

A comparison of two time intervals for test-retest reliability of health status instruments

Robert G. Marx^{a,b,*}, Alia Menezes^b, Lois Horovitz^a, Edward C. Jones^b, Russell F. Warren^a

^a*Sports Medicine and Shoulder Service Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA*

^b*Center for Clinical Outcome Research Hospital for Special Surgery, New York, NY, USA*

Accepted 23 March 2003

Abstract

Studies of test-retest reliability for health-related quality of life instruments have used varying intervals between test administrations. There is no evidence available to aid in the selection of the time interval between questionnaire administrations for a study of test-retest reliability for health status instruments. We compared the test-retest reliability at 2 days and 2 weeks for four knee-rating scales and the eight domains of the SF-36. Seventy patients with disorders of the knee who were in a stable state were randomly allocated to repeat the questionnaires at either 2 days or 2 weeks. There were no statistically significant differences in the test-retest reliability (intraclass correlation coefficient and limits of agreement statistics) for the two time intervals. © 2003 Elsevier Inc. All rights reserved.

Keywords: Reliability; Measurement; Knee; Retest; Scale; Instrument

1. Introduction

Reliability is a critical measurement property for health-related quality of life (QOL) instruments. Reliability refers to the consistency of scores obtained by the same persons when re-examined with the same test on different occasions or with different sets of equivalent sets of items [1]. There are many techniques available to measure reliability, including internal consistency and test-retest reliability.

An instrument that has adequate test-retest reliability gives the same result if an individual is re-tested while remaining in a clinical steady state. The problem with testing reliability by the test-retest method is that there is a potential for learning, carry-over, or recall effects (i.e., the first testing may influence the second) [2]. The length of time between the two test administrations also affects the test-retest reliability. A very short time interval makes the carryover effects due to memory, practice, or mood more likely, whereas a longer interval increases the chances that a change in status could occur [2]. Measuring reliability by the internal consistency method involves dividing the instrument into two equal parts and comparing the score on both halves (i.e., split-half reliability). The Kuder Richardson formula 20 is the average of all of the split-half reliabilities, and

Cronbach's α is an extension of this formula for ordinal data [3].

Test-retest reliability is more relevant in the setting of clinical medicine because the constructs we attempt to measure are heterogeneous. For example, many instruments used by physicians combine apparently diverse domains (i.e., heterogeneous) such as symptoms (e.g., pain, numbness) and disability (e.g., limited mobility, difficulty with activities of daily living) into a single score. A homogenous instrument consists of items that relate to a single domain. Thus, one may expect clinically heterogeneous scales to have poor internal consistency. However, despite the heterogeneity of variables such as health-related QOL, there is evidence that the latter fulfill the criteria for internal consistency despite their apparent heterogeneity [4].

Studies of test-retest reliability for health-related QOL instruments have used varying intervals between test administrations. The interval has ranged from 10 minutes to 1 month [5–13]. Most investigators have chosen an interval ranging from 2 days to 2 weeks. This time frame is generally believed to be a reasonable compromise between recollection bias and unwanted (on the part of the investigator) clinical change.

There is no evidence available to aid in the selection of the time interval between questionnaire administration for a study of test-retest reliability for health status instruments. The goal of this study was to prospectively compare the 2-day and 2-week test-retest reliability of four knee-rating

* Corresponding author. Tel.: 212-606-1866.

E-mail address: marxr@hss.edu (R.G. Marx).

scales and the eight subscales of the SF-36 in a cohort of athletic patients with disorders of the knee.

2. Methods

Patients were recruited in the waiting room of orthopedic surgeons specializing in disorders of the knee or by telephone. The latter group was identified by the research nurse (LH) in a single orthopedic practice (RFW). Patients who were believed to be in a clinically stable state were asked to participate in the study. These individuals were randomly assigned to be re-tested at 2 days or 2 weeks [3,14]. Randomization was performed using blocks of four from a random number generator.

A wide variety of diagnoses was sought to test a broad spectrum of severity and to ensure generalizability to the various conditions that affect the knee in athletic patients. The Lysholm scale [15,16], the American Academy of Orthopaedic Surgeons sports knee rating scale (AAOSSN) [17], the Activity of Daily Living of the Knee Outcome Survey (ADL) [18], the Cincinnati knee rating system [19], and the eight subscales of the SF-36 were studied. The Tegner scale rates patients from 0 to 10 based on their activity level and sports participation [16]. Patients not participating in high-demand sports (Tegner activity rating of 3 or less, indicating that they do not run or participate in sport, with the exception of swimming) were excluded [16,20]. We did not exclude patients based on age alone, although patients with Tegner scores of 4 or greater tended to represent the relatively young age group of patients in most orthopedic sports medicine practices.

The Modified Lysholm scale, as presented by Tegner and Lysholm, is an eight-item questionnaire that was originally designed to evaluate patients after knee ligament surgery [15]. It is scored from 0 to 100 points, based on eight items using a Likert scale [16]. This scale has been used extensively for clinical research studies [21–24].

The Cincinnati Knee Rating System includes 11 components that measure impairments and disabilities [25]. We evaluated the “subjective component” (which includes pain, swelling, and giving-way) and the activity level component because these two parts are most related to disability.

The ADL scale is designed for patients with disorders of the knee ranging from sports-related injury to osteoarthritis [18]. It includes 17 multiple-choice questions divided into two sections—one for symptoms (seven questions) and one for functional disability (10 questions).

The AAOSSN rating scale [17] is comprised of five parts and 23 questions. There is a core section that includes stiffness, swelling, pain, and function (seven questions); a locking or catching on activity section (four questions); a giving way on activity section (four questions); a current activity limitations due to the knee section (four questions); and a pain on activity due to the knee section (four questions).

Three of these scales are 100-point scales. The score for the Cincinnati scale (out of 35) was converted to a score out of 100 (i.e., divided by 35 and then multiplied by 100) to facilitate comparisons.

We also administered version 1.0 of the SF-36, which is a 36-item questionnaire that measures general health [26–28]. Its use has been encouraged in conjunction with knee-specific instruments for studies of patients with knee injuries [24]. Each of the eight subscales was evaluated separately for reliability.

Patients were included if they had a primary disorder of the knee, including patellofemoral disorders (including chondromalacia, patellar dislocation, and patellar tendinitis), knee instability (acute or chronic ligament instability), meniscal injury, or osteochondritis dissecans. Patients were excluded if they were unable to read and write in English. Patients with inflammatory joint disease, tumors, or infection of the knee were excluded. Patients with Tegner Activity Rating Scale of 3 or less (before the injury if there was recent trauma) were excluded. Patients who had had surgical intervention or traumatic injury in the preceding 3 months were excluded.

The patients were believed to be in a stable state and were not expected to change before completing the questionnaires for the second time. Patients did not receive treatment in the interval between the first and second completions of the questionnaires. When the patients completed the questionnaires for the second time, they also completed a transitional index asking them to rate the severity of their knee condition compared with when they completed the questionnaires the first time [14]. They chose from the following 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 to ensure that the clinical status of the patients was stable. Patients were provided with self-addressed stamped envelopes to return the second group of questionnaires. A follow-up phone call was made to remind the patients to complete the second group of questionnaires. The second questionnaire was excluded from the study if it was received by mail more than 1 week after the scheduled date of completion.

The intraclass correlation coefficient (ICC) and the limits of agreement statistic [29–31] were used to compare the scores [14]. The ICC is an index of concordance for dimensional measurements ranging between 0 and 1, where ≥ 0.75 is considered excellent reliability [32]. The limits of agreement statistic was also used as a descriptive measure of agreement. This value is the mean difference between the two tests ± 2 standard deviations (SDs) [30]. Ninety-five percent of the differences between the two test administrations lies within this interval [30]. A difference of 7 points between the limits of agreement statistics for the two groups was believed a priori to indicate clinical significance. The limits of agreement statistic for the 2-day and the 2-week groups were compared with an unpaired *t* test for each

knee-specific instrument and for the eight subscales of the SF-36.

For sample size calculations, the clinically significant difference was 7, and the SD for the knee scales was estimated to be 10, based on previous research [33]. The clinically significant difference was used to estimate a clinically relevant difference between the limits of agreement statistic for the two groups. All scales had a minimum score of 0 and a maximum score of 100. For an α of 0.05 and β of 0.20, 33 subjects were required per group.

If the reliability were found to be higher at 2 days compared with 2 weeks, it would not be possible to determine whether the difference would be due to recall in the 2-day group or clinical change in the 2-week group. These two potential effects could be discounted only if reliability was equal or unexpectedly greater in the 2-week interval group.

3. Results

Of the 108 patients who completed the baseline questionnaire, 82 completed the second. Of these, 70 replied that the status of their knee was unchanged on the transitional index (38 in the 2-day group and 32 in the 2-week group). Five patients in the 2-day group reported change (four stated they were improved, and one stated they were worse). Seven patients in the 2-week group reported change (six stated they were improved, and one stated they were worse).

The mean age of the patients in the 2-day group was 31.8 years (range 15–61 years). There were 17 men and 21 women. The mean Tegner rating was 6.2 (range 4–10). The diagnoses included 26 patients who had sustained anterior cruciate ligament (ACL) injury; two patients with meniscal tears; two with patellofemoral pain; two with osteochondritis dissecans; and one each with osteoarthritis, patellar tendonitis, posterior cruciate ligament rupture, and patellar tendon rupture.

The mean age of the patients in the 2-week group was 38.1 years (range 15–61 years). There were 18 men and 14 women. The mean Tegner rating was 5.8 (range 4–9). The diagnoses included 25 patients who had sustained ACL injury; two patients with meniscal tears; and one each with osteoarthritis, patellofemoral pain, and knee dislocation. There were a large number of patients with ACL injuries in both groups because it is a common condition, the patients are readily identifiable by office records, and the patients are often in a stable state once they recover from the acute phase of injury. Mean baseline scores for the two groups were similar (Table 1).

Twelve patients reported a change in their clinical status. Seven of these patients were in the 2-week group, and five were in the 2-day group. These patients were not included in the analysis.

The ICCs for all four knee scales for both groups were ≥ 0.85 , indicating excellent reliability (Table 2). The ICCs for the subscales of the SF-36 ranged from 0.59 (Role Emotional, 2-day group) to 0.93 (Physical Function, 2-week

Table 1
Mean baseline score \pm SD for both groups

Scale	2-day group	2-week group
Physical Functioning domain of the SF-36	90.7 \pm 17.1	89.4 \pm 15.0
Role Physical domain of the SF-36	88.2 \pm 27.7	93.8 \pm 24.6
Bodily Pain domain of the SF-36	82.5 \pm 22.6	83.3 \pm 19.8
General Health domain of the SF-36	86.3 \pm 16.0	84.9 \pm 17.0
Vitality domain of the SF-36	69.6 \pm 16.2	72.2 \pm 19.9
Social Functioning domain of the SF-36	93.8 \pm 15.9	91.4 \pm 16.3
Role Emotional domain of the SF-36	93.0 \pm 23.5	90.6 \pm 22.8
Mental Health domain of the SF-36	79.7 \pm 11.8	81.3 \pm 15.8
Cincinnati	77.8 \pm 24.8	79.8 \pm 22.3
Lysholm	87.6 \pm 14.1	87.7 \pm 15.1
ADL	89.0 \pm 13.4	88.9 \pm 11.1
AAOSSN	87.4 \pm 14.4	90.8 \pm 10.8

Abbreviations: SD, standard deviation; ADL, Activity of Daily Living of the Knee Outcome Survey; AAOSSN, American Academy of Orthopaedic Surgeons sports knee rating scale.

group). The 95% confidence intervals (CIs) for the ICCs of the 2-day and 2-week groups overlapped for seven of the eight subscales, with the exception of the Physical Function subscale of the SF-36 (Table 2). Therefore, with the exception of the one subscale, the ICCs were not significantly different.

The limits of agreement statistics for the 2-day and 2-week groups were similar for the knee scales and for the subscales of the SF-36 (Table 3, Fig. 1). None of the scales demonstrated a limits of agreement difference of 7 points, which was the a priori limit for clinical significance. The largest discrepancy between the mean difference for the 2-day and 2-week groups was for the Vitality subscale of the SF-36 (Table 3). This difference was statistically significant, although with Bonferroni's correction for multiple comparisons, the P value required for statistical significance ($P = .05/12 = .0042$) was not achieved.

Table 2
Intra-class correlation coefficient (95% CI) for both groups

Scale	2-day group	2-week group
Physical Functioning domain of the SF-36	0.68 (0.46–0.82)	0.93 (0.87–0.97)
Role Physical domain of the SF-36	0.87 (0.76–0.93)	0.72 (0.50–0.85)
Bodily Pain domain of the SF-36	0.84 (0.71–0.91)	0.84 (0.69–0.92)
General Health domain of the SF-36	0.64 (0.41–0.79)	0.72 (0.50–0.85)
Vitality domain of the SF-36	0.86 (0.75–0.92)	0.63 (0.37–0.80)
Social Functioning domain of the SF-36	0.78 (0.62–0.88)	0.80 (0.64–0.90)
Role Emotional domain of the SF-36	0.59 (0.35–0.77)	0.64 (0.37–0.81)
Mental Health domain of the SF-36	0.76 (0.59–0.87)	0.60 (0.33–0.78)
Cincinnati	0.85 (0.72–0.92)	0.94 (0.87–0.97)
Lysholm	0.89 (0.80–0.94)	0.90 (0.80–0.95)
ADL	0.93 (0.87–0.97)	0.93 (0.85–0.97)
AAOSSN	0.92 (0.85–0.96)	0.93 (0.86–0.97)

Abbreviations: CI, confidence interval; ADL, Activity of Daily Living of the Knee Outcome Survey; AAOSSN, American Academy of Orthopaedic Surgeons sports knee rating scale.

Table 3
Limits of agreement statistics for both groups

Scale	2-day group	2-week group	P value
Physical Functioning domain of the SF-36	5.8 ± 31.0	3.4 ± 9.8	0.38
Role Physical domain of the SF-36	5.3 ± 28.8	6.5 ± 40.8	0.79
Bodily Pain domain of the SF-36	7.0 ± 21.1	6.3 ± 17.6	0.76
General Health domain of the SF-36	6.3 ± 25.4	7.2 ± 18.6	0.72
Vitality domain of the SF-36	6.8 ± 12.8	12.0 ± 21.1	0.02
Social Functioning domain of the SF-36	5.3 ± 22.2	5.5 ± 17.9	0.93
Role Emotional domain of the SF-36	7.9 ± 39.3	8.6 ± 38.4	0.88
Mental Health domain of the SF-36	6.3 ± 14.1	8.3 ± 20.2	0.37
Cincinnati	7.6 ± 24.0	5.3 ± 12.1	0.33
Lysholm	3.5 ± 11.4	4.5 ± 11.4	0.46
ADL	2.8 ± 8.8	2.7 ± 7.8	0.95
AAOSSN	3.4 ± 9.2	2.7 ± 6.2	0.47

Abbreviations: ADL, Activity of Daily Living of the Knee Outcome Survey; AAOSSN, American Academy of Orthopaedic Surgeons sports knee rating scale.

4. Discussion

Four knee-specific instruments and one general health-related QOL questionnaire with eight domains were administered to patients to test reliability of the instruments. Patients were randomly assigned to complete the questionnaires at 2 days or 2 weeks. These time intervals were selected because in both cases it was believed that the intervals were too short for clinical change in patients who were believed to be in a stable state. With a minimum of 2 days between

questionnaire completions, sufficient time should have elapsed to minimize the bias associated with the recollection of previous responses.

Overall, the four knee scales and the eight domains of the SF-36 demonstrated excellent test-retest reliability. The clinically significant difference of 7 points occurred for 3 of the 12 scales in the 2-day group and 4 of 12 scales in the 2-week group. However, in no case was the difference in the limits of agreement statistic between the two groups equal to 7 (the greatest difference was 5.2; the next largest was 2.4), indicating that the two time intervals provided a similar measure of reliability. Also, the ICCs were similar for the two groups overall, with the 95% CIs overlapping in all cases except for one.

The number of items varies significantly among the knee-rating scales and the domains of the SF-36. With respect to the latter instrument, the number of items comprising the domains ranges from 2 to 10. The differences in reliability between the 2-day and 2-week groups tended to be greater for domains with a greater number of questions (Physical Function: 10 items; Mental Health: 5 items; Vitality: 4 items) and least for domains with only two items (Bodily Pain and Social Function). However, the number of items comprising the scale did not seem to affect the difference in test-retest reliability at 2 days compared with 2 weeks.

The practical implications of these results are that for studies of test-retest reliability of health status instruments, a 2-day or 2-week time frame between the administration of the two questionnaires does not affect the results of the reliability testing. A time interval ranging from 2 days to 2 weeks is adequate; the patients need not complete the second test on an exact date, but rather within this time frame.

There are limitations to this study. All patients involved were athletic and had disorders of the knee. They were

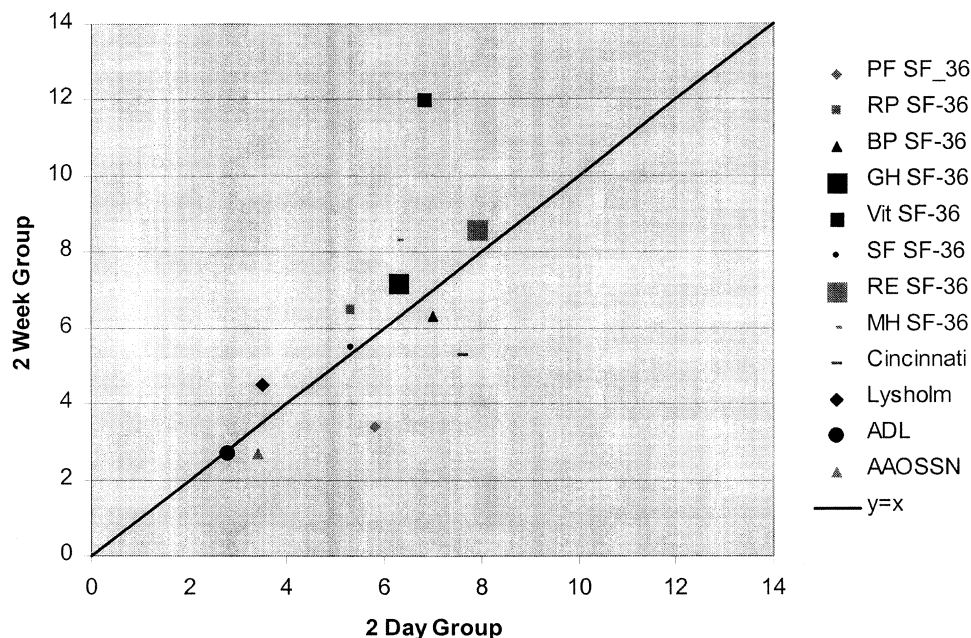


Fig. 1. Limits of agreement statistics for the 2-day and 2-week groups.

relatively young, and therefore the results may not be generalizable to other patient populations and health states. Seventy-six percent of patients who completed the first questionnaires also completed the second. Although phone calls were made to remind patients, the window in which the patients had to complete the second group of instruments was small, which led to a greater number of missing follow-ups. Therefore, the time between the completion of the questionnaires is not known with certainty because the questionnaires were mailed from the study center to the patients and then mailed back for the majority of the cases. However, this is generally a practical reality for other studies of reliability and therefore allows the generalizability of the results to test-retest reliability studies that are carried out using this methodology.

Patients who changed were not included in the analysis. It is possible that the reliability was artificially inflated by excluding patients who felt they had changed. However, the calculation of reliability of patients who believe that they are not in a stable state is not an accurate measure of repeatability because the outcome variable is not expected to be stable.

Multiple questionnaires were administered, each consisting of numerous items. Therefore, the effect of memory may have been minimized. The effect of memory may have been greater if a single questionnaire were used.

In conclusion, there was no clinically or statistically significant difference between the measurement of test-retest reliability performed with a 2-day interval as compared with a 2-week interval for our sample of athletic patients with disorders of the knee who were in a clinically stable state.

Acknowledgments

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

- [1] Anastasi A. Psychological testing. New York: Macmillan; 1988.
- [2] Allen MJ, Yen WM. Introduction to measurement theory. Monterey (CA): Brooks/Cole; 1979.
- [3] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 1989.
- [4] Marx RG, Bombardier C, Hogg-Johnson S, et al. Clinimetric and psychometric strategies for development of a health measurement scale. *J Clin Epidemiol* 1999;52:105–11.
- [5] Martin DP, Engelberg R, Agel J, et al. Comparison of the Musculoskeletal Function Assessment questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile health-status measures. *J Bone Joint Surg Am* 1997;79:1323–35.
- [6] Badia X, Alonso J. Validity and reproducibility of the Spanish Version of the Sickness Impact Profile. *J Clin Epidemiol* 1996;49:359–65.
- [7] Ferris LE, Shamian J, Tudiver F. The Toronto Breast Self-Examination Instrument (TBSEI): its development and reliability and validity data. *J Clin Epidemiol* 1991;44:1309–17.
- [8] Pollard WE, Bobbitt RA, Bergner M, et al. The Sickness Impact Profile: reliability of a health status measure. *Med Care* 1976; 14:146–55.
- [9] Whalen CC, Antani M, Carey J, et al. An index of symptoms for infection with human immunodeficiency virus: reliability and validity. *J Clin Epidemiol* 1994;47:537–46.
- [10] Folsom AR, Jacobs DR Jr., Caspersen CJ, et al. Test-retest reliability of the Minnesota Leisure Time Physical Activity Questionnaire. *J Chronic Dis* 1986;39:505–11.
- [11] Andresen EM, Bowley N, Rothenberg BM, et al. Test-retest performance of a mailed version of the Medical Outcomes Study 36-Item Short-Form Health Survey among older adults. *Med Care* 1996;34:1165–70.
- [12] Loeken K, Steine S, Sandvik L, et al. A new instrument to measure patient satisfaction with mammography: validity, reliability, and discriminatory power. *Med Care* 1997;35:731–41.
- [13] Gerace TA, Smith JC. Children's Type A interview: interrater, test-retest reliability, and interviewer effect. *J Chron Dis* 1985;18:781–91.
- [14] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Control Clin Trials* 1991;12(Suppl 4):142S–58S.
- [15] Lysholm J, Gillquist J. Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. *Am J Sports Med* 1982;10:150–4.
- [16] Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. *Clin Orthop* 1985;198:43–9.
- [17] Academy of Orthopaedic Surgeons. Scoring algorithms for the lower limb outcomes data collection instrument version 2.0. Rosemont (IL): 1998.
- [18] Irrgang JJ, Snyder-Mackler L, Wainner RS, et al. Development of a patient-reported measure of function of the knee. *J Bone Joint Surg Am* 1998;80:1132–45.
- [19] Noyes FR, Barber SD, Mooar LA. A rationale for assessing sports activity levels and limitations in knee disorders. *Clin Orthop* 1989;246:238–49.
- [20] 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;83:1459–69.
- [21] Gauffin H, Pettersson G, Tegner Y, et al. Function testing in patients with old rupture of the anterior cruciate ligament. *Int J Sports Med* 1990;11:73–7.
- [22] Odensten M, Hamberg P, Nordin M, et al. Surgical or conservative treatment of the acutely torn anterior cruciate ligament: a randomized study with short-term follow-up observations. *Clin Orthop* 1985; 198:87–93.
- [23] Roberts TS, Drez D Jr., McCarthy W, et al. Anterior cruciate ligament reconstruction using freeze-dried, ethylene oxide-sterilized, bone-patellar tendon-bone allografts: two year results in thirty-six patients. *Am J Sports Med* 1991;19:35–41. [published erratum appears in *Am J Sports Med*, 19 (1991) 272].
- [24] Shapiro ET, Richmond JC, Rockett SE, et al. The use of a generic, patient-based health assessment (SF-36) for evaluation of patients with anterior cruciate ligament injuries. *Am J Sports Med* 1996; 24:196–200.
- [25] Barber-Westin SD, Noyes FR, McCloskey JW. Rigorous statistical reliability, validity, and responsiveness testing of the Cincinnati knee rating system in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees. *Am J Sports Med* 1999;27:402–16.
- [26] McHorney CA, Ware JE Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;31:247–63.
- [27] McHorney CA, Ware JE Jr., Rogers W, et al. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. *Med Care* 1992;30(Suppl 5):MS253–65.

- [28] Ware JEJ, Snow KK, Kosinski M, et al. SF-36 health survey manual and interpretation guide. Boston: The Health Institute; 1993.
- [29] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
- [30] Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990;20:337–40.
- [31] Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995;346:1085–7.
- [32] Rosner B. *Fundamentals of biostatistics*. Toronto: Duxbury Press; 1995.
- [33] Anderson AF, Federspiel CF, Snyder RB. Evaluation of knee ligament rating systems. *Am J Knee Surg* 1993;6:67–73.