Reliability, Validity, and Responsiveness of Four Knee Outcome Scales for Athletic Patients


This information is current as of February 25, 2005

Subject Collections
Articles on similar topics can be found in the following collections

- Adult Disease (1077 articles)
- Adult Trauma (418 articles)
- Sports (99 articles)
- Knee (283 articles)
- Outcomes (120 articles)

Reprints and Permissions
Click here to order reprints or request permission to use material from this article, or locate the article citation on jbjs.org and click on the [Reprints and Permissions] link.

Publisher Information
The Journal of Bone and Joint Surgery
20 Pickering Street, Needham, MA 02492-3157
www.jbjs.org
Reliability, Validity, and Responsiveness of Four Knee Outcome Scales for Athletic Patients

BY ROBERT G. MARX, MD, MSC, FRCS(C), EDWARD C. JONES, MD, MA, ANSWORTH A. ALLEN, MD, DAVID W. ALTCHEK, MD, STEPHEN J. O’BRIEN, MD, SCOTT A. RODEO, MD, RILEY J. WILLIAMS, MD, RUSSELL F. WARREN, MD, AND THOMAS L. WICKIEWICZ, MD

Investigation performed at the Center for Clinical Outcome Research and the Sports Medicine and Shoulder Service, Hospital for Special Surgery, New York, NY

Background: Many patient-based knee-rating scales are available for the evaluation of athletic patients. However, there is little information on the measurement properties of these instruments and therefore no evidence to support the use of one questionnaire rather than another. The goal of the present study was to determine the reliability, validity, and responsiveness of four knee-rating scales commonly used for the evaluation of athletic patients: the Lysholm scale, the subjective components of the Cincinnati knee-rating system, the American Academy of Orthopaedic Surgeons sports knee-rating scale, and the Activities of Daily Living scale of the Knee Outcome Survey.

Methods: All patients in the study had a disorder of the knee and were active in sports (a Tegner score of ≥4 points). Forty-one patients who had a knee disorder that had stabilized and who were not receiving treatment were administered all four questionnaires at baseline and again at a mean of 5.2 days (range, two to fourteen days) later to test reliability. Forty-two patients were administered the scales at baseline and at a minimum of three months after treatment to test responsiveness. The responses of 133 patients at baseline were studied to test construct validity.

Results: The reliability was high for all scales, with the intraclass correlation coefficient ranging from 0.88 to 0.95. As for construct validity, the correlations among the knee scales ranged from 0.70 to 0.85 and those between the knee scales and the physical component scale of the Short Form-36 (SF-36) and the patient and clinician severity ratings ranged from 0.59 to 0.77. Responsiveness, measured with the standardized response mean, ranged from 0.8 for the Cincinnati knee-rating system to 1.1 for the Activities of Daily Living scale.

Conclusions: All four scales satisfied our criteria for reliability, validity, and responsiveness, and all are acceptable for use in clinical research.

Clinical research studies are placing an increased emphasis on the perspective of the patient with use of health-related quality-of-life instruments. Many patient-based knee-rating scales have been developed for the evaluation of athletic patients. There is little information on the measurement properties of the existing scales and therefore no evidence to support the use of one questionnaire rather than another. Furthermore, without evidence to support the reliability, validity, and responsiveness of these instruments, the results of clinical studies that are based on these instruments are in question.

While instruments that have been used to evaluate patients following knee replacement, shoulder surgery, or trauma have been studied in detail, relatively little work has been done to assess or compare knee instruments for the evaluation of athletic patients.

Previous work in this area has been cross-sectional in design and therefore has not assessed reliability or how the instruments performed over time (that is, responsiveness). Furthermore, those investigations focused exclusively on patients with anterior cruciate ligament deficiency, substantially limiting the generalizability of the results. The goal of the present study was to determine the reliability, validity, and responsiveness of the scales used to measure disability and symptoms in athletic patients with a wide variety of disorders of the knee.

Materials and Methods

Ten orthopaedic surgeons who specialized in sports-related problems of the knee were polled to determine which questionnaires were the most sound and widely used and therefore the most appropriate for study. The goal was to

Copyright © 2001 by The Journal of Bone and Joint Surgery, Incorporated
study questionnaires that measure disabilities rather than impairments. Disabilities are restrictions or the lack of an ability to perform an activity in the usual manner, such as an inability to run, walk, or play a sport. Impairments are defined as a loss or abnormality of psychological, physiological, or anatomical structure or function at the organ level, such as a decrease in range of motion or an increase in translation of a joint. Impairments are important to patients and to surgeons; however, they have less impact on patients’ quality of life than do disabilities.

The Lysholm scale, the American Academy of Orthopaedic Surgeons sports knee-rating scale, the Activities of Daily Living scale of the Knee Outcome Survey, and the Cincinnati knee-rating system were studied. The Cincinnati knee-rating system has been widely used; however, the scale has multiple subscores, including some based purely on impairment. While all components of this scale are relevant to both clinicians and patients, for the purposes of this research we elected to study only the sections relating to disability. We also administered version 1.0 of the Short Form-36 (SF-36) for the validity analyses. The SF-36 is a thirty-six-item questionnaire that measures general health. Its use has been encouraged in conjunction with knee-specific instruments for studies of patients with an injury of the anterior cruciate ligament. Both a physical component scale and a mental component scale can be derived from the SF-36. Other instruments were not evaluated because they had been recently developed or they focused on impairment.

The present study was not a direct statistical comparison of the reliability, validity, and responsiveness of the four knee-rating scales. The goal of the study was to determine whether the measurement properties of each were satisfactory for use in clinical research.

**Patient Recruitment**

The study was approved by the Institutional Review Board, and all subjects gave informed consent. Patients in the waiting rooms of orthopaedic surgeons who specialized in disorders of the knee were recruited for the study during a six-month period. The patients completed the questionnaires in the waiting room prior to seeing their physician. Patients who did not receive treatment in the interval (range, two to fourteen days) between the first and the second administration of the questionnaires. Patients were excluded if they had had surgery or a traumatic injury in the preceding three months.

When the patients completed the questionnaires for the second time, they also completed a transitional index in which they rated the severity of the knee condition at that time compared with the severity when the questionnaires were first administered. They chose from seven responses: “much worse,” “somewhat worse,” “a little worse,” “no change,” “a little better,” “somewhat better,” and “much better.” Only patients who responded “no change” were included in the reliability study. The intraclass correlation coefficient and the limits-of-agreement statistic were used to compare the scores. The intraclass correlation coefficient is an index of concordance for dimensional measurements, ranging between 0 and 1, where ≥0.75 is adequate for patients enrolled in a clinical trial. The limits-of-agreement statistic was also used as a descriptive measure of agreement. This value is the mean difference (and two standard deviations) between the two tests. Ninety-five percent of the differences between the two test administrations will lie within this interval.

**Validity**

Validity is an assessment of whether the instrument actually measures what it is intended to measure. A scale is considered to have face validity when its qualitative attributes are deemed to be adequate by individuals with experience in the field. Content validity is the appraisal of the underlying components of the scale. These two forms of validity were assessed by five orthopaedic surgeons with experience in the field of sports medicine. The experts were not asked to quantify these types of validity but rather to ensure that the instruments were valid from these points of view in order to make certain that...
clinicians would be comfortable using them for the evaluation of athletic patients with disorders of the knee.

Validation is relatively clear-cut when there is a gold standard against which the results can be compared. In cases where there is no gold standard, such as quality of life, we are forced to use “construct validation.” Construct validity is present when the instrument performs as expected in relation to another measurement. We used the patients’ and clinicians’ opinions of severity as well as the other knee-rating scales and the physical component scale of the Short Form-36 (SF-36) as indices to determine construct validity. The patients were asked to rate their condition on a 5-point scale as either “very mildly bothersome,” “mildly bothersome,” “moderately bothersome,” “severely bothersome,” or “very severely bothersome.” Similarly, clinicians were asked to rate the severity of the patient’s knee problem as “very mild,” “mild,” “moderate,” “severe,” or “very severe.”

Six hypotheses were proposed to assess the construct validity of the four knee-specific instruments:

1. Since the knee scales are more specific for abnormalities of the knee, we hypothesized that these instruments would all correlate better with each other than they would with the physical component scale or the mental component scale (Spearman correlation coefficient) and that the physical component scale would correlate more strongly with the knee scales, because it is more specific for physical function, than would the mental component scale.

2. Since the knee scales are more specific for disorders involving the knee, we hypothesized that the knee-rating scales would correlate better with each other (Spearman correlation coefficient) than they would with any of the eight SF-36 subscales.

3. Since certain subscales of the SF-36 are more related to symptoms and disabilities experienced by patients with knee disorders, we hypothesized that the knee scales would correlate better with physical function and role-physical (Spearman correlation coefficient) than they would with vitality or social function and that they would correlate better with general health, bodily pain, vitality, and social function than they would with role-emotional or mental health.

4. Since patient-rated and clinician-rated severity should approximate knee symptoms and disability, we hypothesized that the knee scales would be significantly correlated with clinician-rated and patient-rated severity (Spearman correlation coefficient ≥ 0.6, and p ≤ 0.05).

5. Since patient-rated and clinician-rated severity should approximate knee symptoms and disability, we hypothesized that there would be a difference in the mean scores on the knee-specific instruments for patients who had different patient-rated severity scores as well as for those who had different clinician-rated severity scores; we thought that the instruments would differentiate between at least two of the levels of severity (determined by analysis of variance and the Tukey post hoc honestly significant difference test).

6. Since we included a broad range of diagnoses of varying severity, we hypothesized that there would be no ceiling or floor effects. Ceiling and floor effects have been defined as one-third of the patients receiving the highest or lowest possible score, respectively. For greater sensitivity, we defined ceiling and floor effects as one-third of the patients receiving the highest or lowest 10% of the possible scores.

**Responsiveness**

Patients who were expected to have improvement because of the nature of the diagnosis and the proposed treatment were entered into the responsiveness arm of the study. These patients all had conditions that are known to be successfully treatable in the majority of cases. They were reassessed at a minimum of three months following the initial evaluation. Patients who underwent a reconstructive procedure had a follow-up evaluation at a minimum of six months. Different durations of follow-up were needed because of the different treatments involved. For example, reconstruction of the anterior cruciate ligament requires a longer follow-up to achieve a change in health status than arthroscopic meniscectomy does.

In order to be able to detect that a true difference had occurred, patients were asked to rate their condition as “much worse,” “somewhat worse,” “a little worse,” “no change,” “a little better,” “somewhat better,” and “much better.” Patients who responded “a little better,” “somewhat better,” or “much better” were included in the study of the responsiveness of the instruments, while patients who responded that they were the same or worse were excluded from the responsiveness testing. We included only patients who stated that their condition had improved in order to be certain that the improvement that we believed we were measuring had occurred.

Many statistics are available to determine responsiveness. We elected to use the standardized response mean (the observed change divided by the standard deviation of change) because it has been used widely in previous orthopaedic research and it incorporates the response variance, allowing statistical testing of the response means. Standardized response means for validated orthopaedic instruments have ranged from 0.9 to 1.9.

**The Instruments**

The modified Lysholm scale, as described by Tegner and Lysholm, is an eight-item questionnaire that was originally designed to evaluate patients following knee ligament surgery. It is scored on a 100-point scale, with 25 points for knee stability, 25 points for pain, 15 points for locking, 10 points each for swelling and stair-climbing, and 5 points each for limp, use of a support, and squatting. This scale has been used extensively in clinical research studies.

The first version of the Cincinnati knee-rating system was published in 1983, with additional modifications developed for occupational activities, athletic activities, symptoms, and functional limitations in sports and daily activities. The system has eleven components, including physical examination, laxity of the knee based on instrumented testing, and radiographic evidence of degenerative joint disease. We eval-
uated the subjective component, which includes pain, swelling, and giving-way, as well as the activity-level component, as these two parts are most related to disability.

The Activities of Daily Living scale of the Knee Outcome Survey was developed recently and published with an evaluation of its reliability, validity, and responsiveness. This scale is designed for the evaluation of patients with disorders of the knee ranging from injury of the anterior cruciate ligament to arthrosis. It includes seventeen multiple-choice questions divided into two sections: one for symptoms (seven questions) and one for functional disability (ten questions).

The American Academy of Orthopaedic Surgeons sports knee-rating scale was included in the Musculoskeletal Outcomes Data Evaluation and Management System (MODEMS) for athletic patients with disorders of the knee. This instrument has five parts with a total of twenty-three questions: a core section (seven questions) on stiffness, swelling, pain, and function, and four sections (four questions each) on locking or catching on activity, giving-way on activity, current activity limitations due to the knee, and pain on activity due to the knee. At the present time, we are not aware of any published evidence of the reliability, validity, or responsiveness of this instrument.

Three of the instruments are 100-point scales. The Cincinnati knee-rating system is a 35-point scale that we converted to a 100-point scale, by dividing the score by thirty-five and then multiplying it by 100, to facilitate comparisons.

Data Management and Analysis
The four questionnaires were collated in random order before they were presented to each patient in order to avoid a potential bias due to the sequence in which they were completed. Response forms were completed by the patients, and data entry was accomplished by manually scanning the forms. If the response was not readable by the scanner, the data were entered manually. All analyses were carried out on SPSS software (version 9.0; SPSS Advanced Statistics, Chicago, Illinois) for personal computers.

As already noted, the scoring system for the American Academy of Orthopaedic Surgeons sports knee instrument has five values, or subscales. As it was impractical to use five values for each patient for the analysis, we used an unweighted mean of the five subscales to calculate an overall score on this instrument. We calculated the score for a given subscale if one-half or more of the items in that subscale could be scored. If it was possible to calculate a score for three or more of the subscales, we calculated the mean of the available scores. The Lysholm, Cincinnati, and Activities of Daily Living scales were scored as recommended by the originators of each system.

Sample Size
Sample-size calculation indicated that for \( \alpha = 0.05 \), \( \beta = 0.20 \), \( \rho_{\text{on}} = 0.60 \), and \( \rho_{\text{off}} = 0.85 \), a sample size of forty-two patients was required for the reliability study. As far as we know, there are no studies that describe the sample size for a responsiveness study, but previous authors have used the same number of patients as those used for reliability studies. The baseline questionnaires from both groups of patients were used for the validity analyses, which guaranteed a minimum of eighty patients.

Patient Demographics
Forty-one patients were included in the reliability arm of the study. Twenty of the patients were male, and twenty-one were female. The mean age was 32.6 years (range, fifteen to sixty years). The mean Tegener score was 6.3 points (range, 4 to 10 points). The diagnoses included injury of the anterior cruciate ligament in twenty-eight patients; osteochondritis dissecans in three; arthrosis, a meniscal tear, and patellofemoral joint pain in two each; and patellar tendinitis, injury of the posterior cruciate ligament, knee dislocation, and patellar tendon rupture in one each. The mean time between completion of the baseline questionnaire and the follow-up questionnaire was 5.2 days (range, two to fourteen days).

With the baseline responses from the reliability and responsiveness analyses, a total of 133 patients were included in the validity analysis. There were sixty-nine males and sixty-four females. The mean age was 31.5 years (range, fourteen to sixty-five years). The mean Tegener score was 6.4 points (range, 4 to 10 points). The diagnoses included injury of the anterior cruciate ligament in fifty-seven patients; a patellofemoral disorder in twenty-one; a meniscal tear in seventeen; arthrosis in thirteen; injury of the medial collateral ligament in five; osteochondritis dissecans in four; patellar tendinitis and a patellar tendon ossicle in three each; injury of the posterior cruciate ligament and knee dislocation in two each; and Osgood-Schlatter disease, patellar tendon rupture, symptomatic plica, chondral defect, iliotibial band tendinitis, and quadriceps tendon injury in one each.

Forty-two patients were involved in the responsiveness arm of the study. The mean age was 30.9 years (range, fifteen to sixty-one years). Nineteen of the patients were male, and twenty-three were female. The mean Tegener score was 6.5 points (range, 4 to 9 points). The diagnoses were varied and included a disorder of the patellofemoral joint in fifteen patients; injury of the anterior cruciate ligament in twelve; arthrosis and a meniscal tear in six each; and Osgood-Schlatter disease, injury of the medial collateral ligament, and a patellar tendon ossicle in one each. Twenty-four patients had nonoperative treatment, nine had reconstruction of the anterior cruciate ligament, four had meniscectomy, two had Synvisc (hylan G-F 20) injection, and one each had arthroscopic débridement, meniscal repair, and microfracture.

Results
Reliability Results
The mean baseline scores were 71.6 points for the Cincinnati knee-rating system, 84.1 points for the Lysholm scale, and 85.1 points for the Activities of Daily Living scale. The mean scores for the American Academy of Orthopaedic Surgeons sports knee-rating scale and the Knee Outcome Scales for Athletic Patients were 71.1 points for the Cincinnati knee-rating scale.
system, 84.0 points for the Lysholm scale, 86.1 points for the Activities of Daily Living scale, and 85.9 points for the American Academy of Orthopaedic Surgeons sports knee-rating scale.

The intraclass correlation coefficient for these scales was 0.88 for the Cincinnati knee-rating system, 0.95 for the Lysholm scale, 0.93 for the Activities of Daily Living scale, and 0.92 for the American Academy of Orthopaedic Surgeons sports knee-rating scale. The limits of agreement (mean difference and two standard deviations between the tests) were 3.8 ± 8.0 for the Lysholm scale, 7.8 ± 22.5 for the Cincinnati knee-rating scale, 3.5 ± 9.9 for the Activities of Daily Living scale, and 4.1 ± 9.3 for the American Academy of Orthopaedic Surgeons sports knee-rating scale.

**Validity Results**

The orthopaedic surgeons all considered the scales to have face and content validity. All patients who completed the questionnaires at baseline were included in the construct validity portion of the study since this was a cross-sectional analysis. There were three scenarios in which the baseline responses were used for the validity testing but not for the reliability or responsiveness analysis. Baseline responses of patients who were initially entered into the reliability or responsiveness arm of the study but who did not believe that the status of the knee had improved on the transitional index at follow-up were excluded from the responsiveness study although their baseline responses were used for the validity study. The mean scores for these patients are listed in Table I.

With regard to validity testing, the first and second hypotheses were confirmed as the knee-specific scales all correlated better with each other than they did with the physical component scale, the mental component scale, or any of the SF-36 subscales. As expected, all scales correlated better with the physical component scale than they did with the mental component scale (Table II).

The third hypothesis consisted of two parts. The first was that the physical function and role-physical subscales would correlate better with each knee scale than would the vitality or social function subscales, and the second was that the subscales for general health, bodily pain, vitality, and social function would all correlate better with each knee scale than would the role-emotional and mental health subscales. There were a small number of minor discrepancies from these constructs (Table III).

The fourth hypothesis was confirmed as all knee-rating scales correlated better with each other than they did with the physical component and mental component scales, and any of the SF-36 subscales. The mean scores for these patients are listed in Table I.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Mean Score</th>
<th>Standard Deviation</th>
<th>Lowest Score</th>
<th>Highest Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities of Daily Living scale</td>
<td>76.0</td>
<td>18.3</td>
<td>16.3</td>
<td>100.0</td>
</tr>
<tr>
<td>American Academy of Orthopaedic Surgeons sports knee-rating scale</td>
<td>73.5</td>
<td>19.0</td>
<td>26.9</td>
<td>98.3</td>
</tr>
<tr>
<td>Cincinnati knee-rating scale</td>
<td>53.2</td>
<td>28.7</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Lysholm scale</td>
<td>74.6</td>
<td>18.7</td>
<td>15.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Tegner scale</td>
<td>6.4</td>
<td>1.6</td>
<td>4.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subscales</th>
<th>Mean Score</th>
<th>Standard Deviation</th>
<th>Lowest Score</th>
<th>Highest Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical function</td>
<td>74.6</td>
<td>26.1</td>
<td>5.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Role-physical</td>
<td>61.8</td>
<td>42.9</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>65.7</td>
<td>23.0</td>
<td>12.0</td>
<td>100.0</td>
</tr>
<tr>
<td>General health</td>
<td>84.2</td>
<td>14.4</td>
<td>42.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Vitality</td>
<td>64.4</td>
<td>19.6</td>
<td>5.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Social function</td>
<td>83.1</td>
<td>21.1</td>
<td>12.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Role-emotional</td>
<td>84.2</td>
<td>31.1</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Mental health</td>
<td>76.4</td>
<td>14.2</td>
<td>24.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Physical component scale</td>
<td>46.1</td>
<td>10.2</td>
<td>20.3</td>
<td>61.0</td>
</tr>
<tr>
<td>Mental component scale</td>
<td>53.1</td>
<td>8.7</td>
<td>17.7</td>
<td>67.8</td>
</tr>
</tbody>
</table>
scales correlated well with both clinician and patient ratings of severity (Table II). The minimum correlation between either of these constructs and one of the knee-rating scales was 0.61, and all correlations were significant (p < 0.01) (Table II). Patient-rated severity is more relevant, and all scales correlated better with patient-rated severity than they did with clinician-rated severity. Analysis of variance demonstrated a significant difference, with regard to the scores on each knee-rating scale, among patients with different clinician-rated and patient-rated severity (p < 0.00000001) (Table IV and Figs. 1 through 4). Post hoc testing demonstrated that the scores on the Activities of Daily Living scale, the American Academy of Orthopaedic Surgeons sports knee-rating scale, and the Cincinnati knee-rating system differed significantly between two adjacent levels of clinician-rated severity, but the scores on the Lysholm scale did not. All four scales demonstrated significant differences between two or more adjacent levels of patient-graded severity.

No ceiling or floor effects were demonstrated for any of the scales.

| TABLE II Validity Analysis with Spearman Correlation Matrix for the Four Knee-Rating Scales, the Physical and Mental Component Scales of the Short Form-36 (SF-36), and the Clinician and Patient Ratings of Severity |
|--------------------------------------------------|--------------------------------------------------|------------------|-----------------|---------------------|---------------------|---------------------|
| Physical component scale of SF-36 | — | 0.67† | 0.68† | 0.68† | 0.77† | 0.64† | 0.5† |
| Mental component scale of SF-36 | — | 0.03 | 0.05 | 0.01 | 0.05 | 0.08 | 0.01 |
| American Academy of Orthopaedic Surgeons sports knee-rating scale | 0.67† | — | 0.70† | 0.83† | 0.80† | 0.74† | 0.64† |
| Lysholm scale | 0.68† | 0.05 | 0.70† | — | 0.70† | 0.85† | 0.65† | 0.62† |
| Cincinnati knee-rating scale | 0.68† | 0.01 | 0.83† | 0.70† | — | 0.80† | 0.67† | 0.61† |
| Activities of Daily Living scale | 0.77† | — | 0.80† | 0.85† | 0.80† | — | 0.73† | 0.68† |
| Patient rating of severity | 0.64† | 0.08 | 0.74† | 0.65† | 0.67† | 0.73† | — | 0.58† |
| Clinician rating of severity | 0.59† | 0.01 | 0.64† | 0.62† | 0.61† | 0.68† | 0.58† | — |

*Correlations were significant at p < 0.05. †Correlations were significant at p < 0.01.

| TABLE III Validity Analysis with Spearman Correlations for the Four Knee-Rating Scales and Short Form-36 Subscales |
|--------------------------------------------------|------------------|-----------------|-----------------|---------------------|
| Short Form-36 Subscale | Cincinnati Knee-Rating Scale | Lysholm Scale | Activities of Daily Living Scale | American Academy of Orthopaedic Surgeons Sports Knee-Rating Scale |
| Physical function | 0.68* | 0.66* | 0.72* | 0.66* |
| Role-physical | 0.52* | 0.49* | 0.60* | 0.62* |
| Bodily pain | 0.51* | 0.57* | 0.54* | 0.44* |
| General health | 0.26* | 0.31* | 0.28* | 0.17 |
| Vitality | 0.23* | 0.28* | 0.26* | 0.19† |
| Social function | 0.50* | 0.50* | 0.48* | 0.48* |
| Role-emotional | 0.18† | 0.18† | 0.14 | 0.24* |
| Mental health | 0.24* | 0.29* | 0.21† | 0.19† |

*Correlations were significant at p < 0.01. †Correlations were significant at p < 0.05.
Responsiveness Results
The mean scores improved from 37.9 points at baseline to 65.0 points at the time of follow-up for the Cincinnati knee-rating system, from 64.5 points to 79.6 points for the Lysholm scale, from 66.1 points to 83.6 points for the Activities of Daily Living scale, and from 60.6 points to 81.8 points for the American Academy of Orthopaedic Surgeons sports knee-rating scale. The standardized response means were 0.8 for the Cincinnati knee-rating system, 0.9 for the Lysholm scale, 1.1 for the Activities of Daily Living scale, and 1.0 for the American Academy of Orthopaedic Surgeons sports knee-rating scale. While the American Academy of Orthopaedic Surgeons sports knee-rating scale was quite responsive, six patients had questionnaires that were not able to be scored, leaving only thirty-six patients with questionnaires that could be scored at both baseline and follow-up.

Discussion
The results of a clinical research study are of questionable value if the measure used to evaluate the effectiveness of treatment is not known to be reliable, valid, and responsive. In many clinical research studies, questionnaires are used as primary outcome measures because these instruments accurately reflect symptoms and disabilities that are specific and important to patients.

Anderson et al. compared six knee-ligament rating scales in a study of seventy patients who had had reconstruction of the anterior cruciate ligament five years earlier. They concluded that the International Knee Documentation Committee (IKDC) scale should be used to standardize measurements. However, the authors did not present any data on the reliability, validity, or responsiveness of this scale. In another study, the scores of eight knee-rating scales (including the Lysholm, Hospital for Special Surgery, and IKDC scales) were compared in a group of fifty-six patients who had undergone reconstruction of the anterior cruciate ligament. The authors encouraged the use of the IKDC scale; however, the measurement properties of these scales were not determined.

Other investigators have compared the Lysholm scale and Cincinnati knee-rating system in studies of patients who either had an insufficient anterior cruciate ligament or had reconstruction of the anterior cruciate ligament. They found that patients had higher scores on the Lysholm scale but that there was a linear correlation between the two instruments. These studies did not assess the reliability or responsiveness of the instruments.

Many scales have been developed without patient input or the formal techniques of item generation and item reduction. In addition, scales have been developed for a wide variety of purposes and for specific patient populations. As a result, different scales may be used for similar studies, which could be a cause for differing conclusions.

We evaluated the reliability, validity, and responsiveness to clinical change of four questionnaires that assess disability and symptoms in active patients with disorders of the knee. All of the scales were found to have excellent reliability, with the intraclass correlation coefficient ranging from a low of 0.88 for the Cincinnati knee-rating system to a high of 0.95 for the Lysholm scale. (A coefficient of >0.75 is adequate for pa-
The mean scores on the American Academy of Orthopaedic Surgeons (AAOS) sports knee-rating scale according to the clinician-rated severity. The error bars indicate the standard deviation.

Fig. 2

Clinician Severity Rating

The mean scores on the Cincinnati knee-rating system according to the clinician-rated severity. The error bars indicate the standard deviation.

Fig. 3

patients enrolled in a clinical trial. The limits-of-agreement statistic is a measure of reliability that provides additional information to the intraclass correlation coefficient. This statistic is the mean difference (and two standard deviations) between two measures used to evaluate the same subject. The mean difference among the four scales was extremely low (range, 3.5 to 7.8). Three of the four scales had a 95% confidence interval between 8.0 and 9.9, while the Cincinnati knee-rating system had a confidence interval of 22.5, indicating increased measurement variability or decreased reliability.

The knee scales, the physical component scale and mental component scale of the SF-36, and the patient and clinician severity ratings were used as constructs to evaluate validity. All of the scales were thought to have adequate face and content validity by ten orthopaedic surgeons with experience in the field of sports medicine. Our six hypotheses regarding construct validity were confirmed. Such confirmation is important because the use of joint-specific scales could be questioned if the knee-specific instruments did not correlate with each other to a greater degree than they did with the physical component scale, the mental component scale, or the SF-36 subscales.

Responsiveness is dependent not only on the instrument but also on the magnitude of change actually experienced by the patients. The magnitude of change measured in a cohort of patients is determined by the initial score (lower scores allow more room for improvement), the quality of the intervention, the instrument used to measure the health status, and the statistic used to calculate responsiveness. The responsiveness of measurement scales in orthopaedics has been measured with use of the standardized response mean. The activities of daily living and symptoms subscales of the Cincinnati knee-rating system had standardized response means of 0.72 and 1.56, respectively, in a study of patients who had undergone anterior cruciate ligament reconstruction. Two generic health-status instruments had standardized response means of 0.88 and 1.00 in a study of patients who had undergone hip or knee replacement surgery. The standardized response means for the knee-specific questionnaires in the present study ranged from 0.8 for the Cincinnati knee-rating system to 1.1 for the Activities of Daily Living scale. These values are fairly impressive considering that not all patients underwent surgery and that they were generally not severely disabled at baseline according to their diagnoses and questionnaire scores. Therefore, we concluded that these instruments were all capable of detecting a clinically relevant difference over time.

The standard deviation of the measure is important for calculating sample size when designing a study to compare two treatments with use of a rating scale as the primary outcome measure. Standard deviations can be compared in cases where the minimum and maximum scores are the same. In the present study, the standard deviations ranged from 18.3 to 19.0 for three of the scales, while the Cincinnati knee-rating system had a standard deviation of 28.7, again indicating increased measurement variability for the scale. The standard deviations are relatively large, which is possibly due to the heterogeneity of the baseline population. Other instruments have also demonstrated standard deviations in this range when tested in studies of patients with a variety of diagnoses of varying severity. In the initial report that described the Activities of Daily Living scale, a standard deviation of 20.8 was found for the baseline responses of patients with diagnoses that ranged from tendinitis to osteoarthritis. Conversely, in a study of a
The mean Activities of Daily Living (ADL) scores according to the clinician-rated severity.

The American Academy of Orthopaedic Surgeons (AAOS) knee-rating scale. While all of the scales were well under the realm of acceptable responder burden, the number of items is important, particularly if the questionnaires are administered in conjunction with a generic health-status measure.

The scoring was relatively straightforward for three of the scales; however, the scoring manual provided for the AAOS knee-rating scale suggests the calculation of five subscales. While this information is valuable from a clinical perspective, it is too complicated to have five subscales describing each aspect of knee function for each patient in a research initiative. In addition, this scale was the only one of the four to have the response “cannot do for other reasons.” The scoring manual states that this item should be “dropped,” which we interpreted as “scored as missing.” We elected to calculate the mean of the subscales (to arrive at an overall score for the instrument) if it was possible to score three or more of the subscales. The fact that this scale had more items and more complicated scoring and that more questionnaires could not be scored because of missing responses makes its use somewhat more onerous. However, when patients completed the questionnaire sufficiently so that a score could be calculated, the instrument was reliable, valid, and responsive according to our criteria.

One limitation of the present study is that the results are not generalizable to patients who are not active in sports (that is, those who have a Tegner rating of <4 points). Another potential limitation is that we chose to measure a variety of disorders of the knee. We did so to allow generalizability of the results to a wide variety of patients and treatments. In effect, our cohort of active patients with disorders of the knee is a relatively homogeneous group from the perspective of general orthopaedic and medical health. While it is possible that a given scale would perform differently in a group of patients with a single diagnosis, it is unlikely that this discrepancy would be very large. Lastly, we used only the subjective components of the Cincinnati knee-rating system, and our results do not apply to the other components of this system, which mainly measure impairments.

The Activities of Daily Living scale was well understood by patients, could be completed in a relatively short time-period, and had slightly better construct validity and responsiveness than the other scales. This finding is possibly due to the clear wording of the instrument or to the fact that it evaluates a very wide variety of symptoms and disabilities compared with the others. The latter allows an investigator to use this instrument for studies involving various knee diagnoses, as was its intended purpose. We recommend this instrument for the study of disorders of the knee in athletic patients.

The AAOS knee-rating scales also satisfied our criteria for reliability, validity, and responsiveness, and all are acceptable for use in clinical research. The four scales have many areas of overlap, and the development of statistical methods to compare the results from one scale with those from another is an important area for future research. Additional work to evaluate some of the newer, well-designed knee-specific tools, such as the quality-of-life outcome measure for chronic anterior cruciate ligament deficiency and the Knee Injury and Osteoarthritis Outcome Score (KOOS), is required.

Robert G. Marx, MD, MSc, FRCS(C)
Edward C. Jones, MD, MA
Answorth A. Allen, MD
David W. Altchek, MD
Stephen J. O’Brien, MD
Scott A. Rodeo, MD
Riley J. Williams, MD
Russell F. Warren, MD
Thomas L. Wickiewicz, MD
marxr@hss.edu

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. No funds were received in support of this study. Dr. Marx was supported by an American Academy of Orthopaedic Surgeons Health Services Research Fellowship and a Royal College of Physicians and Surgeons of Canada Detweiler Travelling Fellowship.

References


Downloaded from www.ejbjs.org on February 25, 2005


