Published in final edited form as:
J Orthop Sports Phys Ther. 2012 ; 42(8): 716–723. doi:10.2519/jospt.2012.4038.

Comparison of Reliability and Responsiveness of Patient-Reported Clinical Outcome Measures in Knee Osteoarthritis Rehabilitation

Valerie J. Williams, DPT, MS[1], **Sara R. Piva, PT, PhD, OCS, FAAOMPT**[2], **James J. Irrgang, PT, PhD, ATC, FAPTA**[3], **Chad Crossley, DPT, MS**[1], and **G. Kelley Fitzgerald, PT, PhD, FAPTA**[4]

[1]Staff Physical Therapist, Centers for Rehabilitation Services, Pittsburgh, PA

[2]Assistant Professor, Department of Physical Therapy, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA

[3]Associate Professor, Director of Clinical Research, Department of Orthopaedic Surgery, University of Pittsburgh School of Medicine, Pittsburgh PA

[4]Associate Professor, Department of Physical Therapy, School of Health and Rehabilitation Sciences, Director of Physical Therapy Clinical and Translational Research Center, University of Pittsburgh, Pittsburgh PA

Abstract

Study Design—Secondary analysis, pre-treatment:post-treatment observational study.

Objective: C—ompare the reliability and responsiveness of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), the Knee Outcome Survey-Activities of Daily Living Scale (ADLS), and the Lower Extremity Functional Scale (LEFS) in individuals with knee osteoarthritis (OA).

Background—The WOMAC is the current standard in patient-reported measures of function in patients with knee OA. The ADLS and LEFS have been designed for potential use in patients with knee OA. If the ADLS and/or LEFS are to be considered viable alternatives to the WOMAC in measuring patient-reported function in individuals with knee OA, then they should have measurement properties that are comparable to the WOMAC. It would also be important to determine whether either of these instruments may be superior to the WOMAC in terms of reliability or responsiveness in this population.

Methods—Data from 168 subjects with knee OA who participated in a rehabilitation program were used in the analyses. Reliability and responsiveness of each outcome measure were estimated at 2, 6, and 12 month follow-up time points. Reliability was estimated by calculating the intraclass correlation coefficient (2,1) for subjects who were unchanged in status from baseline at each follow-up time point, based on a global rating of change score. To examine responsiveness, the standard error of the measure (SEM), minimum detectable change (MDC), minimum clinically important difference (MCID), and the Guyatt responsiveness index (GRI) were calculated for each outcome measure at each follow-up time point.

Results—All 3 outcome measures demonstrated reasonable reliability and responsiveness to change. Reliability and responsiveness tended to decrease somewhat with increasing follow up

Corresponding Author Information: G. Kelley Fitzgerald, PT, PhD, FAPTA, University of Pittsburgh, Department of Physical Therapy, 6035 Forbes Tower, Pittsburgh, PA 15260. Phone: 412-383-6643, Fax: 412-648-5970, kfitzger@pitt.edu.

time. There were no substantial differences between outcome measures for reliability or any of the 3 measures of responsiveness at any follow-up time point.

Conclusions—The results do not indicate that one outcome measure is superior to another with regards to reliability and responsiveness when applied to subjects with knee OA. We believe all 3 instruments are appropriate outcome measures to examine change in functional status of patients with knee OA.

Keywords

Function; Measurement; Psychometrics; Clinimetrics; Physical Therapy

INTRODUCTION

Patient-reported outcome measures are commonly used to assess symptoms, functional status, or change in disability as a result of treatment in patients with knee osteoarthritis (OA). The current standard disease specific instrument for knee and hip OA is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC).[15,23] The WOMAC is a self-reported instrument that includes 24 questions divided into 3 sections; pain (5 questions), stiffness (2 questions), and physical function (17 questions). The pain category assesses pain elicited during activities of daily living (ADL), while the stiffness category assesses the amount of stiffness elicited after staying in certain positions and the time of day it is experienced. The physical function category measures the patients' ability to perform certain activities including; going from sit to stand, walking, stair negotiation, putting on socks, etc. The WOMAC is scaled using either a 10 cm visual analog scale (VAS), or 5 point Likert Scale (0–4).[4,23,26] The WOMAC total score is determined by combining the scores from all three sections (range: 0–240 for VAS scale, 0–96 for 5 point Likert scale version). Higher scores represent greater problems with pain and function. It has been well documented that the WOMAC is reliable, valid, and responsive to change in patients with hip and knee OA.[4,6,7,26]

Other patient-reported instruments have also been used as a means of assessing functional status in patients with knee OA. Two of these are the Knee Outcome Survey-Activities of Daily Living Scale (ADLS) and the Lower Extremity Functional Scale (LEFS). The ADLS is a 14 item knee specific patient-reported measure that can be used to assess functional status in patients with a variety of knee disorders, including knee OA.[13,20] It consists of 14 questions devised to ascertain limitations in daily activity imposed by symptoms such as pain, swelling, and instability (6 questions), and difficulty performing functional activities such as walking, going up and down stairs and raising from a chair (8 questions).[13] The ADLS score ranges from 0–100 with higher scores representing better function. Individual items are scored using a 6-point system similar to the Likert Scale. Although it has been shown to be reliable, valid, and responsive in patients with a variety of knee conditions, including knee OA,[13,17] a thorough examination of these properties specifically for use in individuals with knee OA has not been reported.

The Lower Extremity Functional Scale (LEFS) is a region specific patient-reported measure of function that can be used to assess functional status in patients with a variety of lower extremity disorders, including knee OA.[5] The LEFS consists of 20 questions designed to assess the degree of "difficulty" of specific functional tasks. It is a 5-point Likert like scale (0=extreme difficulty, 4=no difficulty). The LEFS score ranges from 0 to 80 with higher scores representing better function. Although the LEFS has been found to be reliable, valid, and responsive in multiple populations with lower extremity dysfunction,[5,25,28] similar to the ADLS, a more in depth evaluation of these properties specifically in individuals with knee OA is warranted.

The purpose of this study is to report the reliability and responsiveness of the ADLS and the LEFS in individuals with knee OA, and to compare these clinimetric properties to those for the WOMAC. Given that the WOMAC is the current standard in patient-reported measures of function in patients with knee OA, if the ADLS and/or LEFS are to be considered viable alternatives for this population, reliability and responsiveness characteristics should at least be comparable to those of the WOMAC.

METHODS

Subjects

The subjects included in this study were taking part in a randomized clinical trial studying the effects of 2 different exercise approaches in people with knee OA.[8] Subjects were included in the study if they met the 1986 American College of Rheumatology clinical criteria for knee OA[1] and had grade II or greater Kellgren and Lawrence radiographic OA of the tibiofemoral joint.[16] Subjects with only patellofemoral OA were excluded. Subjects were also excluded if they could not safely participate in the treatment programs, had a history of 2 or more falls within the previous year, or were unable to ambulate 100 feet without the use of an assistive device or rest break. Other exclusionary criteria included total knee arthroplasty, uncontrolled hypertension, history of cardiovascular disease, history of neurological disorders affecting the extremities, corticosteroid injection into the quadriceps or patellar tendon within the past month or 3 within the past year, quadriceps or patellar tendon rupture, patellar fracture, severe visual impairment, or pregnancy. All subjects signed an informed consent approved by the University of Pittsburgh Institutional Review Board prior to participating in the study.

One hundred and eighty three subjects, (122 females (67%) and 61 males (33%)), 63.9±8.8 years of age with an average body mass index of 30.5±6.5 were enrolled in the clinical trial. Of these, only those with complete baseline data and complete questionnaire data for at least 1 follow up at 2 months, 6 months or 12 months were included in the analysis for the current study. One hundred sixty eight subjects met this criterion. Complete data sets were considered to be full completion of the ADLS, LEFS, WOMAC, and Global Rating of Change (GRC) questionnaires at the time point of interest.

Procedures

After completing baseline questionnaires (ADLS, LEFS, WOMAC), subjects participated in 12 sessions of a supervised exercise program conducted over 6 to 8 weeks, depending on the subject's schedule. The exercise program has been described in detail elsewhere and consisted of lower extremity strengthening, stretching, aerobic, balance, and agility exercises.[8] Follow up questionnaire data were collected at 2, 6, and 12 months after enrollment in the study. Subjects also completed the GRC questionnaire at each of these follow up time points. The GRC allows the subject to rate the extent to which they perceive their condition has changed over time using a 15 level scale. The GRC ranges from 1("a very great deal better"), to 8 ("about the same"), to 15 ("a very great deal worse"). The GRC was used as an external anchor to determine responsiveness.[14]

Statistical Analysis

Statistical analyses were performed using SPSS 16.0 and Microsoft Excel 2007. For the WOMAC, we used the total score in the analysis. The LEFS and WOMAC scores were transformed to a 0 to 100 scale so that higher scores for all outcome measures indicated fewer symptoms and a high level of function. The ADLS data did not have to be transformed. Change scores from baseline to each follow-up time point were calculated for each patient-reported outcome measure in a manner that positive change scores indicated

improvement. Outcome measures at each time point, as well as the change scores, were all found to be normally distributed using the Shapiro-Wilk test and visual observation of histograms.

Reliability—Test-retest reliability of each questionnaire was determined at each follow-up time point by calculating intra-class correlation coefficients (ICC), using model (2,1) with 95% confidence intervals for those subjects who reported they had not changed based on the GRC. We considered subjects unchanged if their GRC scores were 7 ("a tiny bit better (almost the same)"), 8 ("about the same"), and 9 ("a tiny bit worse (almost the same)").

Responsiveness—To investigate responsiveness, subjects were classified as improved or not improved on the basis of the GRC score. Those with a GRC score between 1 ("a very great deal better") and 5 ("somewhat better") were categorized as improved, and those with scores between 6 ("a little bit better") and 15 ("a very great deal worse") were categorized as not improved.

Internal Responsiveness—Internal responsiveness is usually measured within the context of randomized trials or repeated measures designs, and its statistics are based on the distribution of the data. Internal responsiveness was assessed for each measurement tool at each time point. Three indices of internal responsiveness were calculated including the standard error of the measurement (SEM), minimum detectable change (MDC), and Guyatt's responsiveness index (GRI).[10,12,27] The SEM was calculated to determine the measurement error as follows; $SEM=S_x*\sqrt{(1-r_{xx})}$, where S_x is the standard deviation at baseline of the measurement tool from the total sample and r_{xx} is the reliability coefficient for that measurement tool.[3] The SEMs are reported in the same units as the data for the measurement tool, and is reflective of the instrument's measurement precision.[21,27] The MDC was calculated as the amount of change needed to be certain, within a defined level of statistical confidence, that change is beyond measurement error.[18,21] The MDC was calculated at the 90% and 95% level of statistical confidence: $MDC_{90} =SEM_x*1.64*\sqrt{(2)}$, and $MDC_{95} =SEM_x*1.96*\sqrt{(2)}$. The GRI was calculated as the ratio of mean change of those improved divided by the standard deviation of change of those not improved.[10,12]

External Responsiveness—External responsiveness reflects the ability of an instrument to detect a clinically meaningful change based on an external anchor. First, Spearman's correlation coefficient (rho) was calculated between change scores of each functional questionnaire at each time point and the corresponding GRC score to determine the strength of the relationship between the change for each patient-reported outcome measure and the external anchor. Next, receiver operating characteristic (ROC) curves (and the area under the curve (AUC) and associated 95% CI for the AUC) were generated for each outcome measure at each time point. The ROC curve plots the sensitivity of the outcome measure on the Y-axis and 1 minus the specificity of the outcome measure on the X-axis. Traditionally a minimum clinically important difference (MCID) is identified by finding the point on the ROC curve that maximizes both specificity and sensitivity. We used the Youden Index (sensitivity+specificity-1) to identify this cut-point.[19] The Youden Index is considered to be a good quantitative estimate of a cut-point that maximizes correct classification and minimizes incorrect classification when sensitivity and specificity are weighted equally.[19]

The Youden Index described above is recommended when equal weight is given to sensitivity and specificity in identifying a cut-point for an MCID.[19] In some cases, clinicians may be interested in maximizing specificity in identifying an MCID. Higher specificity would indicate lower probability of falsely classifying someone as improved. Consequently the MCID that has a high specificity is desirable to minimize falsely classifying someone as

improved when they are not. Therefore, we also elected to identify the MCID that maximized specificity by identifying the change score that had a specificity of 0.8 (or closest to 0.8 in some cases) for differentiating between those subjects who perceived improvement from those who perceived no improvement.

RESULTS

Demographic data for the subjects included in the analyses are provided in Table 1. Table 2 provides descriptive statistics for the baseline and follow-up scores and change scores for each measurement tool at each time point. Responsiveness assessments were based on the data from 159 subjects at the 2 month time point, 153 subjects at the 6 month time point, and 142 subjects at the 12 month time point.

Reliability

Reliability was based on the data from 26 subjects at the 2 month time point and 28 subjects at the 6 month and 12 month time points who reported they had not changed based on the GRC. The ICCs and their 95% confidence intervals are presented in Table 3. All of the measurement tools had good reliability, with ICCs ranging from 0.75 to 0.93. The reliability was higher when measured over a shorter period of time. As the time increased between baseline and follow-up measures, the reliability generally decreased and the 95% CIs for the reliability statistics were wider.

Internal Responsiveness

The SEM, MDC, and GRI for each measurement tool at each time point are displayed in Table 3. The SEM for each tool ranged from 4.52 to 7.60 for the ADLS, from 6.35 to 8.17 for the LEFS, and from 5.09 to 7.21 for the WOMAC over the course of the study. There did not appear to be one tool with a consistently larger or smaller SEM. As the length of follow-up increased, the SEM for each outcome measure increased slightly, with the exception of the LEFS for which the SEM for the 2 months and 6 months follow-up were similar. The MDCs for the 90 and 95% CIs for all of the outcome measures followed a similar pattern to the SEMs, increasing over time. In examining the point estimates and the 95% confidence intervals of the GRI for each time point in Table 3, the point estimates for each instrument are contained within the 95% confidence intervals between each instrument. Therefore there does not appear to be a significant difference in responsiveness between the instruments at any time point, based on the GRI.

External Responsiveness

At 2 months 65% of the subjects perceived themselves to be improved, based on the GRC data. The percentage was 64% at 6 months and 57% and 12 months. Spearman's Rho values and the AUC are displayed in Table 3. Spearman's Rho correlation coefficients ranged from −.30 to −.53 indicating that the change scores on the outcome measures had small to moderate correlations with the GRC. The Figure displays the ROC curve for the ADLS, LEFS, and WOMAC at 2 months, and the AUC from each instrument was .71, .69, and .70 respectively. The ROC curves for 6 and 12 months were very similar in shape. The MCID values using the Youden Index (method 1) and the specificity equal to .80 method (method 2) to identify the MCID are provided in Table 4.

DISCUSSION

Reliability and Internal Responsiveness

To our knowledge, this is the first study to compare the reliability and responsiveness of the WOMAC, LEFS, and ADLS specifically in subjects with knee OA. We found each of these patient reported outcome measures to have similar levels of reliability and responsiveness.

The reliability and responsiveness of each of the outcome measures decreased over time, which is likely due to greater measurement error over time. As the length of follow-up increased, the decreased reliability was concomittant with decreased measurement precision as evidenced by larger SEMs and MDCs for all 3 instruments. The higher SEMs at increasing time points from baseline indicate that over time a higher change score is needed to reliably detect a statistically meaningful change, which is probably a function of both a decrease in the ICC and an increase in the variability of the scores. This pattern would hold true for each of the outcome measures examined in this study. The increasing MDCs show that as time progresses a larger change score is needed to be 90% and 95% confident that true change has occurred. For example, the MDC_{95} of the WOMAC increases from 14.1at 2 months, to 15.0 at 6 months, and finally to 18.5 at 12 months. Greco et al,[9] performed a study comparing the reliability and responsiveness of 4 outcome measures, one of which was the WOMAC total score, in patients who had undergone surgery for articular cartilage lesions of the knee. These authors found that as time increased from 6 months to 12 months follow up the MDC_{95} for the WOMAC total score increased from 10.9 to 15.3.[9] This demonstrates the importance of considering the time frame and patient population of the study when choosing an MDC.

External Responsiveness

When using the ROC curve analysis to determine the MCID it is important to consider the AUC when interpreting the results. The AUC represents the probability that the outcome measure would be able to distinguish an individual who perceives improvement from an individual who perceives no improvement. If the AUC is relatively small, then the confidence one may have in the calculated MCID may be low, because it indicates a lower level of ability for the change score in the outcome measure to accurately discriminate between responders and non-responders based on the external anchor. Hosmer et al,[11] have reported criteria for classifying the AUC as a measure of the strength of the relationship between the change score in the outcome measure and external anchor for discriminating between responders and non-responders based on AUC values. An AUC between 0.7 and 0.8 demonstrates acceptable discrimination. An area between 0.8 and 0.9 demonstrates excellent discrimination, and an area between 0.9 and 1.0 indicates outstanding discrimination. The AUC for our data at all time points were all close to 0.7, which is considered to be acceptable for finding a minimal clinically important difference.[11] The reason our AUCs were not higher is likely because our outcome measures were only moderately correlated with the GRC.

Because the AUCs in our ROC plots were at the lower end of what could be considered adequate discrimination of the outcome measures, the confidence in our calculated MCID using the Youden Index is somewhat limited. As a result of this, we propose an alternative which is to determine the MCID where specificity is equal to .80. We propose this alternative to derive MCID values to allow researchers to decide which MCID they may wish to select based on the intended use of the outcome measure. To illustrate this, consider the MCID for the WOMAC for the 0 to 2 month follow-up period reported in Table 4. If a clinician wanted to select an MCID that yielded the best combination of sensitivity and specificity, they would select an MCID of 4, calculated using the Youden Index (method 1).

If, however, one wanted to have greater confidence in identifying patients who perceived themself to be improved, a change score of greater than 8.8 calculated using the specificity equal to .80 approach (method 2) might be considered. Method 1 has only reasonable sensitivity and specificity, whereas method 2 has good specificity at the expense of low sensitivity. The limitations of using the MCID values are reflected by the likelihood ratios that are close to one. The positive likelihood ratios were at most 2.3, indicating that changes above the MCID are somewhat limited in identifying subjects who improved.

Previous studies have determined the MCID of the same patient-reported measures used in the present study. Angst et al,[2] investigated the responsiveness of the WOMAC in subjects with OA of the knee and hip. The MCID was the percent change in the WOMAC score corresponding to a small change in a global rating scale 3 months after physical therapy intervention. The MCID to show improvement was equal to a 17–22% change from baseline.[2] Tubach and colleagues[24] derived the MCID for improvement in the WOMAC function scale in subjects with knee OA after 4 weeks of non-steroidal anti-inflammatory drugs. The MCID was the change score in which most subjects stated they had a good improvement based on a 5 point Likert scale.[24] The methods used in studies by Angst et al and Tubach et al do not allow calculation of the accuracy of the MCID cut-off value to detect patient perceived improvement. Accuracy can be determined by the sensitivity and specificity of the MCID cut-off by using the ROC curve method. Watson et al[25] determined the MCID for the LEFS in adults with anterior knee pain following rehabilitation. They reported the AUC to be .77, and a change of 12 points in the LEFS maximized sensitivity and specificity (both at .67).[25] Greco and colleagues[9] determined the MCID of the WOMAC total score in patients with articular cartilage lesions of the knee that underwent surgical repair. They reported the AUC to be .71, and MCID of 11.5 for both 6 and 12 months follow-up. Sensitivity and specificity of the MCID values were .79 and .57 at 6 months, and .84 and .55 at 12 months respectively.[9] Piva et al[20] determined the MCID of the ADLS in younger patients with patellofemoral pain following 2 months of rehabilitation to be 7.1, with an AUC of .83, corresponding to both sensitivity and specificity of .78. Conceptually, the MCID should be greater than the MDC so that one could be sure the change score was beyond measurement error. We suggest that when the accuracy of the MCID is questionable, and when the MDC is available, perhaps the use of the MDC may be more appropriate than the use of MCID.

Stratford et al[22] compared responsiveness between the WOMAC physical function subscale and the LEFS by examining the differences in standardized response means for the 2 instruments in subjects who had undergone total knee arthroplastic surgery. Measurements were taken within 16 days after surgery and then greater than 20 days after the first post-operative assessment. Both instruments were found to be responsive, however, there was no evidence of superiority of one instrument over the other. Although the sample is slightly different and the method of assessing responsiveness was different from our study, the findings by Stratford et al[22] are similar to those reported in our study.

Because there are so many factors to be considered when determining the psychometric properties of an outcome measure, providing the reader with multiple approaches may be valuable. First, the reliability, SEM, and MDC should be reported to evaluate the tool's ability to accurately measure change. When considering external responsiveness using ROC curves to determine a change which was meaningful to the patient it would be most helpful to know the strength of the relationship between the outcome measure and the external anchor, the AUC with its 95% CI, and the sensitivity and specificity of the MCID. Preferably, multiple cut-offs of change should be presented to give a more comprehensive understanding of the change seen in patients. The results of this study also support the concept that the psychometric properties of a patient-reported outcome measure are not a

fixed property of the outcome measure itself but rather an interaction between the patient-reported outcome measure and the circumstances in which the outcome measure is used (i.e. the patient population and intervention under study and the length of follow-up). As such, clinicians and researchers should use the psychometric characteristics of a patient-reported outcome measure that were established under the conditions that are most similar to the conditions in which the outcome measure will be used.

There may be concern that because our analyses were performed on subjects participating in a randomized trial and not on a broader, consecutively enrolled cohort of subjects with knee OA, our results may not completely apply to the broader population of people with knee OA across the entire disease spectrum. While this concern may be valid, Table 1 indicates that our sample did include patients across all levels of radiographic severity, the gender distribution is comparable to what is reported in epidemiological studies, and the mean age and body mass index is consistent with patients with knee OA who are typically seen in physical therapy clinics. The means for the function scores indicate moderate disability. Although we did have some subjects with more severe and less severe disability, a broader observation, such as would be provided in a broader longitudinal cohort study might yield different results. Nevertheless, we believe our results can relate to most patients receiving physical therapy for knee OA.

CONCLUSION

Based on the results of this study, the WOMAC, ADLS, and LEFS demonstrated similar reliability and responsiveness to change in patients with knee OA. Therefore our study indicates that all 3 outcome measures have similar psychometric properties when applied to subjects with knee OA. We believe all 3 instruments are appropriate outcome measures to examine change in functional status of patients with knee OA.

Acknowledgments

National Institute of Arthritis and Musculoskeletal and Skin Diseases 1-R01-AR048760

References

1. Altman R, Asch E, Bloch D, et al. Development of criteria for the classification and reporting of osteoarthritis. Arthritis Rheum. 1986; 29:1039–1049. [PubMed: 3741515]
2. Angst F, Aeschlimann A, Beat MA, et al. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. J Rheumatol. 2002; 29:131–138. [PubMed: 11824949]
3. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine. 2000; 25:3192–3199. [PubMed: 11124736]
4. Bellamy N, Watson-Buchanan W, Goldsmith CH, et al. Validation study of WOMAC: A health status instrument for measuring clincically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol. 1988; 15:1833–1840. [PubMed: 3068365]
5. Binkley JM, Stratford PW, Lott SA, et al. The Lower Extremity Functional Scale (LEFS): Scale development, measurement properties, and clinical application. Phys Ther. 1999; 79:371–383. [PubMed: 10201543]
6. Bischoff-Ferrari HA, Vandechend M, Bellamy N, et al. Validation and patient acceptance of a computer touch screen version of the WOMAC 3.1 osteoarthritis index. Ann Rheum Dis. 2005; 64:80–84. [PubMed: 15231508]
7. Davies GM, Watson DJ, Bellamy N. Comparison of the responsiveness and relative effect size of the Western Ontario and McMasters Universities osteoarthritis index and the short-form medical

outcomes study survey in a randomized clinical trial of osteoarthritis patients. Arthritis Care Res. 1999; 12:172–179. [PubMed: 10513507]

8. Fitzgerald GK, Piva SR, Gil AB, et al. Agility and perturbation training techniques in exercise therapy for improving pain and function in subjects with knee osteoarthritis: A randomized clinical trial. Phys Ther. 2011; 91:452–469. [PubMed: 21330451]
9. Greco NJ, Anderson AF, Mann BJ, et al. Responsiveness of the International Knee Documentation Committee subjective knee form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. Am J Sports Med. 2010; 38:891–902. [PubMed: 20044494]
10. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative insturments. J Chron Dis. 1987; 40:171–178. [PubMed: 3818871]
11. Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. 2. New York: John Wiley & Sons, Inc; 2000.
12. Husted JA, Cook RJ, Farewell VT, et al. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000; 53:459–468. [PubMed: 10812317]
13. Irrgang JJ, Snyder-Mackler L, Wainner RS, et al. Development of a patient-reported measure of function of the knee. Journal of Bone & Joint Surgery - American Volume. 1998; 80-A:1132–1145.
14. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials. 1989; 10:407–415. [PubMed: 2691207]
15. Jette AM, McDonough CM, Ni P, et al. A functional difficulty and functional pain instrument for hip and knee osteoarthritis. Arthritis Res Ther. 2009; 11:1–12.
16. Kellgren JH, Lawrence JS. Radiological Assessment of Osteo-Arthrosis. Ann Rheum Dis. 1957; 16:494–502. [PubMed: 13498604]
17. Marx RG, Jones EC, Allen AA, et al. Reliability, validity and responsiveness of four knee outcome scales for athletic patients. J Bone Joint Surg Am. 2001; 83-A:1459–1469. [PubMed: 11679594]
18. Nunnally, JC.; Bernstein, IH. Psychometric Theory. New York: McGraw Hill; 1994.
19. Perkins NJ, Schisterman EF. The inconsistency of optimal cut-points using two criteria based on the receiver operator curve. Am J Epidemiol. 2006; 163:670–675. [PubMed: 16410346]
20. Piva SR, Gil AB, Moore CG, et al. Responsiveness of the activities of daily living scale of the knee outcome survey and numeric pain rating scale in patients with patellofemoral pain. J Rehabil Med. 2009; 41:129–135. [PubMed: 19229444]
21. Portney, LG.; Watkins, MP. Applications to Practice. 2. Upper Saddle River, NJ: Prentice Hall Health; 2000. Foundations of Clinical Research.
22. Stratford PW, Kennedy DM, Hanna SE. Condition-specific Western Ontario McMaster Osteoarthritis Index was not superior to region specific Lower Extremity Function Scale at detecting change. J Clin Epidemiol. 2004:1025–1032. [PubMed: 15528053]
23. Sun Y, Sturmer T, Gunther KP, et al. Reliability and validity of clinical outcome measurements of the hip and knee: A review of the literature. Clin Rheumatol. 1997; 16:185–198. [PubMed: 9093802]
24. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis:the minimum clinically important improvement. Ann Rheum Dis. 2005; 64:29–33. [PubMed: 15208174]
25. Watson CJ, Propps M, Ratner J, et al. Reliability and responsiveness of the lower extremity functional scale and the anterior knee pain scale in patients with anterior knee pain. J Orthop Sports Phys Ther. 2005; 35:136–146. [PubMed: 15839307]
26. Whitehouse SL, Crawford RW, Learnmouth ID. Validation for the reduced Western Ontario and McMaster Universities osteoarthritis index function scale. J Ortho Surg. 2008; 16:50–53.
27. Wywrich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 2000; 6:39–49. [PubMed: 10807023]
28. Yeung TSM, Wessel J, Stratford P, et al. Reliability, validity, and responsiveness of the lower extremity functional scale for inpatients of an orthopaedic rehabilitation ward. J Orthop Sports Phys Ther. 2009; 39:468–477. [PubMed: 19487822]

KEY POINTS

Findings

There were no differences found in reliability or measures of responsiveness between the WOMAC, LEFS, or ADLS in our sample of people with knee OA. Both reliability and responsiveness were reduced somewhat with increasing follow-up time in all 3 instruments.

Implications

All 3 outcome measures are similarly reliable and responsive measures to detect change in functional status following a rehabilitation program in people with knee OA.

Caution

Because our analyses were performed on subjects participating in a randomized trial and not on a broader, consecutively enrolled cohort of patients with knee OA, our results may not ideally apply to the broader population of people with knee OA across the entire disease spectrum.

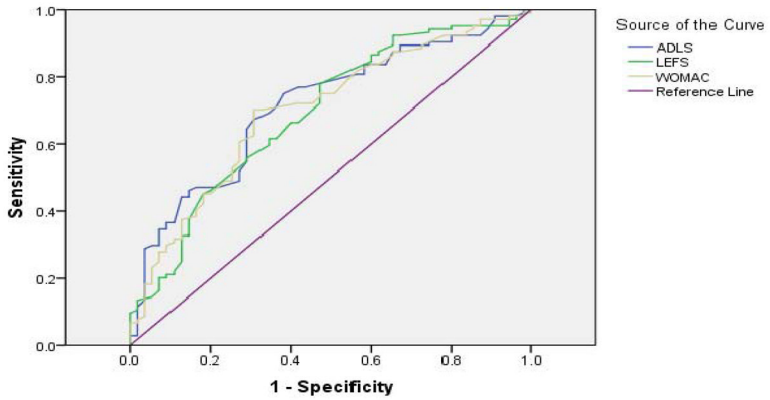

Figure 1.

Receiver operating characteristic curves for the changes in Knee Outcome Survey Activities of Daily Living Scale (ADLS), Lower Extremity Functional Scale (LEFS), and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for the 2 Month Follow-up. Curves are generated from normalized data where 0 is the lowest and 100 is the highest. Area under the curve for each instrument is as follows: ADLS=0.71, LEFS = 0.69, WOMAC = 0.70

Table 1

Demographic information for subjects included in the analyses N=168

Variable	Mean (SD) or Frequency (%)
Gender	
Female	110 (65.5%)
Male	58 (34.5%)
Age, y	64.4 (8.7)
Body Mass Index, kg/m^2	30.1 (6.2)
Radiographic Severity	
Tibiofemoral radiographs KL	
Grade	
1	1 (0.6%)
2	23 (13.7%)
3	83 (49.4%)
4	61 (36.3%)
Baseline ADLS (0–100)	68.3 (17.3)
Baseline LEFS (0–80)	52.3 (13.2)
Baseline WOMAC (0–96)	27.0 (15.2)

KL Grade = Kellgren and Lawrence Grade

ADLS = Knee Outcome Survey Activities of Daily Living (higher score is better function)

LEFS = Lower Extremity Function Scale (Original scale, non-transformed, higher score is better function)

WOMAC = Western Ontario and McMasters University Osteoarthritis Index (Original scale, non-transformed, lower score is better function)

Table 2

ADLS, LEFS and WOMAC Scores at 2, 6 and 12 Months. Data reported as mean (standard deviation).

		Improved				Not-Improved				All Subjects			
	Time Point	**N (%)**	**Baseline**	**Final**	**Change Score**	**N (%)**	**Baseline**	**Final**	**Change Score**	**N (%)**	**Baseline**	**Final**	**Change Score**
ADLS	2 months	104 (65%)	68.4 (16.8)	77.0 (16.5)	8.6 (12.5)	55 (35%)	68.0 (17.1)	68.9 (18.9)	0.9 (10.6)	159	68.2 (16.8)	73.8 (17.8)	5.6 (12.6)
	6 months	98 (64%)	69.2 (15.9)	77.8 (16.1)	8.6 (13.8)	55 (36%)	67.7 (20.0)	65.0 (21.0)	−2.7 (11.8)	153	68.6 17.4	73.2 (19.0)	4.6 (14.1)
	12 months	81 (57%)	69.2 (16.4)	81.3 (14.8)	12.1 (12.7)	61 (43%)	69.1 (17.8)	68.6 (20.9)	−.6 (14.3)	142	69.2 (17.0)	75.8 (18.7)	6.6 (14.7)
LEFS	2 months	104 (65%)	65.5 (14.7)	73.3 (17.3)	7.7 (11.9)	55 (35%)	65.4 (18.5)	65.6 (18.7)	.18 (11.0)	159	65.5 (16.0)	70.6 (18.1)	5.1 (12.1)
	6 months	98 (64%)	66.6 (13.1)	75.2 (15.5)	8.6 (13.1)	55 (36%)	63.7 (21.5)	61.5 (23.5)	−2.2 (14.1)	153	65.5 (16.6)	70.3 (19.8)	4.7 (14.4)
	12 months	81 (57%)	66.8 (14.7)	77.4 (15.8)	10.6 (13.0)	61 (43%)	65.5 (18.6)	64.2 (22.2)	−1.3 (15.3)	142	66.2 (16.4)	71.7 (19.8)	5.5 (19.2)
WOMAC	2 months	104 (65%)	71.9 (15.8)	81.0 (15.0)	9.1 (11.2)	55 (35%)	71.9 (15.6)	74.0 (16.5)	2.1 (9.2)	159	71.9 (15.7)	78.6 (15.9)	6.7 (11.0)
	6 months	98 (64%)	73.3 (14.0)	82.6 (13.6)	9.3 (12.6)	55 (36%)	69.9 (18.4)	69.3 (20.1)	−.6 (10.9)	153	72.0 (15.8)	77.8 (17.4)	5.7 (12.9)
	12 months	81 (57%)	73.2 (15.3)	83.9 (13.3)	10.7 (12.2)	61 (43%)	72.9 (15.3)	73.0 (19.8)	.1 (12.6)	142	73.1 (15.2)	79.2 (17.2)	6.1 (13.4)

All values represent transformed scores where 0 (lowest) to 100 (highest)

ADLS = Knee Outcome Survey Activities of Daily Living Scale

LEFS = Lower Extremity Function Scale

WOMAC = Western Ontario and McMasters University Osteoarthritis Index

Table 3

Measures of Reliability and Responsiveness for the ADLS, LEFS and WOMAC at 2, 6 and 12 Months

		ICC (95% CI)	SEM [+]	SEM [++]	$MDC_{95\%}$ [+]	$MDC_{95\%}$ [++]	$MDC_{90\%}$ [+]	$MDC_{90\%}$ [++]	Rho	AUC(95% CI)	GRI(95%CI)
0–2 months	ADLS	.93 (.85, .97)	4.5	4.5	12.5	12.5	10.5	10.5	−.41*	.71 (.63, .80)	.81 (.49, 1.14)
	LEFS	.81 (.63, .91)	6.9	5.6	19.2	15.4	16.1	12.9	−.30*	.69 (.61, .78)	.70 (.38, 1.03)
	WOMAC	.90 (.76, .95)	5.1	4.8	14.1	13.4	11.8	11.2	−.33*	.70 (.62, .79)	1.0 (.64, 1.35)
0–6 months	ADLS	.87 (.75, .94)	6.2	6.2	17.2	17.2	14.4	14.4	−.52*	.74 (.66, .82)	.73 (.40, 1.06)
	LEFS	.85 (.71, .93)	6.4	5.1	17.6	14.1	14.7	11.8	−.46*	.71 (.62, .79)	.61 (.31, .91)
	WOMAC	.88 (.76, .94)	5.4	5.8	15.0	16.1	12.6	13.4	−.51*	.72 (.64, .80)	.85 (.52, 1.20)
0–12 months	ADLS	.80 (.61, .90)	7.6	7.6	21.0	21.0	17.6	17.6	−.53*	.74 (.66, .82)	.85 (.54, 1.15)
	LEFS	.75 (.53, .88)	8.2	6.5	22.6	18.1	19.0	15.2	−.47*	.72 (.63, .80)	.69 (.40, .99)
	WOMAC	.81 (.63, .91)	6.7	7.2	18.5	20.0	15.5	16.7	−.44*	.70 (.61, .79)	.85 (.53, 1.16)

[+] Values reported are transformed scores with ranges from 0 to 100, where higher scores represent better function.

[++] Values are reported in outcome measure original scale where higher scores represent better function for the ADLS and LEFS, and lower scores represent better function for the WOMAC.

* correlation significant at the 0.01 level

$SEM = S_X * \sqrt{(1 - r_{XX})}$, where S_X=standard deviation at baseline of the measurement tool from the total sample and r_{XX}=reliability coefficient for that measurement tool

MDC 95%=SEM_X*1.96*√(2),

MDC 90%=SEM_X*1.64*√(2)

Rho=Spearman's correlation coefficient between each outcome measure and the GRC

AUC= probability that the outcome measure would be able to distinguish an individual who perceives them self to be improved from an individual who does not

GRI= mean change of those improved/standard deviation of change of those not improved

Table 4

MCID Values for the ADLS, LEFS and WOMAC Using Two Methods

			MCID [+]	Sensitivity	Specificity	+ LR	− LR	MCID [++]
0–2 months	ADLS	Method 1	2.2	.75	.62	2.0	.2	2.2
		Method 2	8.4	.47	.82	2.6	.4	8.4
	LEFS	Method 1	.6	.78	.53	1.7	.2	0.5
		Method 2	6.3	.46	.80	2.3	.5	5.5
	WOMAC	Method 1	4.0	.70	.69	2.3	.2	−5.5
		Method 2	8.8	.45	.80	2.3	.5	−8.5
0–6 months	ADLS	Method 1	5.6	.59	.78	2.7	.3	5.6
		Method 2	7.2	.57	.80	2.9	.4	7.2
	LEFS	Method 1	7.5	.52	.80	2.6	.4	1.5
		Method 2	7.5	.52	.80	2.6	.4	8.5
	WOMAC	Method 1	6.8	.57	.80	2.9	.4	−6.5
		Method 2	6.8	.57	.80	2.9	.4	−6.5
0–12 months	ADLS	Method 1	5.0	.72	.66	2.1	.2	5.0
		Method 2	10.6	.57	.80	2.9	.4	10.6
	LEFS	Method 1	1.3	.77	.59	1.9	.2	5.5
		Method 2	12.5	.41	.80	2.1	.5	12.5
	WOMAC	Method 1	1.6	.79	.54	1.7	.1	−1.5
		Method 2	12.0	.37	.80	1.9	.5	−11.5

[+]MCID = Minimum Clinically Important Difference values reported are transformed scores with ranges from 0 to 100, where higher scores represent better function.

[++] MCID = Minimum Clinically Important Difference values reported are in original scale of outcome measure where higher scores represent better function for the ADLS and LEFS, and lower scores represent better function for the WOMAC.

+LR = Positive Likelihood Ratio

−LR = Negative Likelihood Ratio

ADLS = Knee Outcome Survey Activities of Daily Living Scale

LEFS = Lower Extremity Function Scale

WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index

Method 1 = Minimum Clinically Important Difference identified using Youden Index

Method 2 = Minimum Clinically Important Difference identified using value with specificity at least .80