Original Research

Item-level and Composite-level Interrater Reliability of Functional Movement Screen™ Scores Following Condensed Training in Novice Raters

Brent A Harper, PT, DPT, DSc[1] [a], Stephen M Glass, PhD[2]

[1] Physical Therapy, Chapman University (CA), [2] Phyiscal Therapy, Radford University

BACKGROUND

The Functional Movement Screen™ (FMS™) is a clinical instrument designed to use movement behaviors to screen individuals for injury risk. Current rater certification programs focus on extensive, individualized training, which may not be appropriate in all screening contexts.

PURPOSE

The purpose of this research was to examine the effect of a two-hour FMS^{TM} training seminar on measures of reliability between previously untrained scorers.

STUDY DESIGN

Repeated measures, descriptive cohort study.

METHODS

Four novice raters completed a two-hour training course administered by an FMS™-certified, licensed physical therapist. The novices and the instructor then scored a group of 16 individuals on the seven FMS™ component tests on two separate occasions. Interrater reliability was assessed for FMS™ component scores using Fleiss' kappa and Krippendorff's α. Interrater reliability for the FMS™ composite score was assessed using a two-way ICC for agreement (a priori significance level=0.05).

RESULTS

Reliability ranged from fair to almost perfect (kappa) for Deep Squat (0.61 Day 1, 0.79 Day 2), Shoulder Mobility (0.90 Day 1, 1.00 Day 2), Active Straight Leg Raise (0.53 Day 1, 0.69 Day 2), and Trunk Stability Push Up (0.48 Day 1, 0.49 Day 2) on both testing occurrences (p<0.05). Reliability (kappa) was fair for Inline Lunge (0.24 Day 1, 0.39 Day 2), and poor for Hurdle Step (Day 1 -0.01, Day 2 no result) and Rotary Stability (Day 1 -0.03, Day 2 -0.01). Results for Krippendorff's α were similar, with unacceptable interrater reliability for Hurdle Step (Day 1 -0.01, Day 2 1.00), Inline Lunge (Day 1 0.31, Day 2 0.39), and Rotary Stability (Day 1 -0.02, Day 2 -0.01). Interrater composite score reliability (ICC) was good (0.79 Day 1, 0.84 Day 2; both p<0.05).

CONCLUSIONS

Findings suggest that a brief training seminar may be sufficient to ensure acceptable reliability in many, but not all, of the FMS™ component tests and composite score.

a **Corresponding author:**
   Brent Harper
   Chapman University, Crean College of Health and Behavior Sciences, Department of Physical Therapy
   9401 Jeronimo Road, Irvine, CA 92618
   E-mail: brharper@chapman.edu Phone: 714-516-5946; Fax: 949-206-0012

Levels of Evidence

Level 2b

BACKGROUND

The toll of musculoskeletal injuries is difficult to quantify, but is likely substantial among nations across the economic spectrum. The ramifications of musculoskeletal injury are far-reaching and include costs related to healthcare as well as impact on quality of life, future health, and workplace productivity, to name a few.[1] Physical activity, despite its readily apparent benefits to physical health, increases one's exposure to potentially injurious events and is often implicated in initiating the cycle of injury-related personal and societal costs. Recent epidemiological studies of sport-related injury in the U.S. estimate 8.6 million Americans report an activity-related injury each year.[2] Preserving the benefits of physical activity while avoiding adverse outcomes requires a balance between participation and, where possible, minimizing exposure.[3]

One potential method for reducing such exposures involves screening for or modifying high-risk movement behaviors. The developers of the FMS™ proposed that the practice of sports medicine was lacking with respect to injury risk screening.[4,5] They describe a gap between 1) the pre-participation medical clearance exam, and 2) performance testing designed to guide sport-related training or tactical decisions. Their solution, which has since gained considerable traction, involves the screening of fundamental movement behaviors as an indicator of potential activity-related injury risk and as an initial means of identifying possible avenues of remediation.

Initial research on the FMS™ indicated that it may help prospectively discriminate individuals at high vs. low risk for activity-related injury on the basis of a standardized movement assessment battery.[6] This observation has led to an increased focus on the application of movement screens, both as a predictor of risk and to support the design of training programs. Additional movement assessment instruments developed to date have sought to address a range of populations and specific activity-based needs.[7–11] These developments, and the accelerating pace of research on the topic of movement quality, attest to the continued interest in applying such instruments clinically.

Notwithstanding, the proliferation of movement screens as a pre-participation tool has led to a concomitant increase in the demand for raters and the lack of demonstrated competence with visual observation when evaluating movement. As the scale of application increases for the FMS™ and similar clinical instruments, there is a potential for their reliability to suffer within and across studies. This may stem from variability in rater expertise, individual raters adopting personal preferences in rating style, or the mutual influence of different screening systems featuring similar component tests. Any such source of error has the potential to affect clinical and scientific interpretation of the associated rating systems. Alternatively, one may increase confidence in their meaning to the extent such sources of error can be addressed. A feasible method of calibrating clinical movement assessments (or the raters who rate them) may

help ensure data quality and insulate these instruments from reliability concerns associated with scale of application.

Assessing practical methods by which raters with varying levels of experience as a movement professional—and varying levels of exposure to specific movement assessment instruments—can achieve greater reliability in applying movement quality assessments. This may be particularly useful in high-volume settings, in which effects related to rater variation have a greater likelihood of obscuring meaningful trends.

The subject of FMS™ reliability among raters of varying experience has been partially addressed by previous work. While specific findings vary by study, authors appear to conclude more often than not that the instrument is reliable for the purposes investigated.[12,13] Even so, valid concerns have been raised about the conclusiveness of the research,[14] the analytical approaches involved,[15] and the psychometric properties of the FMS™ as a rating instrument.[16] Establishing reliability of the FMS™ and similar movement quality assessment scales should be considered an ongoing effort. The body of literature addressing FMS™ interrater reliability has thus far given little attention to expediently calibrating or "synchronizing" item and composite scores across novice raters, which is a priority in high-volume applications or any time multiple raters are involved. This study examined the effect of a brief training seminar—administered by a licensed physical therapist who is FMS™-certified—on interrater reliability of FMS™ scores among individuals with no prior exposure to the instrument or its scoring criteria. Such a seminar could feasibly be administered prior to large scale testing endeavors to reduce measurement noise. Data was analyzed at the level of the component scores and the composite score, in each case using models that account for the type of data and number of raters. The purpose of this research was to examine the effect of a two-hour FMS^TM training seminar on measures of reliability between previously untrained scorers. It is hypothesized that a brief, standardized training seminar will be sufficient to achieve good to strong interrater reliability for all FMS™ components.

METHODS

EXPERIMENTAL APPROACH TO THE PROBLEM

Component (i.e., item) and composite FMS™ scores were acquired on two occasions from a group of five raters. The raters consisted of four novice second-year physical therapy students with no prior FMS™ training or experience, and one expert who was FMS™ certified with three years' experience using FMS™ and has been a licensed physical therapist for 20 years. The novice raters participated in a two-hour training seminar provided by the expert rater eight days prior to the initiation of data collection. The training session consisted of initially viewing each of the seven screening tests, totaling approximately 75 minutes, of the FMS™ scoring video (Functional Movement Systems). Ad-

ditionally, the seven movement patterns, three clearing tests, examiner verbal instructions, and scoring criteria were explained in detail by the expert rater. Summary sheets for each FMS™ movement were provided to the raters, including written and visual descriptions of scoring from zero to three for each movement pattern. Novice raters then performed, practiced, and scored each of the seven movement patterns and three clearing tests.

A sample of 16 subjects was scored twice by each rater with four days between each session. On both occasions, a researcher read the scripted instructions used the same materials as used in the training session to have the subjects perform each test. The tests were scored in real-time by all raters simultaneously and subsequently analyzed to establish reliability.

SUBJECTS

A total of sixteen subjects (12 females [23.33 ± 1.61 years, 164.68 ± 5.94 cm, 61.97 ± 9.33 kg] and four males [23.75 ± 1.71 years, 181.61 ± 10.47 cm, 88.22 ± 20.18 kg]) participated in this study. Participation was open to healthy adults without restrictions to physical activity. Prior to participation, subjects signed an informed consent form approved by the university Institutional Review Board.

PROCEDURES

Participants reported to the testing site on Day 1 of testing, and returned to repeat the test four days later (Day 2) at the same location. Upon arrival, participants were instructed in the performance of each movement pattern in the order specified by Cook et al.[4,5] The standardized order of movement patterns and tests was as follows: 1) Deep Squat (DS), 2) Hurdle Step (HS), 3) Inline Lunge (ILL), 4) Shoulder Mobility (SM), 5) Shoulder Clearing Tests, 6) Active Straight Leg Raise (ASLR), 7) Trunk Stability Push Up (TSPU), 8) Spinal Extension Clearing Test, 9) Rotary Stability (RS) (prior to changes of 2020), 10) Spinal Flexion Clearing Test. Test order and verbal instructions were scripted for criteria to meet scores of "grade 3" or "grade 2" and each subject completed each test position regardles of rater's score. All raters observed and scored the same subject at the same time. Raters were permitted to move about the testing room and to request that participants perform additional repetitions of any test, but were not permitted to discuss scores. These same procedures were repeated four days later. Participants were instructed not to practice the test behaviors between the first and second testing occasions. Summary sheets for each FMS™ movement were provided, including written and visual descriptions of scoring for each movement pattern. Novice raters performed, practiced, and scored each of the seven movement patterns and three clearing tests. Prior to data collection, interrater reliability for novice raters for the DS, HS, and ILL movement patterns was rated and found to have excellent reliability after viewing and scoring video clips of these three movement patterns. These three movement patterns were selected by the researchers due to the increased complexity of the grading criteria for those movement patterns when compared to the other movement patterns.

Each item was rated by all participants in real-time based on the originally published scoring criteria as instructed during the training seminar. Raters were additionally instructed to record the lower of two scores as the component score for any test in which a bilateral asymmetry was noted, and to assign a component score of 0 in any test which pain was reported or if an associated clearing test was positive (i.e. evoked pain).

STATISTICAL ANALYSES

Interrater reliability was analyzed separately for each Day 1 component score and also for the Day 1 composite score, the latter of which is simply a sum of the component scores. To account for the number of raters (n > 2) and the structure of the component data, Krippendorff's α and Fleiss' Kappa were computed. Note, Krippendorff's α is designed for ordinal data whereas Fleiss' kappa is designed for categorical data. To facilitate comparison with previously published data intraclass correlation coefficients (ICC) was computed for each component score, although, it should be noted, that ICC may not be appropriate for ordinal data. For the composite score, interrater reliability was assessed using ICC. All ICC coefficients were calculated using two-way ICC models for agreement. Interrater reliability for Day 2 scores was calculated separately using the same methods described for Day 1. All statistical analyses were conducted using R version 3.6.1 (the R Foundation; Vienna, Austria) at an a priori significance level of α = 0.05. Coefficients were interpreted in accordance with published guidelines.[17,18] Specifically, ICC was interpreted as poor (0.00 – 0.40), fair/good (0.40 – 0.75), excellent (0.75 – 1.00). Krippendorff's α was interpreted as unacceptable, (0.00 – 0.65), tentatively acceptable (0.65 – 0.80), or acceptable (0.80 – 1.00). Finally, Fleiss' Kappa was interpreted as slight (0.00 – 0.20), fair (0.21 – 0.40), moderate (0.41 – 0.60), substantial (0.61 – 0.80), or almost perfect (0.81 – 1.00).

RESULTS

Score counts for each combination of Rater * Day * Test Item are shown in Table 1. Interrater reliability on Day 1 and Day 2 are summarized in Tables 2 and 3, respectively. The results vary considerably depending on the statistical test that was utilized. Interpreting Krippendorff's α, Day 1 interrater reliability was unacceptable for Hurdle Step, Inline Lunge, Active Straight Leg Raise, and Rotary Stability; tentatively acceptable for Deep Squat; and acceptable for Shoulder Mobility. Based on Fleiss' Kappa, Day 1 interrater reliability was poor for Hurdle Step and Rotary Stability (p > 0.05); fair for Inline Lunge and Trunk Stability Push Up; moderate for Active Straight Leg Raise; substantial for Deep Squat; and almost perfect for Shoulder Mobility. Day 1 ICCs indicated poor interrater reliability for Hurdle Step (p > 0.05), Rotary Stability (p > 0.05), and Inline Lunge; fair/good interrater reliability for Active Straight Leg Raise, and Trunk Stability Push Up; and excellent reliability for Deep Squat and Shoulder Mobility.

Table 1. FMS™ item score tallies by rater for each day.

Score	0					1					2					3				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
Day 1																				
DS	0	0	0	0	0	2	2	3	2	6	10	11	10	9	6	4	3	3	5	4
HS	0	0	0	0	0	0	0	0	0	0	16	16	15	16	16	0	0	1	0	0
ILL	0	0	0	0	0	0	0	0	1	1	13	12	12	12	15	3	4	4	3	0
SM	1	1	1	1	1	0	0	0	0	0	5	5	7	5	5	10	10	8	10	10
ASLR	0	0	0	0	0	4	2	3	2	2	9	7	11	10	6	3	7	2	4	8
TSPU	1	1	1	1	1	10	8	5	8	11	5	6	5	5	3	0	1	5	2	1
RS	0	0	0	0	0	1	0	1	0	0	15	16	15	16	16	0	0	0	0	0
Day 2																				
DS	0	0	0	0	0	3	3	3	3	6	11	10	10	10	8	2	3	3	3	2
HS	0	0	0	0	0	0	0	0	0	0	16	16	16	16	16	0	0	0	0	0
ILL	0	0	0	0	0	0	0	0	0	0	13	10	11	13	15	3	6	5	3	1
SM	0	0	0	0	0	0	0	0	0	0	4	4	4	4	4	12	12	12	12	12
ASLR	0	0	0	0	0	5	4	4	4	3	5	5	6	6	4	6	7	6	6	9
TSPU	0	0	0	0	0	9	7	5	7	9	6	8	4	7	6	1	1	7	2	1
RS	0	0	0	0	0	0	0	0	0	0	16	16	15	16	16	0	0	1	0	0

Raters R1-R4 are the novice raters. R5 is the expert rater. DS = Deep Squat; HS = Hurdle Step; ILL = Inline Lunge; SM = Shoulder Mobility; ASLR = Active Straight Leg Raise; TSPU = Trunk Stability Push Up; RS = Rotary Stability.

Table 2. Interrater reliability statistics for Day 1 FMS™ item scores.

Outcome	Coefficient	Statistic	Sig
ICC			
DS	0.75	$F_{(15, 60)} = 16.96$	<0.01*
HS	0.00	$F_{(15, 60)} = 1.00$	0.467
ILL	0.32	$F_{(15, 62)} = 3.58$	<0.01*
SM	0.96	$F_{(15, 60)} = 138.14$	<0.01*
ASLR	0.68	$F_{(15, 38)} = 14.95$	<0.01*
TSPU	0.68	$F_{(15, 32)} = 15.42$	<0.01*
RS	-0.02	$F_{(15, 59)} = 0.92$	0.549
Krippendorff's α			
DS	0.74	--	--
HS	-0.01	--	--
ILL	0.31	--	--
SM	0.91	--	--
ASLR	0.64	--	--
TSPU	0.68	--	--
RS	-0.02	--	--
Fleiss' Kappa			
DS	0.61	z = 10.51	<0.01*
HS	-0.01	z = -0.16	0.873
ILL	0.24	z = 3.38	<0.01*
SM	0.90	z = 13.47	<0.01*
ASLR	0.53	z = 8.95	<0.01*
TSPU	0.48	z = 8.88	<0.01*
RS	-0.03	z = -0.32	0.746

DS = Deep Squat; HS = Hurdle Step; ILL = Inline Lunge; SM = Shoulder Mobility; ASLR = Active Straight Leg Raise; TSPU = Trunk Stability Push Up; RS = Rotary Stability.

Interpreting Krippendorff's α for Day 2, interrater reliability was acceptable reliability for Deep Squat, Hurdle Step, Shoulder Mobility, and Active Straight Leg Raise; tentatively acceptable reliability for Trunk Stability Push Up; and unacceptable reliability for Inline Lunge and Rotary Stability. Fleiss' kappa indicated poor agreement for Rotary Stability (p > 0.05); fair agreement for Inline Lunge; moderate agreement for Trunk Stability Push Up; substantial agreement for Deep Squat and Active Straight Leg Raise; and almost perfect agreement for Shoulder Mobility. Day 2 ICCs indicated poor interrater reliability for Rotary Stability (p > 0.05); fair/good interrater reliability for Inline Lunge and Trunk Stability Push Up; and excellent interrater reliability for Deep Squat, Shoulder Mobility, and Active Straight Leg Raise. Day 2 interrater ICC for Hurdle Step could not be calculated.

Finally, interrater ICC for the composite score was excellent on both days (Day 1 ICC = 0.79, Day 2 ICC = 0.84; Table 4). Intraclass correlation coefficients (two-way models for agreement) calculated separately for Day 1 and Day 2 FMS™ composite scores.

DISCUSSION

The results of this study indicate that interrater FMS™ item score reliability was variable following a standardized two-hour training seminar in raters previously unfamiliar with the FMS™. We elaborate on specific FMS™ components in the following paragraphs. Additionally, we observed that interrater reliability of the composite score was excellent. One caveat that bears mentioning before further discussion is the lack of variability within certain component ratings. Specifically, nearly all raters assigned a score of "2" for every participant—on both days—in the Hurdle Step and Rotary Stability tests. Depending on the statistical test, this may result in a finding that agreement between raters is either essentially perfect or cannot be calculated. Whichever the case, these models should be interpreted with caution.

Results concerning the composite score are fairly consistent with previous findings.[13] For example, Onate et al.[19] observed an interrater ICC of 0.98 for the FMS™ composite score, and Smith et al.[20] observed interrater ICCs of 0.87 and 0.89, respectively, on two separate days of testing. The authors conclude that the composite score can be rated reliably by judges of varying levels of experience. While this observation does strengthen the case for composite scor-

Table 3. Interrater reliability statistics for Day 2 FMS™ item scores.

Outcome	Coefficient	Statistic	Sig
ICC			
DS	0.86	$F_{(15, 47)} = 38.02$	<0.01*
HS	--	--	--
ILL	0.42	$F_{(15, 61)} = 4.89$	<0.01*
SM	1.00	$F_{(15, 59)} > 1000$	<0.01*
ASLR	0.85	$F_{(15, 58)} = 32.99$	<0.01*
TSPU	0.68	$F_{(15, 25)} = 17.26$	<0.01*
RS	0.00	$F_{(15, 60)} = 1.00$	0.467
Krippendorff's α			
DS	0.85	--	--
HS	1.00	--	--
ILL	0.39	--	--
SM	1.00	--	--
ASLR	0.83	--	--
TSPU	0.72	--	--
RS	-0.01	--	--
Fleiss' Kappa			
DS	0.79	z = 13.53	<0.01*
HS	--	--	--
ILL	0.39	z = 4.94	<0.01*
SM	1.00	z = 12.65	<0.01*
ASLR	0.69	z = 12.25	<0.01*
TSPU	0.49	z = 8.21	<0.01*
RS	-0.01	z = -0.16	0.873

DS = Deep Squat; HS = Hurdle Step; ILL = Inline Lunge; SM = Shoulder Mobility; ASLR = Active Straight Leg Raise; TSPU = Trunk Stability Push Up; RS = Rotary Stability.

Table 4. Interrater reliability for Day 1 and Day2

Outcome	ICC	Statistic	Sig
Day 1	0.79	$F_{(15,59)} = 21.52$	<0.01*
Day 2	0.84	$F_{(15,40.1)} = 34.84$	<0.01*

Intraclass correlation coefficients (two-way models for agreement) calculated separately for Day 1 and Day 2 FMS™ composite scores.

ing of the FMS™, and perhaps movement quality screens in general, recent publications have highlighted serious limitations concerning this metric. Multiple factor analyses[21,22] have identified a non-unidimensional structure and/or unacceptably low internal consistency. These observations call into question the psychometric validity of the composite score independently of whether or not a reliable score can be obtained.

In contrast, FMS™ item/component scores present a more granular perspective of movement quality and may be less vulnerable to criticism concerning their psychometric qualities. The study's findings for Rotary Stability were again consistent with Onate et al.,[19] who observed that a kappa statistic could not be calculated due to lack of variability. This study's remaining results show a pattern of in-

terrater agreement that is more or less similar to that of Onate et al. for the item scores, albeit a lower coefficient in all cases except Shoulder Mobility. This may be due in part to the use of Fleiss' kappa where Onate et al. used Cohen's kappa. (The latter was not an option in this study design because of the number of raters involved.) Minick et al.[23] also used a two-rater kappa and reported generally higher agreement than this study found. Particularly noteworthy in their findings were considerably higher levels of observed agreement for Hurdle Step and Rotary Stability. Shultz et al.[18] evaluated interrater reliability of FMS™ item scores using Krippendorff's α and found unacceptable agreement in all cases *except* Hurdle Step, for which agreement was in the "acceptable" range. This may be partially attributable to the study population (DI varsity athletes), but does stand in

contrast to the present findings.

The clinical interpretation of agreement depends on the choice of reliability statistic. This study endeavored to make the case that ICC should not be used for assessing reliability of ordinally scaled items such as the FMS™ component scores. In those cases, kappa (Fleiss or Cohen) and Krippendorff's α are better suited models. In the dataset for this study, Active Straight Leg Raise and Trunk Stability Push Up—along with the Deep Squat, to a lesser extent—are perhaps the best examples of how ICC results may give the impression of an unrealistically high level of reliability. However, ambiguity of interpretation remains even when comparing results from kappa and α models. For instance, where Active Straight Leg Raise and Inline Lunge are considered "unacceptable" by α standards, the authors of this study would judge them as having moderate and fair agreement, respectively, based on their kappa models (referring to Day 1 results).

Based on the combined results for this study, the best candidates for inclusion in a high-volume screening effort following a brief, introductory training seminar would be: Shoulder Mobility, Active Straight Leg Raise, Deep Squat, and Trunk Stability Push Up. With one exception, each of these FMS™ components achieves a level of reliability that could be considered at least "moderate" (kappa) or "tentatively acceptable" (α) on both days. Active Straight Leg Raise, the exception, misses the α cutoff for being considered "tentatively acceptable" on Day 1 by a slim margin. These findings could be useful for those planning large-scale screens. Further, they might suggest a refinement of scoring criteria to the less reliable items or, at least, more focused training prior to their use.

Before concluding, this study highlights one potentially telling observation. The interrater reliability models feature five raters, one of whom was designated an "expert" and the rest "novices". The rater designations are not accounted for in the models, but are specified in the Table 1 caption. In several cases, it appears that the cluster of novice raters disagrees systematically with the expert (e.g., DS, ILL). For example, the expert rater assigned a Deep Squat score of 1 to six subjects on both Day 1 and Day 2. In contrast, only two or three subjects were assigned a Deep Squat score of 1 by the novice raters. The expert rater also stands alone in assigning more 2's and fewer 3's on the Inline Lunge (both days) when compared with the novices, the latter of whom agree more closely with each other than they do with the expert. These systematic biases existed despite checking for interrater reliability on DS, HS, and ILL during the training session. It may represent opportunities to firm up reli-

ability by modifying the training method, such as using live subjects rather than video, and by devoting additional training such that consensus is achieved with the criterion rater prior to data collection.

LIMITATIONS

There are several limitations in the current study. First, scoring by all raters was performed in real-time. While this better simulates the conditions under which the FMS™ would be administered, simultaneous assessment by five raters may have affected scores by virtue of requiring raters to view test subjects from different vantage points. This may be especially true for multidimensional tests such as the Inline Lunge, for which scores are likely to be more sensitive to viewing angle. The second limitation concerns the test subjects themselves. These individuals comprised a limited (n = 16) convenience sample of graduate students. Third, subjects may have scored differently from day 1 to day 2; however, the test subjects were blinded to their scores. Although raters may have recalled scores from Day 1, biasing their Day 2 scores, it is unlikely due to the number of scripted movement patterns tested and since re-testing was four days later. As such, our findings should be considered preliminary pending further work involving diverse samples with a greater number of observations.

CONCLUSIONS

A two-hour training session on the scoring and administration of the Functional Movement Screen™ in previously untrained raters produced acceptable interrater reliability in the Shoulder Mobility, Active Straight Leg Raise, Deep Squat, and Trunk Stability Push Up tests. Based on the results of the current study, the authors are not able to conclude that the remaining tests—Hurdle Step, Rotary Stability, and Inline Lunge—are comparably reliable after similar training. A brief training seminar could be used prior to high-volume movement screens to provide reliable measurements involving multiple raters, particularly where rater experience is limited.

REFERENCES

1. Finkelstein E, Corso PS, Miller TR. *The Incidence and Economic Burden of Injuries in the United States*. Oxford; New York: Oxford University Press; 2006.

2. Sheu Y, Chen LH, Hedegaard H. Sports- and recreation-related injury episodes in the United States, 2011-2014. *Natl Health Stat Report*. 2016;(99):1-12.

3. Marshall SW, Guskiewicz KM. Sports and recreational injury: the hidden cost of a healthy lifestyle. *Inj Prev*. 2003;9(2):100-102. doi:10.1136/ip.9.2.100

4. Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movements as an assessment of function - part 1. *N Am J Sports Phys Ther*. 2006;1(2):62-72.

5. 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.

6. 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.

7. Thelen MD, Koppenhaver SL, Hoppes CW, Shutt C, Musen JL, Williams MK. Reliability of a novel return to duty screening tool for military clinicians. *US Army Med Dep J*. 2015:14-23.

8. Rogers SA, Hassmen P, Roberts AH, Alcock A, Gilleard WL, Warmenhoven JS. Development and reliability of an athlete introductory movement screen for use in emerging junior athletes. *Pediatr Exerc Sci*. 2019;31(4):448-457. doi:10.1123/pes.2018-0244

9. McKeown I, Taylor-McKeown K, Woods C, Ball N. Athletic ability assessment: a movement assessment protocol for athletes. *Int J Sports Phys Ther*. 2014;9(7):862-873.

10. Kritz M. Development, reliability and effectiveness of the Movement Competency Screen (MCS). 2012.

11. 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. doi:10.1111/j.1600-0838.2010.01267.x

12. Bonazza NA, Smuin D, Onks CA, Silvis ML, Dhawan A. Reliability, validity, and injury predictive value of the functional movement screen: A systematic review and meta-analysis. *Am J Sports Med*. 2017;45(3):725-732. doi:10.1177/0363546516641937

13. Cuchna JW, Hoch MC, Hoch JM. The interrater and intrarater reliability of the functional movement screen: A systematic review with meta-analysis. *Phys Ther Sport*. 2016;19:57-65. doi:10.1016/j.ptsp.2015.12.002

14. McCunn R, Aus der Funten K, Fullagar HH, McKeown I, Meyer T. Reliability and association with injury of movement screens: A critical review. *Sports Med*. 2016;46(6):763-781. doi:10.1007/s40279-015-0453-1

15. Glass SM, Ross SE. Modified functional movement screening as a predictor of tactical performance potential in recreationally active adults. *Int J Sports Phys Ther*. 2015;10(5):612-621.

16. Li Y, Wang X, Chen X, Dai B. Exploratory factor analysis of the functional movement screen in elite athletes. *J Sports Sci*. 2015;33(11):1166-1172. doi:10.1080/02640414.2014.986505

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

18. Shultz R, Anderson SC, Matheson GO, Marcello B, Besier T. Test-retest and interrater reliability of the functional movement screen. *J Athl Train*. 2013;48(3):331-336. doi:10.4085/1062-6050-48.2.11

19. Onate JA, Dewey T, Kollock RO, et al. Real-time intersession and interrater reliability of the functional movement screen. *J Strength Cond Res*. 2012;26(2):408-415. doi:10.1519/JSC.0b013e318220e6fa

20. Smith CA, Chimera NJ, Wright NJ, Warren M. Interrater and intrarater reliability of the functional movement screen. *J Strength Cond Res*. 2013;27(4):982-987. doi:10.1519/JSC.0b013e3182606df2

21. Kazman JB, Galecki JM, Lisman P, Deuster PA, O'Connor FG. Factor structure of the functional movement screen in marine officer candidates. *J Strength Cond Res*. 2014;28(3):672-678. doi:10.1519/JSC.0b013e3182a6dd83

22. Kelleher LK, Beach TAC, Frost DM, Johnson AM, Dickey JP. Factor structure, stability, and congruence in the functional movementy screen. *Measurement in Physical Education and Exercise Science*. 2018;22(2):109-115.

23. 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. doi:10.1519/JSC.0b013e3181c09c04