*Journal of Athletic Training* 2013;48(3):331–336
doi: 10.4085/1062-6050-48.2.11

www.natajournals.org

*original research*

# Test-Retest and Interrater Reliability of the Functional Movement Screen

**Rebecca Shultz, PhD*; Scott C. Anderson, MA†; Gordon O. Matheson, MD, PhD*; Brandon Marcello, PhD†; Thor Besier, PhD‡**

*Department of Orthopaedic Surgery, Stanford University School of Medicine, CA; †Department of Athletics, Physical Exercise, and Recreation, Stanford University, Palo Alto, CA; ‡Auckland Bioengineering Institute, University of Auckland, New Zealand

***Context:*** The Functional Movement Screen (FMS) is a popular test to evaluate the degree of painful, dysfunctional, and asymmetric movement patterns. Despite great interest in the FMS, test-retest reliability data have not been published.

***Objective:*** To assess the test-retest and interrater reliability of the FMS and to compare the scoring by 1 rater during a live session and the same session on video.

***Design:*** Cross-sectional study.

***Setting:*** Human performance laboratory in the sports medicine center.

***Patients or Other Participants:*** A total of 21 female (age = 19.6 ± 1.5 years, height = 1.7 ± 0.1 m, mass = 64.4 ± 5.1 kg) and 18 male (age = 19.7 ± 1.0 years, height = 1.9 ± 0.1 m, mass = 80.1 ± 9.9 kg) National Collegiate Athletic Association Division IA varsity athletes volunteered.

***Intervention(s):*** Each athlete was tested and retested 1 week later by the same rater who also scored the athlete's first session from a video recording. Five other raters scored the video from the first session.

***Main Outcome Measure(s):*** The Krippendorff α (K α) was used to assess the interrater reliability, whereas intraclass correlation coefficients (ICCs) were used to assess the test-retest reliability and reliability of live-versus-video scoring.

***Results:*** Good reliability was found for the test-retest (ICC = 0.6), and excellent reliability was found for the live-versus-video sessions (ICC = 0.92). Poor reliability was found for the interrater reliability (K α = .38).

***Conclusions:*** The good test-retest and high live-versus-video session reliability show that the FMS is a usable tool within 1 rater. However, the low interrater K α values suggest that the FMS within the limits of generalization should not be used indiscriminately to detect deficiencies that place the athlete at greater risk for injury. The FMS interrater reliability may be improved with better training for the rater.

***Key Words:*** repeatability, preparticipation screening, injury-prevention screening

**Key Points**

- Within the limits of generalizability, the Functional Movement Screen (FMS) had good test-retest reliability but low interrater reliability.
- Given that FMS scores are not comparable between raters and these differences may influence the score's clinical utility, the FMS should be used cautiously to detect deficiencies that place the athlete at greater risk for injury.
- Including a population with a wider range of FMS scores may improve the reliability of the scoring system.

Screening tools identify the presence or absence of an identified risk factor, which then requires follow up.[1] One class of screening tools has been developed to identify functional movement deficiencies that may place individuals at increased risk for musculoskeletal injury.[2–5] This type of screening tool, which typically assesses a particular movement pattern, is used with a training intervention designed to address the identified "faulty" mechanics through specific corrective exercise.[3] These tools rely on the notion that functional limitations may predispose an athlete to injury.

One popular screening tool of this class is the Functional Movement Screen (FMS).[6,7] It consists of 7 fundamental movements that assess mobility and stability. The creators also developed a series of corrective exercises that are prescribed based on the level and type of faulty movement patterns achieved while performing the FMS and identified from an individual's FMS score. Together with these corrective exercises, the FMS is promoted to reduce the risk of sport-related musculoskeletal injury.

Although the FMS is popular, data about the reproducibility and validity of its measurements are lacking. Researchers[8] investigating the interrater reliability of the FMS have compared FMS scores by 2 experts (FMS creators) and 2 novice raters (1 year of training). Their 2 sets of raters had a high level of agreement, with 14 of 17 tests demonstrating excellent reliability.[8] They demonstrated the FMS has high interrater reliability when trained individuals complete its scoring. However, interrater is only one type of reliability; another important reliability test is test-retest reliability, which is used to assess biological variability, instrumentation error by the participant, and error by the rater.[9] The reproducibility of a test needs to be established before its validity can be determined, so

**Table 1. Comparison of Raters who Scored the Functional Movement Screen of All 39 Athletes to Establish Interrater Reliability**

| Task | Experience | Average of Total Scores | Average SD | Difference from Rater 1 |
|---|---|---|---|---|
| Rater 1: student | <1 mo | 17.20 | 1.34 | 0.00 |
| Rater 2: physical therapist | 6–9 mo | 16.80 | 1.47 | 0.40 |
| Rater 3: athletic trainer | 3–4 y, noncertified Functional Movement Screen user | 16.90 | 1.39 | 0.30 |
| Rater 4: strength and conditioning coach | 3 y | 17.70 | 1.50 | −0.50 |
| Rater 5: strength and conditioning coach | 2 y | 16.20 | 1.23 | 1.00 |
| Rater 6: athletic trainer | <1 y | 17.40 | 1.50 | −0.20 |

research to investigate FMS scoring metrics[10–12] or its use as an outcome measure in injury prevention[13] is premature.

Therefore, the primary purpose of our study was to assess the test-retest reliability of the FMS and to determine the interrater reliability across a group of raters. Our secondary purpose was to compare the scoring by 1 rater for a live session and the same session on video.

## METHODS

### Participants

A total of 39 (21 women: age = 19.6 ± 1.5 years, height = 1.7 ± 0.1 m, mass = 64.4 ± 5.1 kg; 18 men: age = 19.7 ± 1.0 years, height = 1.9 ± 0.1 m, mass = 80.1 ± 9.9 kg) National Collegiate Athletic Association Division IA varsity athletes who competed in swimming, soccer, volleyball, cross country, or gymnastics volunteered to participate in this study. Six raters scored the FMS performed by the athletes (1 undergraduate student, 1 physical therapist, 2 athletic trainers, and 2 strength and conditioning coaches; Table 1). Five of 6 raters were trained by a certified FMS administrator; 1 undergraduate student was self-taught and demonstrated her ability to a certified FMS instructor before data collection. All participants (N = 39) provided written informed consent, and the study was approved by the Stanford University Ethics Committee.

### Equipment

We recorded each athlete's movement with 2 digital video cameras (model PV-GS500, Panasonic, Osaka, Japan) recording at 30 frames per second and positioned in the sagittal and frontal planes. The cameras were positioned on tripods in the same location and height for all tests (Figure 1). Siliconcoach Pro 7 (Siliconcoach Ltd, Dunedin, New Zealand) video-analysis software was used to simultaneously capture the 2 video streams and save them to the hard drive of a laptop computer. This software also was used in the data analysis to enable the rater to view the movement frame by frame or in slow motion at a deinterlaced 60 frames per second.

### Procedures

The FMS comprises 7 tasks: overhead squat, hurdle step, in-line lunge, active hamstrings, shoulder mobility, trunk stability, and rotary stability. Each task is scored from 0 to 3, and the maximum total score is 21. Each task has specific criteria outlined to help the rater differentiate among 0, 1, 2, and 3 scores. Any pain identified during the movement or clearing tests automatically results in a score of 0. An athlete only receives a 3 if the movement meets all the criteria outlined in the manual.[6,7]

### Reliability Analysis

Each participant attended 2 sessions that were conducted 1 week apart,[14] during which he or she completed the full FMS protocol. All athletes had performed the FMS with their athletic trainer or performance coach, none of whom were raters in this study, so a familiarization session was unnecessary. The test was not randomized because the FMS has a standardized testing sequence. One rater, who was not an author, administrated all testing sessions for all athletes to ensure consistency in the instruction set and video capturing. This same rater evaluated each athlete during these testing sessions (live rater). The live rater also scored the athlete's first video session to compare the scores from the live session with the scores from the video-based session. This rater was blinded to all previous scores from the live session. The 6 raters analyzed the videos recorded during the first session for 39 athletes and evaluated each athlete immediately after receiving the video (interrater).

### Statistical Analysis

**Test-Retest Reliability and Live-Versus-Video Session Reliability.** Intraclass correlation coefficients (ICCs) were used to assess the test-retest or live-versus-video session reliability. The ICCs evaluate the relative reliability by assessing the variance due to how the scores differ from each other, which is calculated by dividing the variability among scores by the total variance.[9] The total variance includes both random and systematic errors.[9] The ICC varies from 0 to 1, where 1 is considered *perfectly reliable*; for this study, an ICC greater than 0.75 was considered *excellent*, from 0.4 to 0.75 was considered *fair to good*, and less than 0.4 was considered *poor*.[15]

**Interrater Repeatability.** The interrater reliability for the total FMS score (interval; maximum score = 21) and the individual scores (ordinal; maximum score = 4) was assessed by calculating the Krippendorff α (K α).[16] The K α requires a value of .8 to be considered acceptable or .65 if tentative conclusions are deemed acceptable.[17] The FMS task scores are a multicategory system that depends on the hierarchical nature of the test. The K α was developed to establish a reliability measure that works well with more than 2 raters and different types of data (interval, ordinal) and can correctly handle missing data.[16] The total FMS scores were categorized as *interval data*, whereas the individual tasks were categorized as *ordinal data*. The 95% confidence intervals (CIs) were calculated by bootstrapping (n = 1000). Means and standard deviations were calculated for each rater across the sample, and the mean of the 6

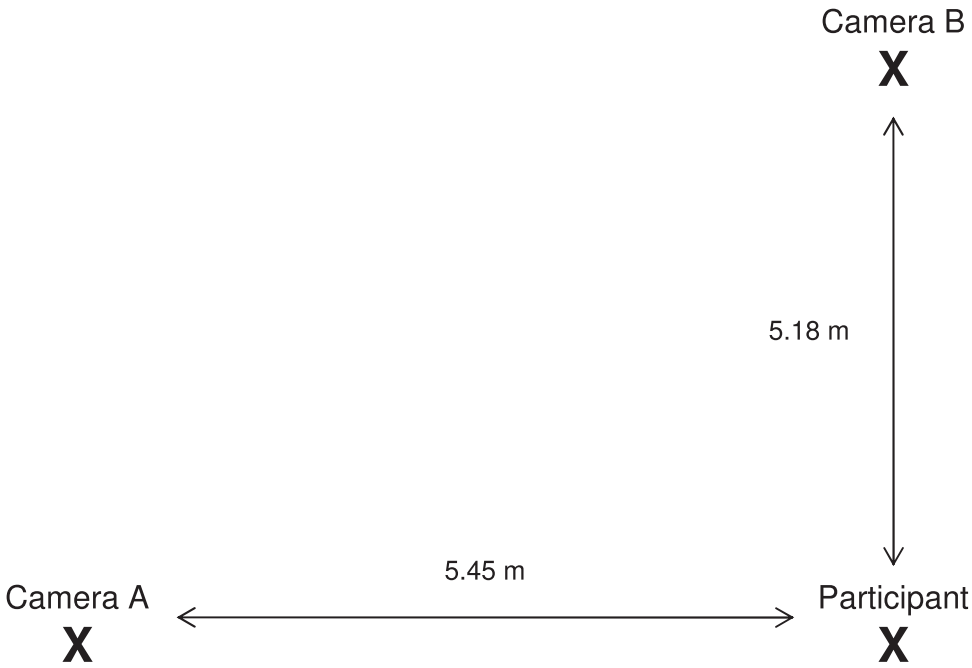


**Figure 1. The laboratory setup during the data-collection phase. Participants maintained their hips over the X. Camera A recorded the athlete from the sagittal view. Its height was 80 cm and distance from the participant was 5.45 m. Camera B recorded the athlete from the frontal view. Its height was 48 cm for the overhead squat, hurdle-step, and in-line lunge tasks and 80 cm for the shoulder-mobility (zoomed in to heels of participant), active-hamstrings, trunk-stability, and rotary-stability tasks. Its distance from the participant was 5.18 m.**

raters then was calculated. The ICCs also were used to assess the reliability of the total FMS scores given by raters who have similar experience using the FMS. Two groups of 2 raters had similar experience: those with less than 1 year of experience and those with more than 2 years of experience.

## RESULTS

In 8 of 1638 (0.5%) possible scoring opportunities, a rater could not evaluate 1 of the 7 tasks, and for less than 3% of the scoring opportunities, a rater reviewed the video before providing a score due to a technical issue or because the rater believed that not enough information was available to score the athlete. The K $\alpha$ method eliminates any participant who does not have scores from 2 or more raters. Given this criterion, the interrater reliability of the total scores was calculated using the scores of 36 athletes. Only 36 athletes were included in the analysis for the test-retest reliability because 3 athletes did not return for their second sessions due to scheduling difficulties ($n = 2$) or an injury ($n = 1$; Figure 2).

The reliability for the live-versus-video sessions was consistently excellent (ICC = 0.92, 95% CI = 0.855, 0.959). The test-retest reliability was good, but the 95% CIs were much wider, indicating poorer precision (ICC = 0.6, 95% CI = 0.35, 0.77; Figure 3). The interrater reliability was considered poor (K $\alpha$ = .38, 95% CI = 0.35, 0.41). To further examine the interrater reliability, we calculated ICCs across raters with similar experience. The raters with less than 1 year of experience (1 physical therapist and 1 athletic trainer) had fair reliability (ICC = 0.44, 95% CI = 0.12, 0.67), whereas the raters with more than 2 years of experience and the same professional background had poor reliability (ICC = 0.177, 95% CI = –0.15, 0.46). The total FMS scores of all participants were within a range of 14 to 20 (Figure 4). Interrater reliabilities for the individual tasks are provided in Table 2. The in-line lunge was the least reliable task (K $\alpha$ = .1), whereas the hurdle step was the most reliable task (K $\alpha$ = .95).

## DISCUSSION

Our objectives were to assess the test-retest reliability of the FMS and to compare the scoring by 1 rater during a live session and the same session on video. The relatively good reliability score found for the test-retest analysis (ICC = 0.6) established that the FMS is a reliable test when the same rater is using it. However, the poor interrater reliability (K $\alpha$ = .38) showed caution should be taken when comparing FMS scores across raters. One interesting observation was that the raters with less experience (the athletic trainer and physical therapist) had fair reliability, whereas the raters with more than 2 years of experience had poor reliability. In a similar study, researchers assessing some of the FMS tasks found the reliability of the in-line lunge to be good,[14] whereas it was poor in our study. With a range of experience and professions, drawing any conclusions from these findings is difficult, and future researchers should determine whether profession or experience influences the interrater reliability.

We cannot determine whether these results are due to the ambiguity of the scoring criteria or the need for improved rater training. The potential difficulty in assessing these movements and the low ICCs in our study may be due to the uncertainty in the scoring criteria of a complex task that involves multiple joints and complex physical qualities, such as balance, coordination, and core stability.[14] However, as did Minick et al,[8] we believe that the rater's training is an important component. Minick et al[8] involved

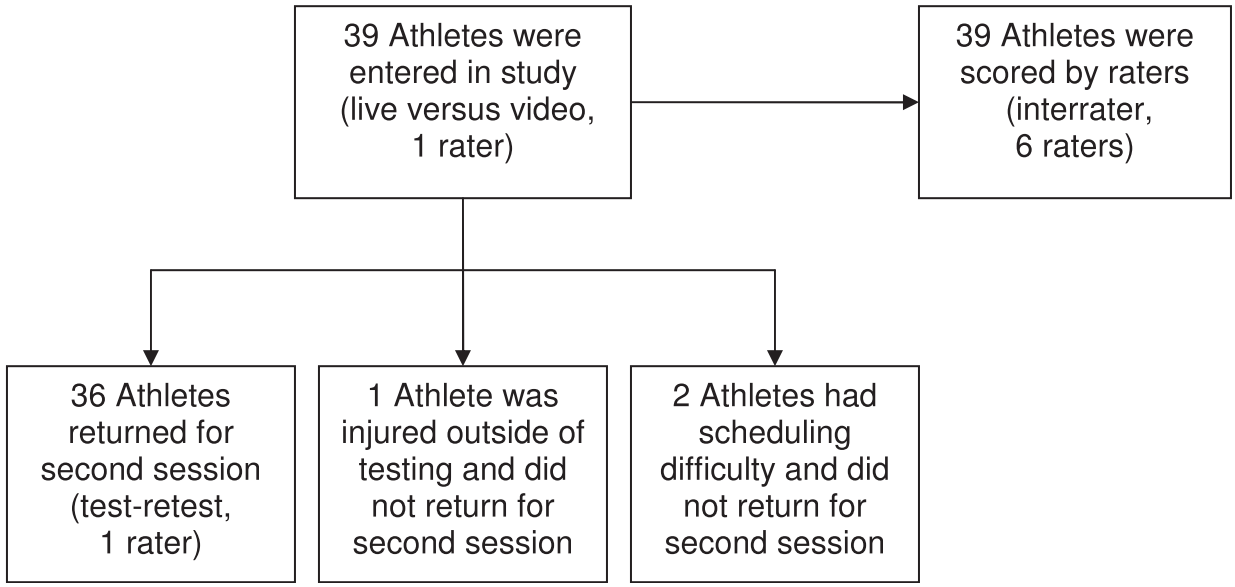


**Figure 2. Diagram shows the flow of the athletes through the study.**

expert and experienced raters who demonstrated better reliability than what we calculated. The experienced raters in the study of Minick et al[8] and the raters with more than 2 years of experience in our study had similar experience but very different reliability values. We expect that the number of FMSs performed by the 2 groups influenced their ratings. In future research, investigators should determine if the number of FMSs performed influences ratings because we suspect that the number of tests that the rater has completed is more important than his or her years of experience.

Using ICCs to calculate the test-retest reliability, we can evaluate the systematic error and the random error.[9] The systematic error was an important component in our study because scoring might be influenced by a learning effect. The potential learning effect of the athlete is one limitation of our study that may have negatively influenced the test-retest reliability. Another limitation was that 1 person administered each of the tests to maintain consistency; however, using this design, we cannot determine the variability of different administrators. This likely would add further variability to the outcomes. Specifically regarding the interrater reliability, our study also was limited by the homogeneity of the group of elite athletes (Figure 2). Each of the total scores from all 6 raters was within a range of 14 to 20, classifying our athletes as highly functioning. The K $\alpha$ calculates the variation within the range of scores included in the analysis; therefore, with no scores less than 14, the range of 0 to 14 was not included in

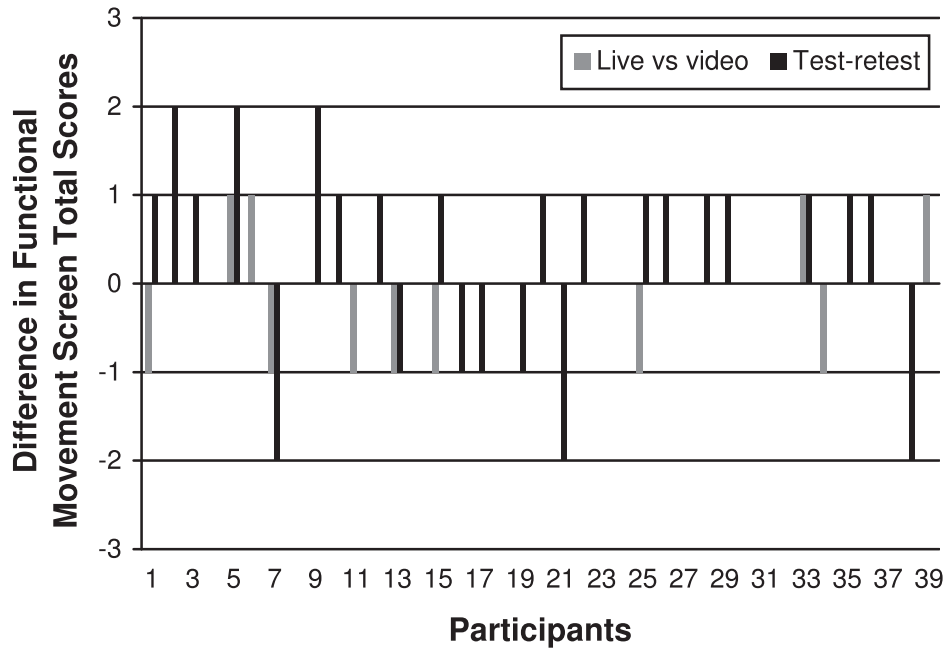


**Figure 3. Differences between the sessions for every participant for the live-versus-video (intraclass correlation coefficient [ICC] = 0.92) and the test-retest (ICC = 0.6) reliability.**

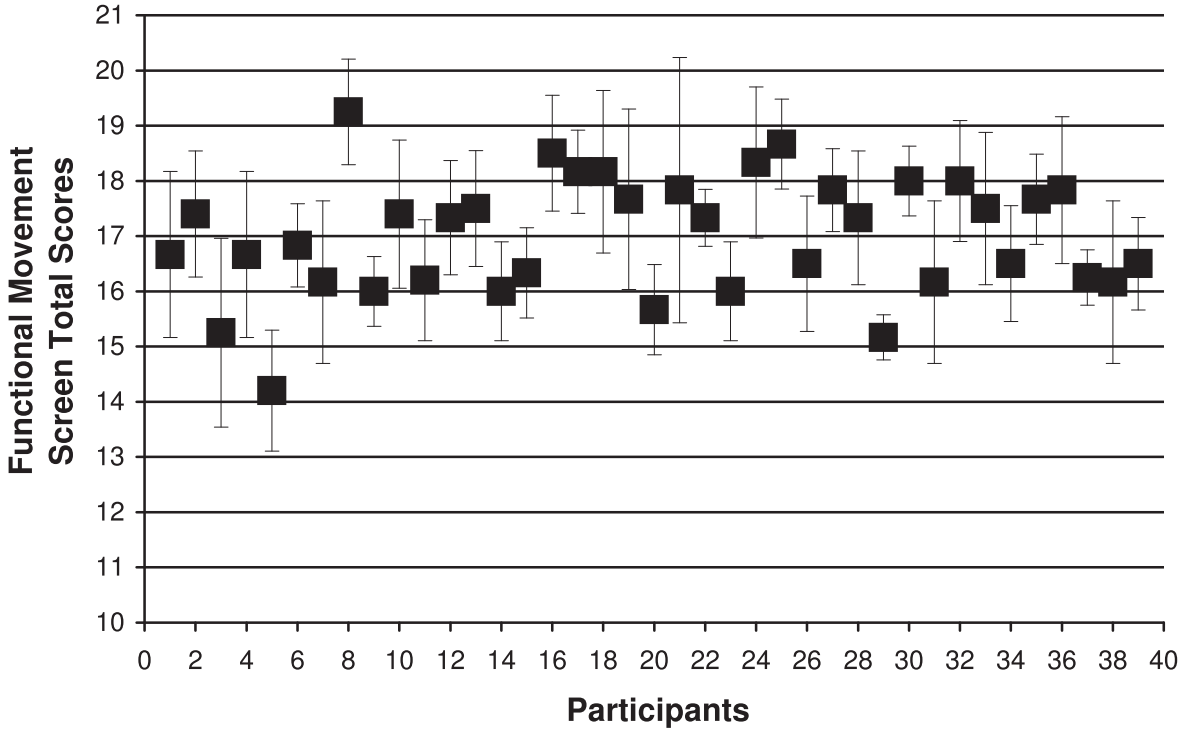


**Figure 4. Interrater reliability. Average Functional Movement Screen total scores given by each rater (N = 6) for all participants. The total scores ranged from 14 to 20, which implied that this athletic population was highly functioning. The SD error bars demonstrated the variance of the raters' scores across a participant.**

the analysis. Including individuals, such as recreational athletes or patients with injuries, who have a wider range of FMS scores would improve the reliability of the scoring system. Raters also were allowed to use features in the software to slow down the video. Given that this is not the traditional way of performing a FMS, it is a limitation. However, with such high live-versus-video session reliability, we believe this did not affect the results, and this method should be used more often for FMS.

Using video to train and test the ability of the rater to score various athletes is time efficient and should provide more consistent and objective scoring. The developers of the test have noted the benefit of interactive software programs for the evaluation of FMS.[18] This method offers rapid feedback and may reduce data-collection and -analysis times. It enables multiple ratings by different people if necessary, allows for data to be archived for postassessment, and can be used with video-analysis software to make the measurements more objective (eg, measuring distances, angles). Video also can provide valuable information for clinicians when an athlete is injured. Ensuring that the video-analysis score is comparable with a score given during a live session is important.[8] Performing the screen in a live session or scoring a video session did not affect the total score given to the athlete (ICC = 0.92). Some readers may consider the use of video to be a limitation because the FMS traditionally is scored live. This excellent reliability demonstrates that the 2 methods are comparable.

We also assessed the interrater reliability of the individual task (Table 2). One interesting observation was that the in-line lunge task was the least reliable, and the hurdle step task was the most reliable task, but they had the same standard deviation because the mean and standard deviation do not assess the consistency of each score. For example, 1 rater might rate the first 5 participants as 2, 3, 3, 2, 2, and another rater might rate the same participants as 3, 2, 2, 3, 2 for a certain task.

Our results are important because they establish a foundation for further work aimed at establishing the validity of the test. Having a highly repeatable test, which we did not find when multiple users were involved, is the first phase of establishing validity. This type of test needs to be validated for the results to be clinically useful. For now, users of the FMS need to proceed with caution when forming conclusions from the results of this screen and discussing a change in the athlete's functional-movement patterns that multiple users assessed. In the future, researchers should focus on the influence of rater and administrator training. When using the FMS for clinical

**Table 2. Functional Movement Screen Scores From All Raters (N = 6) for All Participants (N = 39)[a]**

| Task | Functional Movement Screen Total Scores Across Raters, Mean ± SD | Interrater Reliability |
|---|---|---|
| Total score | 17.00 ± 1.10 | 0.38 |
| Overhead squat | 2.18 ± 0.44 | 0.41 |
| Hurdle step | 2.38 ± 0.41 | 0.95 |
| In-line lunge | 2.62 ± 0.41 | 0.10 |
| Active hamstrings | 2.36 ± 0.19 | 0.63 |
| Shoulder mobility | 2.36 ± 0.32 | 0.64 |
| Trunk stability | 2.69 ± 0.32 | 0.31 |
| Rotary stability | 2.41 ± 0.38 | 0.25 |

[a] The interrater reliability for each task also is provided as the Krippendorff α.

purposes, clinicians and researchers should perform their own reliability tests with their own staff and population to have confidence in the screen. Investigators also should test a population whose FMS scores range from low (0–7) to high (15–21).

## ACKNOWLEDGMENTS

We thank Alex Sox-Harris for his help with the statistical analysis. We thank Lindsey Dame, Myra Tara, Floyd VitoCruz, Tammy Moreno, Lesley Moser, Katie Mooney, and Devan McConnell and all the Stanford University student-athletes who participated.

## REFERENCES

1. Bonita R, Beaglehole R, Kjellstrom T. Screening. In: *Basic Epidemiology*. 2nd ed. Geneva, Switzerland: World Health Organization; 2006:110.
2. Myer GD, Ford KR, Khoury J, Succop P, Hewett TE. Clinical correlates to laboratory measures for use in non-contact anterior cruciate ligament injury risk prediction algorithm. *Clin Biomech (Bristol, Avon)*. 2010;25(7):693–699.
3. Padua DA, Marshall SW, Boling MC, Thigpen CA, Garrett WE Jr, Beutler AI. The Landing Error Scoring System (LESS) is a valid and reliable clinical assessment tool of jump-landing biomechanics: the JUMP-ACL study. *Am J Sports Med*. 2009;37(10):1996–2002.
4. Tabor MA, Davies GJ, Kernozek TW, Negrete RJ, Hudson V. A multicenter study of the test-retest reliability of the lower extremity functional test. *J Sport Rehabil*. 2002;11(3):190–201.
5. van Mechelen W, Hlobil H, Kemper HC. Incidence, severity, aetiology and prevention of sports injuries: a review of concepts. *Sports Med*. 1992;14(2):82–99.
6. Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movement as an assessment of function. Part 1. *N Am J Sports Phys Ther*. 2006;1(2):62–72.
7. Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movements as an assessment of function. Part 2. *N Am J Sports Phys Ther*. 2006;1(3):132–139.
8. Minick KI, Kiesel KB, Burton L, Taylor A, Plisky P, Butler RJ. Interrater reliability of the functional movement screen. *J Strength Cond Res*. 2010;24(2):479–486.
9. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005; 19(1):231–240.
10. Chorba RS, Chorba DJ, Bouillon LE, Overmyer CA, Landis JA. Use of a functional movement screening tool to determine injury risk in female collegiate athletes. *N Am J Sports Phys Ther*. 2010;5(2):47–54.
11. Hoover D, Killian CB, Bourcier B, Lewis S, Thomas J, Willis R. Predictive validity of the functional movement screen in a population of recreational runners training for a half marathon. *Med Sci Sports Exerc*. 2008;40(5):S219.
12. Kiesel K, Plisky PJ, Voight ML. Can serious injury in professional football be predicted by a preseason functional movement screen? *N Am J Sports Phys Ther*. 2007;2(3):147–158.
13. Kiesel K, Plisky P, Butler R. Functional movement test scores improve following a standardized off-season intervention program in professional football players. *Scand J Med Sci Sports*. 2011;21(2): 287–292.
14. Frohm A, Heijne A, Kowalski J, Svensson P, Myklebust G. A nine-test screening battery for athletes: a reliability study. *Scand J Med Sci Sports*. 2012;22(3):306–315.
15. Streiner DL, Norman GR. *Health Measurements: A Practical Guide to Their Development and Use*. 2nd ed. New York, NY: Oxford University Press; 1995.
16. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas*. 2007; 1(1):77–89.
17. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: SAGE Publications; 2004.
18. Jaffe L, Cook G. One frame at a time. *Training Conditioning*. 2006; 16(8). http://www.training-conditioning.com/2007/03/08/one_frame_at_a_time/index.php. Accessed October 5, 2012.

*Address correspondence to Rebecca Shultz, PhD, Stanford University, Department of Orthopaedic Surgery, Stanford University School of Medicine, 341 Galvez Street, Stanford, CA 94305. Address e-mail to rshultz@stanford.edu.*