Rheumatol Int (2015) 35:1015–1020
DOI 10.1007/s00296-014-3201-1

Rheumatology INTERNATIONAL

ORIGINAL ARTICLE - VALIDATION STUDIES

Psychometric properties of the Mayo Elbow Performance Score

Derya Celik

Received: 10 September 2014 / Accepted: 19 December 2014 / Published online: 31 December 2014

Abstract To translate and culturally adapt the Mayo Elbow Performance Score (MEPS), a widely used instrument for evaluating disability associated with elbow injuries, into Turkish (MEPS-T) and to determine psychometric properties of the translated version. The MEPS was translated into Turkish using published methodological guidelines. The measurement properties of the MEPS-T (construct validity and floor and ceiling effects) were tested in 91 patients with elbow pathology. The reproducibility of the MEPS-T was tested in 59 patients over 7–14 days. The responsiveness of the MEPS-T was tested in a subgroup of 46 patients diagnosed with lateral epicondylitis and who received conservative treatment for 6 weeks. The interclass correlation coefficient (ICC) was used to estimate the test–retest reliability. The construct validity was analyzed with the disabilities of the arm, shoulder and hand (DASH), Visual Analog Scale (VAS) and the Short Form 36 (SF-36). Effect size (ES) was used to assess the responsiveness. The distribution of floor and ceiling effects was determined. The MEPS-T showed very good test–retest reliability (ICC 0.89). The correlation coefficients between the MEPS-T and DASH and VAS were −0.61 and −0.53, respectively ($p < 0.001$). The highest correlations were between the MEPS-T and the mental component summary ($r = 0.47$, $p = 0.001$) and role emotional ($r = 0.45$, $p = 0.001$). The MEPS-T ES, 0.50, was moderate (95 % CI 0.33–0.62). We observed no ceiling or floor effects. The MEPS-T represents a valid, reliable and moderately responsive instrument for evaluating patients with elbow disease.

D. Celik (✉)
Division of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Istanbul University, 34740 Bakırköy, Istanbul, Turkey
e-mail: ptderya@hotmail.com

Keywords MEPS-T · Reliability · Validity

Introduction

Patient-reported outcome (PRO) measures provide insights from the patient's perspective into the impact of disease and treatment on their health and quality of life. PRO measures are categorized as generic or disease- or joint-specific. Generic measures often reflect health-related quality of life questions that are relevant across different diseases and populations. In contrast, specific measures include areas of importance related to a specific disease. In clinical studies, both generic and disease-specific measures are often included, with disease-specific measures often considered the primary outcome [1].

Numerous PRO measures to evaluate elbow dysfunction have been described, but there is no universal agreement regarding which PROs should be used because many of them lack reliability data [2]. This problem may be due to the fact that it is difficult for any single scoring system to adequately capture the impact of disease and treatments related to the full spectrum of elbow pathology. The PROs that have been used to assess elbow diseases include the Mayo Elbow Performance Score (MEPS), Oxford elbow score (OES), Disabilities of the arm, shoulder and hand (DASH), Visual Analog Scale (VAS) and the patient-rated tennis elbow evaluation (PRTEE) [3–6]. Short-Form Health Survey (SF-36) is a generic score that can be used to establish a health profile of the patients with elbow pathology [7]. The MEPS, designed to measure pain, stability, range of motion and the patient's ability to accomplish functional tasks, is one of the most commonly used physician-based and joint-specific elbow rating system [3].

Before instruments that evaluate outcome measures can be used in different regions of the world, they must be translated, culturally adapted, and retested to ensure

Table 1 Demographics of the patients

	$n = 91$ %
Age, years (mean ± SD)	42.9 ± 11.9 years
Female/male	49/42
Duration of symptoms	8.1 ± 1.2 months
Involved dominant/non-dominant	70/21
Occupation	
Housewife	26 (28.5)
Housekeeper	6 (6.5)
Government official	10 (10.9)
Laborer	8 (8.8)
Teacher	9 (9.9)
Massage therapist	4 (4.4)
Nurse/caregivers	7 (7.7)
Student	7 (7.7)
Banker	5 (5.5)
Secretary	3 (3.3)
Turner	6 (6.6)
Diagnosis	
Lateral epicondylitis	55 (60.4)
Medial epicondylitis	5 (5.5)
Olecranon bursitis	4 (4.4)
Contracture	9 (9.9)
Osteoarthritis	11 (12.0)
Radial head fracture	7 (7.7)

Values expressed as mean ± SD or *n*

the validity of the revised instruments [8]. In addition, the cross-cultural adaptations may contribute to a better understanding of the measurement properties of the outcome measures. Therefore, the purpose of this study was to translate and culturally adapt the English version of the MEPS into Turkish and investigate the reliability, validity and responsiveness of the translated version.

Methods

Translation and cross-cultural adaptation

Translation and cross-cultural adaptation of the MEPS was performed in five stages, as described by Beaton [8]. The Turkish version of the MEPS was named "MEPS-T."

Participants

Informed consent was obtained from all of the participants in the study; the informed consent form was approved by Istanbul University Research Foundation (Ethics committee approval date: December 23, 2011, IRB study protocol: 2011/2092-880). The study included patients seen between March 2012 and January 2013 at Istanbul University Department of Orthopedics and Traumatology. The eligibility criteria were as follows: (1) 18 years of age or older; (2) elbow pathology including lateral and medial epicondylitis, bursitis, contractures, osteoarthritis or radial head fracture and (3) the ability to read and write in Turkish. The diagnoses were established by a physician based on the patient history, physical examination and diagnostic imaging results. Maudsley's and Cozen's Test for lateral epicondylitis and Golfer's elbow test for medial epicondylitis were performed. X-rays has been taken to diagnose arthritis and fractures of the elbow joint. Range of motion was evaluated for the presence of contracture (Table 1). The patients with a history of inflammatory arthritis, neuropathic pain and gross structural abnormality of the elbow or any acute condition were excluded. In the first assessment, 91 patients with elbow pathology completed the MEPS-T (see the Appendix) and the previously validated Turkish versions of the DASH, SF-36 and VAS [9, 10].

Administration of PRO measures

The physical therapists administered the questionnaires in a random order to the patients in a waiting room after the patient's appointment with an orthopedic surgeon. The "range of motion" and "instability" subscales of the MEPS-T were assessed by the same physical therapist in the first and second assessments. The second assessment, in which the patients were asked to complete the MEPS-T again, occurred 7–14 days after the first MEPS-T to determine the test–retest reliability of the MEPS-T. To minimize the risk of short-term clinical change, no treatment was provided during this period. Responsiveness was assessed in a subgroup of 46 patients diagnosed with lateral epicondylitis who had conservative treatment for 6 weeks at the clinic. The patients were assessed at baseline and after 6 weeks of treatment.

Statistical analysis

All statistical analyses were performed with Stata version 11. (Stata Corp. LP., TX., USA). Descriptive statistics were calculated for all variables. These included frequency counts and the percentage for nominal variables and measures of central tendency (means and medians) and dispersion (standard deviations and ranges) for continuous variables. The measurement properties analyzed in this study for the instruments included internal consistency, test–retest reliability, construct validity and ceiling and floor effects.

Test–retest reliability

Test–retest reliability represents a scale's capability of yielding consistent results when administered on separate occasions during a period when an individual's status has remained stable [11]. The patients who reported "no change" in their condition between the first and second assessments were included in the analysis of test–retest reliability. Interclass correlation coefficient (ICC) was calculated using a 2-way mixed model ANOVA. The values of 0.4 or greater were considered satisfactory (specifically, $r = 0.81$–1.0 was excellent, 0.61–0.80 was very good, 0.41–0.60 was good, 0.21–0.40 was fair and 0.00–0.20 was poor) [12, 13].

Agreement

Agreement was assessed with the standard error of measurement (SEM) and minimal detectable change (MDC). The ICC was used to calculate the SEM, which is an index of measurement precision. The SEM was calculated as $\text{SD} \times \sqrt{(1-\text{ICC})}$. The MDC refers to the minimal amount of change that is within measurement error. The SEM was used to determine the MDC at the 95 % limits of confidence ($\text{MDC}_{95\,\%}$) and was calculated using the formula $1.96 \times \sqrt{2} \times \text{SEM}$ [14].

Validity

Validity is represented by the extent to which a score retains its intended meaning and interpretation [15]. In this study, we examined three aspects of validity: construct, convergent/divergent and content validity. Evidence for construct validity of the Turkish MEPS-T was provided by determining its relationship with the DASH, VAS and the PCS of the SF-36. The PF, RP and PCS of the SF-36 domains were used to assess convergent validity. Evidence for divergent validity was provided by determining the relationships with the MH, RE and MCS domains of the SF-36. Pearson correlation coefficients were calculated to assess construct and convergent/divergent validity. Content validity was assessed by the distribution of the scores and occurrence of ceiling and floor effects. Floor and ceiling effects of the MEPS-T at the first and second completion of the form were assessed by calculating the proportion of patients scoring the minimum or maximum values on the scale relative to the total number of patients. We considered scores between 0 and 10 % being minimum scores and scores between 90 and 100 % to be maximum scores. Floor and ceiling effects were considered to be relevant if greater than 30 % of the patients had a score at the limits of the scale [16].

Responsiveness

Responsiveness determines whether an instrument can detect clinical changes. Effect size (ES) was determined by calculating the differences in the means of baseline and follow-up data, divided by the standard deviation at baseline. A value between 0.20 and 0.50 was considered to be small effects, between 0.51 and 0.80 moderate effects, and between higher than 0.80 large effects [14].

Results

Translation and cross-cultural adaptation

No difficulties were encountered in translating the questionnaire, and the back translation corresponded very well to the original version. The questions were very simple to understand for the patients, so there was no need for cultural adaptation.

Measurement properties and testing

Table 1 provides the demographic and clinical characteristics of the patients. The descriptive statistics for the scores at baseline and at the second assessment of the MEPS-T are provided in Table 2. The mean ± SD duration of symptoms

Table 2 Descriptive statistics for the patient-reported outcome measures

	Mean ± SD	95 % CI
MEPS-T1 (first assessment)	58.2 ± 12.6	53.9–61.9
MEPS-T2 (second assessment)	59.7 ± 13.2	55.6–63.7
DASH	44.3 ± 17.7	39.6–50.9
VAS	3.6 ± 3.3	6.3–8.0
SF-36 (PF)	58.9 ± 22.3	51.7–65.5
SF-36 (RP)	21.5 ± 13.7	10.8–31.4
SF-36 (BP)	41.1 ± 22.1	47.1–56.1
SF-36 (GH)	51.0 ± 21.6	43.9–57.8
SF-36 (VT)	51.1 ± 23.7	44.7–57.6
SF-36 (SF)	69.8 ± 23.5	62.4–77.2
SF-36 (RE)	40.1 ± 20.2	27.3–52.9
SF-36 (MH)	58.0 ± 18.4	52.1–63.3
SF-36 (PCS)	36.1 ± 8.7	33.2–39.0
SF-36 (MCS)	43.6 ± 9.7	40.7–46.6

The Turkish version of the patient-reported outcome measures was used in this study

BP bodily pain, *GH* general health perceptions, *MCS* mental component scale, *MH* mental health, *PCS* Physical Component Scale, *PF* physical functioning, *RE* emotional role functioning, *RP* physical role functioning, *SF* social function, *VT* vitality

Table 3 Correlation between MEPS and other outcome measures in the literature and present study

Outcomes	MEPS-T
Oxford	
Pain	0.68*
Function	0.77*
Social–psychological condition	0.77*
SEV	0.59*
ASES	0.83*
Present study	
DASH	−0.61**
VAS	−0.53**
SF-36 (PF)	0.18
SF-36 (RP)	0.25
SF-36 (BP)	0.58**
SF-36 (GH)	0.37*
SF-36 (VT)	0.32*
SF-36 (SF)	0.38*
SF-36 (RE)	0.35*
SF-36 (MH)	0.35*
SF-36 (PCS)	0.33*
SF-36 (MCS)	0.43**

MEPS-T Mayo Elbow Performance Score—Turkish, *DASH* disabilities of the arm, shoulder and hand, *VAS* visual analog scale, *SEV* subjective elbow value, *BP* bodily pain, *GH* general health perceptions, *MCS* mental component scale, *MH* mental health, *PCS* Physical Component Scale, *PF* physical functioning, *RE* emotional role functioning, *RP* physical role functioning, *SF* social function, *VT* vitality

* p <0 .05); level of significance is only reported for the data of the current study

** $p < 0.01$); level of significance is only reported for the data of the current study

was 8.1 ± 1.2 months. Ninety-one patients (42 males; mean ± SD age: 49.2 ± 11.9 years; range 18–67 years) completed all of the questionnaires at the first assessment. Thirty-two of these patients did not return to the clinic for the second assessment. Therefore, of the 91 patients who participated at the first assessment, 59 patients (28 males; mean age: 42.8 ± 10.6 years; range 20–65 years) participated in the second assessment for the test–retest reliability analysis. Responsiveness was analyzed in the 46 patients (23 males; age: 42.8 ± 8.0 years; range 31–58) diagnosed with lateral epicondylitis.

Test–retest reliability

The average ± SD interval between the two assessments was 9.4 ± 2.4 days. The test–retest assessment had an ICC of 0.89, indicating excellent reliability.

Agreement

The SEM and MDC were 4.1 and 11.3, respectively.

Construct validity

The MEPS-T results correlated well with the results obtained using the DASH and VAS ($r = -0.61$ and $r = -0.53$, respectively; $p < 0.001$). The correlations between the results using the MEPS-T and the SF-36 are presented in Table 3. The MEPS-T was most strongly associated with the BP and MCS scales ($r = 0.58$ and $r = 0.43$, respectively; $p < 0.05$) of the SF-36. However, the MEPS-T showed poor and fair correlation with the PF and RP scales of the SF-36 ($r = 0.18$ and $r = 0.25$, respectively).

Floor and ceiling effects

The floor and ceiling effects and the number of items answered were identical during the test and retest examinations. None of the patients' scores were at the maximal or minimal value of the overall MEPS-T, indicating that there was no floor or ceiling effect. However, the subscales of the MEPS-T that were analyzed depended on the diagnosis. The "range of motion" and "stability" subscales of the MEPS-T showed high ceiling effects in patients with lateral epicondylitis. Of the 55 patients in the subgroup, 31 and 42 % reported maximal scores in the "range of motion" and "stability" subscales, respectively.

Responsiveness

For the 46 patients with lateral epicondylitis, the baseline scores of the MEPS-T were compared with the scores obtained after 6 weeks of treatment. The mean ± standard deviation of the baseline and post-treatment MEPS-T scores were 68.7 ± 14.4 and 76.0 ± 14.0, respectively, which resulted in a moderate (ES of 0.50, 95 % CI 0.33–0.62).

Discussion

This study test–retest reliability, validity and responsiveness data for the MEPS-T are provided. Based on our sample, the MEPS-T demonstrated acceptable levels of reliability, validity and responsiveness as a PRO questionnaire for Turkish-speaking individuals.

The test–retest reliability of the MEPS-T was excellent (ICC = 0.89), comparable to that reported previously

by Cusick et al [17]. The time interval between repeat measurements is an important issue when determining test–retest reliability. In general, the interval between repeat administrations for a PRO measure should be relatively brief (3–7 days) when the condition being measured is expected to change rapidly [11]. However, short test–retest intervals carry the risk of patients ''becoming familiar with the questions'' and simply answering based on memory of the first assessment. Although longer intervals can decrease this possibility, other factors need to be considered to prevent bias in such studies. Because the pain and function subscales of the MEPS consist of only nine questions, patients could easily remember the questions over a short time interval. In this study, an interval of 7–14 days was chosen to decrease the likelihood of this possibility and also to ensure an individual's condition had not changed. Similarly, Cusick et al. used a 2- to 3-week interval for retest assessment for the MEPS. The MDC was determined to be 11.3, indicating that a change of less than this value on repeated administrations of the MEPS-T should be considered a reflection of measurement error rather than a true change in the patient's condition.

Recent studies attempting to validate the MEPS have focused on determining the relationship of MEPS with PROs, including the OES, subjective elbow value (SEV), American Shoulder and Elbow Surgeons (ASES) and) [17–19]. In these studies, the highest levels of association were with the ASES and the function and social–psychological conditions of the OES ($r = 0.83$, $r = 0.77$, $r = 0.77$, respectively). Schneeberger et al. [19] used SEV for validity and found a very good correlation value ($r = 0.59$). In the present study, the DASH and the VAS were used for validity estimation and found to have a very good ($r = -0.61$) and good ($r = -0.53$) correlation, respectively. To determine convergent and divergent validity, we determined the level of associations between the scores on the MEPS-T and the eight domains and two summary scores for the SF-36. The MEPS-T was more strongly related to concurrent measures of MCS ($r = 0.43$) and BP ($r = 0.58$) than to concurrent measures of PF ($r = 0.18$) and PCS ($r = 0.33$). There is no literature with which to compare our results.

Ceiling effects occur when a measure's highest score is unable to assess a patient's level of ability. This can be especially common for PROs used on multiple occasions, thereby decreasing the likelihood that the testing instrument has accurately measured the intended subscales. In this study, the patients' "range of motion" and "instability" subscales were already high at the baseline because these symptoms are not typical in patients with lateral epicondylitis. Although many recent studies have used MEPS to assess lateral epicondylitis [20–23], we believe that MEPS is of limited use for lateral epicondylitis and it is not the best tool to use to assess patients with this condition. A disease-specific PRO such as the PRTEE should be considered for assessment of lateral epicondylitis.

Responsiveness, based on the completion of the MEPS-T at baseline and after 6 weeks of treatment, indicated an ES of 0.50 (95 % CI 0.33–0.62). Responsiveness has previously been reported after different elbow surgeries with a standardized response mean (SRM) of 1.26 and ES between 0.98 and 2.71 [19, 24], which is considered high compared to our result. These findings also suggest that MEPS-T is not the ideal PRO measure to assess patients with lateral epicondylitis.

One limitation of the study is that this is the first translation and cross-cultural adaptation study using the MEPS. In addition, physicometric properties of the original English version of the MEPS have not been reported. Therefore, we could not compare our results with those of previous studies.

Conclusion

The MEPS-T is brief and easy to administer and interpret, with a minimal investment of time required for the clinician or researcher. The MEPS-T is a reliable, valid and moderately responsive instrument that can be used as a PRO measure for Turkish-speaking individuals with elbow disease.

Clinical massages

The MEPS-T has sufficient reliability, validity and responsiveness, with values similar to those reported. The MEPS-T can be used as a PRO measure for Turkish-speaking individuals with various elbow pathologies.

Acknowledgments The author would like to thank Nilgun Turkel and Gulten Cetik for their excellent work during data collection.

Conflict of interest None.

Appendix

See Table 4.

Table 4 Mayo Dirsek Performans Skoru

Puan	
Ağrı (45 puan)	
Yok	45
Hafif	30
Orta	15
Şiddetli	0
Hareket açıklığı (20 puan)	
>100° fleksiyon	20
50–100° fleksiyon	15
<50° fleksiyon	5
Stabilite (10 puan)	
Stabil	15
Hafif instabilite (<10°varus-valgus laksitesi)	10
Tam instabilite (≥10°varus-valgus laksitesi)	5
Günlük Fonksiyon (25 puan)	
Saç tarayabilme	5
Yemek yiyebilme	5
Hijyen aktivitelerini yapabilme	5
Üstünü giyebilme	5
Ayakkabı giyebilme	5
Toplam puan	100

Mükemmel 90–100 puan; İyi 75–89 puan; Orta 60–74 puan; Kötü 60 puanın altında

References

1. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42
2. Longo UG, Franceschi F, Loppini M, Maffulli N, Denaro V (2008) Rating systems for evaluation of the elbow. Br Med Bull 87:131–161
3. Morrey BF, An KN (1993) Functional evaluation of the elbow. In: Morrey BF (ed) The elbow and its disorders. WB Saunders, Philadelphia, pp 74–83
4. Dawson J, Doll H, Boller I, Fitzpatrick R, Little C, Rees J et al (2008) The development and validation of a patient-reported questionnaire to assess outcomes of elbow surgery. J Bone Joint Surg Br 90:466–473
5. Macdermid J (2005) Update: the patient-rated forearm evaluation questionnaire is now the patient-rated tennis elbow evaluation. J Hand Ther 18:407–410
6. Hudak PL, Amadio PC, Bombardier C (1996) Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med 29:602–608
7. Ware JE Jr, Sherbourne CD (1992) The MOS 36-item short-form health survey [SF-36] Conceptual framework and item selection. Med Care 30:473–483
8. Beaton DE, Bombardier C, Guillemin F, Ferraz MB (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 25:3186–3191
9. Duger T, Yakut E, Oksuz C (2006) The reliability and validity of Turkish version of DASH questionnaire. Physiother Rehabil 17:99–107
10. Kocyigit H, Aydemir O, Fisek G (1999) Reliability and validity of Turkish version of short form SF-36. Med Treat 12:102–106
11. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF (2003) A comparison of two time intervals for test–retest reliability of health status instruments. J Clin Epidemiol 56:730–735
12. Kane RL (1997) Outcome measures. Understanding health care outcomes research. Aspen, Gaithersburg, pp 17–18
13. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
14. De Vet HC, Terwee CB, Bouter LM (2003) Current challenges in clinimetrics. J Clin Epidemiol 56:1137–1141
15. Irrgang JJ, Marx RG (2007) Clinical outcomes in sport and exercise physical therapies. In: Kolt GS, Synder-Mackler L (eds) Physical therapies in sports and exercise. Elsevier, Edinburgh, pp 206–219
16. Nunnally JC, Bernstein IR (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York
17. Cusick MC, Bonnaig NS, Azar FM, Mauck BM, Smith RA (2014) Throckmorton TW^2. Accuracy and reliability of the Mayo Elbow Performance Score. J Hand Surg Am 39:1146–1150
18. de Haan J, Goei H, Schep NW, Tuinebreijer WE, Patka P, den Hartog D (2011) The reliability, validity and responsiveness of the Dutch version of the Oxford elbow score. J Orthop Surg Res 30(6):39
19. Schneeberger AG, Kösters MC, Steens W (2014) Comparison of the subjective elbow value and the Mayo Elbow Performance Score. J Shoulder Elbow Surg 23:308–312
20. Raeissadat SA, Rayegani SM, Hassanabadi H, Rahimi R, Sedighipour L, Rostami K (2014) Is Platelet-rich plasma superior to whole blood in the management of chronic tennis elbow: one year randomized clinical trial. BMC Sports Sci Med Rehabil 18(6):12
21. Dzugan SS, Savoie FH 3rd, Field LD, O'Brien MJ, You Z (2012) Acute radial ulno-humeral ligament injury in patients with chronic lateral epicondylitis: an observational report. J Shoulder Elbow Surg 21:1651–1655
22. Kim JW, Chun CH, Shim DM, Kim TK, Kweon SH, Kang HJ, Bae KH (2011) Arthroscopic treatment of lateral epicondylitis: comparison of the outcome of ECRB release with and without decortication. Knee Surg Sports Traumatol Arthrosc 19:1178–1183
23. Garg R, Adamson GJ, Dawson PA, Shankwiler JA, Pink MM (2010) A prospective randomized study comparing a forearm strap brace versus a wrist splint for the treatment of lateral epicondylitis. J Shoulder Elbow Surg 19:508–512
24. Dawson J, Doll H, Boller I, Fitzpatrick R, Little C, Rees J, Carr A (2012) Specificity and responsiveness of patient-reported and clinician-rated outcome measures in the context of elbow surgery, comparing patients with and without rheumatoid arthritis. Orthop Traumatol Surg Res 98:652–658