They are reasonably valid for wrist and finger disorders but may not be sufficiently valid for nerve disorders. They were all found to be equally useful and relevant to the disorders in our patients. The PEM was found to be the easiest questionnaire to complete, taking the least time.

References

Betchtol CO (1954). Grip test: use of dynamometer with adjustable handle spacings. Journal of Bone and Joint Surgery, 36A: 820–824.

Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1: 307–310.

Chung KC, Pillsbury MS, Walters MR et al. (1998). Reliability and validity testing of the Michigan Hand Outcomes questionnaire. Journal of Hand Surgery, 23A: 575–587.

Cronbach L (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16: 287–334.

Deyo RA (1984). Measuring functional outcomes in therapeutic trials for chronic disease. Controlled Clinical Trials, 5: 223–240.

Deyo RA, Diehr P, Patrick DL (1991). Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clinical Trials, 12: 142S–158S.

Dias JJ, Bhowal B, Wildin CJ et al. (2001). Assessing the outcome disorders of the hand. Is the patient evaluation measure reliable, valid, responsive and without bias? Journal of Bone and Joint Surgery, 83B: 235–240.

Engelberg R, Martin DP, Agel J et al. (1996). Musculoskeletal function assessment instrument: criterion and construct validity. Journal of Orthopedic Research, 14: 182–192.

Fitzpatrick R, Davey C, Buxton MJ et al. (1998). Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment, 2: 1–46.

Gartland JJ, Werley CW (1951). Evaluation of healed Colles' fractures. Journal of Bone and Joint Surgery, 33A: 895–907.

Guyatt GH, Walter S, Norman G (1987). Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Diseases, 40: 171–178.

Guyatt GH, Deyo RA, Charlson M et al. (1989). Responsiveness and validity in health status measurement: a clarification. Journal of Clinical Epidemiology, 42: 404–408.

Hobby JL, Watts C, Elliot D (2005). Validity and responsiveness of the patient evaluation measure as an outcome measure for carpal tunnel syndrome. Journal of Hand Surgery, 30B: 350–354.

Hudak PL, Amadio PC, Bombardier C (1996). Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand: the upper extremity collaborative group). American Journal of Industrial Medicine, 29: 602–608.

Kendall MG, Stuart A, 3rd Edn. The advanced theory of statistics volume 2: inference and relationship, London, Griffin, 1973.

Levine DW, Simmons BP, Koris MJ et al. (1993). A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. Journal of Bone and Joint Surgery, 75A: 1585–1592.

Macey AC, Burke FD (1995). Outcomes of hand surgery. Journal of Hand Surgery, 20B: 841–855.

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86: 420–428.

Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use, 2nd Edn., Oxford, Oxford University Press, 1995.

Wright JG, Feinstein AR (1992). A comparative contrast of climimetric and psychometric methods for constructing indexes and scales. Journal of Clinical Epidemiology, 45: 1201–1218.

Received: 29 January 2006
Accepted after revision: 5 July 2007
Mr. Joseph J Dias, University Hospitals of Leicester, UK. Tel.: +44 0116 2563089; fax: +44 0116 2502696.
E-mail: Joseph.dias@uhl-tr.nhs.uk.

doi:10.1177/1753193407087121 available online at http://jhs.sagepub.com