IJSPT
INTERNATIONAL JOURNAL OF SPORTS PHYSICAL THERAPY

Mann KJ, O'Dwyer N, Bruton MR, Bird SP, Edwards S. Movement Competency Screens Can Be Reliable In Clinical Practice By A Single Rater Using The Composite Score. *IJSPT*. 2022;17(4):593-604.

Original Research

Movement Competency Screens Can Be Reliable In Clinical Practice By A Single Rater Using The Composite Score

Kerry J. Mann[1] [a], Nicholas O'Dwyer[2], Michaela R. Bruton[3], Stephen P. Bird[4], Suzi Edwards[2]

[1] School of Allied Health, Exercise and Sports Science, Charles Sturt University, [2] The Discipline of Exercise Science, The University of Sydney, [3] School of Exercise Sciences, Australian Catholic University, [4] School of Health and Medical Sciences, University of Southern Queensland

Keywords: Movement screening, injury risk, pre-elite youth athletes

https://doi.org/10.26603/001c.35666

International Journal of Sports Physical Therapy

Vol. 17, Issue 4, 2022

Background

Movement competency screens (MCSs) are commonly used by coaches and clinicians to assess injury risk. However, there is conflicting evidence regarding MCS reliability.

Purpose

This study aimed to: (i) determine the inter- and intra-rater reliability of a sport specific field-based MCS in novice and expert raters using different viewing methods (single and multiple views); and (ii) ascertain whether there were familiarization effects from repeated exposure for either raters or participants.

Study Design

Descriptive laboratory study

Methods

Pre-elite youth athletes (n=51) were recruited and videotaped while performing a MCS comprising nine dynamic movements in three separate trials. Performances were rated three times with a minimal four-week wash out between testing sessions, each in randomized order by 12 raters (3 expert, 9 novice), using a three-point scale. Kappa score, percentage agreement and intra-class correlation were calculated for each movement individually and for the composite score.

Results

Fifty-one pre-elite youth athletes (15.0±1.6 years; *n*=33 athletics, *n*=10 BMX and *n*=8 surfing) were included in the study. Based on kappa score and percentage agreement, both inter- and intra-rater reliability were highly variable for individual movements but consistently high (>0.70) for the MCS composite score. The composite score did not increase with task familiarization by the athletes. Experts detected more movement errors than novices and both rating groups improved their detection of errors with repeated viewings of the same movement.

Conclusions

Irrespective of experience, raters demonstrated high variability in rating single movements, yet preliminary evidence suggests the MCS composite score could reliably assess movement competency. While athletes did not display a familiarization effect after

[a] **Corresponding author:**
Dr Kerry Mann
School of Allied Health, Exercise & Sports Sciences
Charles Sturt University
Panorama Avenue
Bathurst NSW 2795 Australia
FAX +61 2 6338 4065
Phone +61 2 6338 4579
E-mail kmann@csu.edu.au

performing the novel tasks within the MCS for the first time, raters showed improved error detection on repeated viewing of the same movement.

Level of Evidence

Cohort study

INTRODUCTION

Due to substantial financial and social benefits in reducing injury prevalence, simple and accessible movement competency screens (MCSs) are popular among the sporting community.[1] The primary aim of most screens quantifying the movement competency of athletes is to identify movement limitations that may provide indicators of injury risk.[1] Early identification of risk[2] may in turn allow implementation of training programs to lower injury rates among youth sporting populations.[3] However, successful implementation of MCS protocols within sporting communities require them to be simple to implement, validated, reliable, cost-effective, and relevant to sport-specific injuries.

Since MCSs requires the ability of raters to identify movement errors,[4–7] an essential requirement is high inter- and intra-rater reliability.[8] These factors are critical for movement screening as a result of the observational variance that can occur with subjective rating.[9] Methodological limitations have cast doubt on studies that demonstrated good MCS reliability,[10,11] because many of these studies[11–13] have employed intra-class correlation coefficients (ICCs) to assess reliability. This statistical procedure should be applied only to continuous scalar data, not ordinal data[14] such as that reported in many MCSs. A more appropriate statistical method for assessing reliability in ordinal data is the kappa score, which removes 'chance agreement' from the analysis.[14] Although some authors have adopted the use of the kappa score to determine reliability,[10,15,16] these studies have claimed high reliability despite reporting low kappa scores [e.g. inline lunge k=0.45,[15] rotary stability final score k=0.43,[10] hurdle step total score k=0.31].[16]

The way raters view the MCS also might influence the tool's reliability, yet this appears not to have been investigated to date. Many sporting teams and clinicians have adopted simple, real-time, single-viewing and manual grading methods.[4–6] However, it may be difficult to manually rate multiple error cues that occur simultaneously in real-time.

While rater reliability has been widely studied, an additional consideration is the effect of familiarization of both athlete and rater. When an athlete firstly performs a MCS, they may never have performed some of the movements, while similarly, a novice rater may have no experience in rating them. Hence, familiarization effects may be present, but it is currently unknown whether athletes or raters require familiarization prior to MCS performance.

This study aimed to: (i) determine the inter- and intra-rater reliability of a sport specific field-based MCS in novice and expert raters using different viewing methods (single and multiple views); and (ii) ascertain whether there were familiarization effects from repeated exposure for either raters or participants. It was hypothesised that the MCS would display high inter- and intra-rater reliability for both novice and expert raters; the MCS score would change with repeated exposure of athletes and raters due to familiarization effects; and viewing the performance of the movement multiple times while focussing on different error criteria each time would increase reliability.

METHODS

SUBJECTS

Fifty-one pre-elite youth athletes who had never performed a MCS were recruited from a Regional Academy of Sport in rural Australia. Informed consent was obtained from all participants and their guardians/parents prior to data collection and all methods were approved by the Charles Sturt University Human Research Ethics Committee.

EXPERIMENTAL APPROACH TO THE PROBLEM

Athletes performed a MCS on three separate occasions, with a minimal four-week wash out between testing sessions. Performances of each movement screening were recorded, then viewed and rated three times each in randomized order by 12 raters (3 expert, 9 novice), using a three-point scale. Both inter- and intra-rater reliability were calculated using; types of raters (novice and expert) and viewing type (single and multiple views). Each group of raters (Figure 1) was limited to a total of three raters, as increasing the number beyond this sample size is not suggested to affect statistical power.[14]

PROCEDURES

Each athlete performed a MCS comprising nine dynamic movements on three separate occasions (data trial 1, 2, 3), with a minimum four-week washout period between trials. Of the 51 athletes initially screened, 43 completed two sessions and 37 completed all three screening trials; non-participation was due to absence from training. The dynamic movements included within the screen were amended from previous screening methodologies and literature to include: Tuck Jump,[6] Overhead Squat,[4] Single Leg Squat (left and right),[17] Dip Test (left and right),[17] Forward Lunge (left and right)[18] and Prone Hold[19] (See Supplementary Material). Performance of each movement by each participant was videotaped in the sagittal and frontal planes at 240 Hz (ZR-200, Casio Computer Co., Ltd, Tokyo, Japan).

Twelve individuals (n = 3 expert, n = 9 novice) rated the performance of the 51 athletes using the videos. Raters were divided into four groups based on three variables (Figure 1). The first variable was rater experience (expert or novice). An expert (E) rater was defined as an exercise and sport science professional with a minimum of one year of experience completing greater than 150 movement screens, while a novice (N) rater was defined as an individual with less than one year experience in screening. The second variable was method of viewing (single or multiple). A single

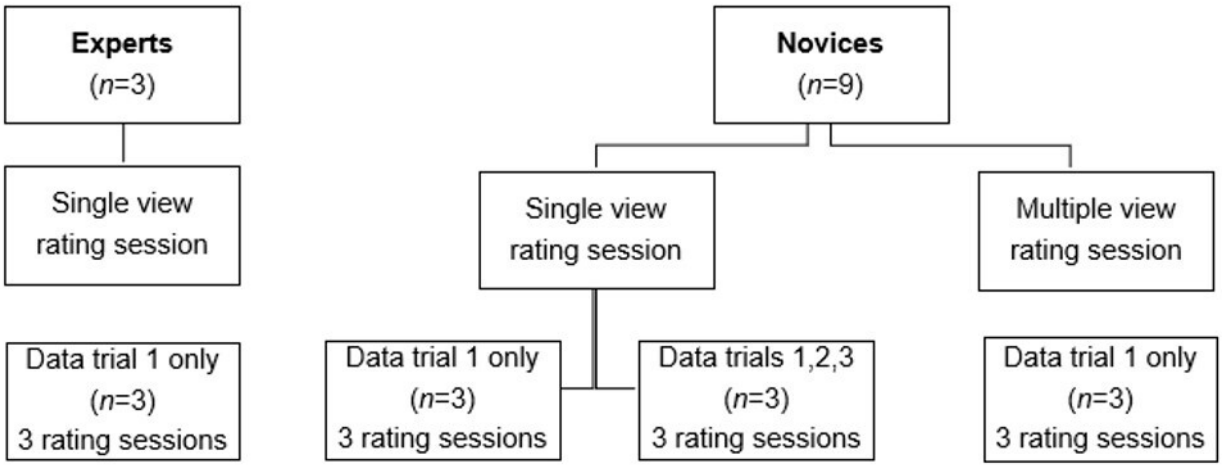

Figure 1. Raters divided into groups based on novice/expert status, MCS viewing method and video data viewed.

(Single) viewing involved the rater watching the sagittal and frontal plane videos of a movement task once, and assessing all the criteria during that viewing. A multiple (Multiple) viewing involved the rater watching the sagittal and frontal plane videos of a movement and assessing two criteria, then re-watching the videos and assessing two different criteria, and repeating this until all criteria were assessed. The third variable was the athlete trial data viewed, either data trial 1 viewed three times in separate rating sessions, or data trials 1, 2 and 3, each viewed once in separate rating sessions.

Four rating groups were formed based on these three variables (Figure 1). Novice (n=3) and expert raters (n=3) undertook single video viewings of data trial 1 only, in three separate sessions. Different novices (n = 3) undertook single video viewings, in separate sessions, from data trials 1, 2 and 3. Different novices (n=3) undertook multiple video viewings from data trial 1 only, in three separate sessions.

Novices and experts were compared to determine the effect of rating experience on detection of movement errors and reliability. MCS scores for data trials 1, 2 and 3 were compared to determine whether a familiarization effect was evident for athletes performing the movement tasks over repeated attempts. Three ratings of the video from trial 1 were carried out (in separate sessions) to determine whether a familiarization effect (increased detection of errors) was evident for the raters assessing the same movements over repeated sessions and to assess the reliability of their ratings. Single and multiple viewings of movement videos were compared to determine whether reliability was altered by simplifying the rater's task through reducing the number of criteria assessed on each viewing session.

Each rater categorized each movement task by identifying the presence of errors and counting them to yield a score of 1, 2 or 3 (1 = 3+ errors; 2 = 1-2 errors; 3 = no errors), a zero for pain is typically applied however was not applicable in this study. These individual MCS scores were then summed to give a composite score for all nine movements (maximum 27).

STATISTICAL ANALYSES

A series of repeated measures analyses of variance (ANOVAs) were conducted to determine significant differences ($p<0.05$) in total movement composite scores across repeated screenings by raters and repeated performances by athletes, i.e. to establish whether there was a familiarization effect for raters or athletes, respectively.

Percentage agreement, kappa and intra-class correlation coefficients (ICCs) were calculated for ratings of each of the nine movements to determine intra- and inter-rater reliability as a pairwise comparison between each rater and analysis method. Kappa was defined as slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect (0.81-1.00), with a negative Kappa representing less agreement than expected with chance.[20] Percentage agreement was calculated as the proportion of occasions on which both raters agreed (i.e. the sum of the occasions the raters agreed divided by the total number of occasions), expressed as a percentage.[21] To define percentage agreement, the following categories were used: poor (<50%), moderate (51-79%) and excellent (⩾80%).[20] Pearson's ICC (2,1) was used to indicate the relationship between scalar data[22] and defined as poor (<0.40), fair/good (0.40-0.75) and excellent (>0.75).[23]

Statistical procedures assessing MCS reliability often inappropriately employ ICCs to determine the reliability of ordinal (categorical) data.[11,12] This is an incorrect application of ICCs, which are appropriate only for scalar data.[14] The present study employed ICCs to assess the reliability of the MCS composite score, a scalar measure (as seen in Tables 1-3), however, ICCs were also presented for individual movement tasks in the MCS, only for comparison with previous research. The reliability of ordinal scores for individual movements of the MCS was assessed using both kappa scores to assess "true" agreement[14] and percentage agreement.[24] The measures for the nine movements were compared across sessions via t-tests to assess intra- and inter-rater reliability. Repeated measure ANOVAs and t-tests

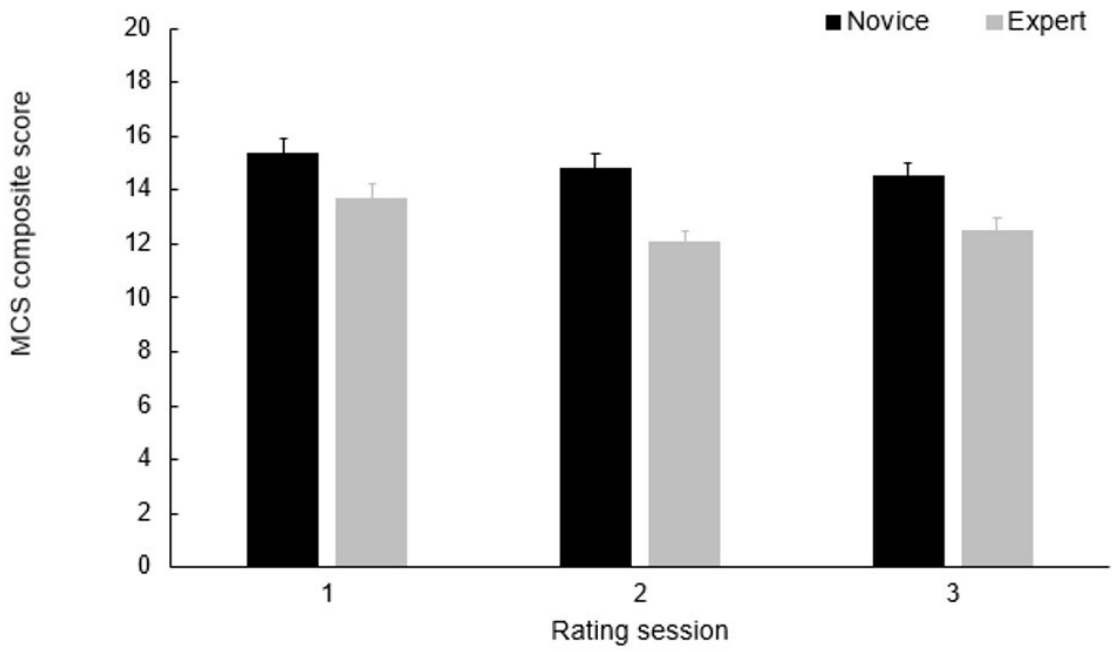

Figure 2. Novice versus expert MCS score pattern across three viewing sessions. Vertical bars denote standard errors.

were performed in Statistica (v13.6, StatSoft Inc., Tulsa, OK, USA) and statistical analyses of reliability were performed using SPSS statistical package (Version 17.0.1, SPSS Inc, Chicago, IL).

RESULTS

ATHLETE FAMILIARIZATION

The MCS composite scores achieved by athletes (15.0±1.6 years; *n*=33 athletics, *n*=10 BMX and *n*=8 surfing) for their three performance trials were analyzed separately for novices and experts because of incomplete data for one expert rater. MCS composite scores assigned by novices showed significant differences across performance trials ($F_{2,72}$ = 10.89, p<0.001, η^2=0.23) and between individual raters ($F_{2,72}$ = 184.55, p<0.001, η^2=0.84). Similarly, composite scores assigned by experts showed significant differences across trials ($F_{2,68}$ = 10.18, p<0.001, η^2=0.23) and between raters ($F_{1,34}$ = 475.32, p<0.001, η^2=0.93). Post hoc analyses for novice and expert raters revealed no clear pattern in the direction of movement competency scores across trials.

INFLUENCE OF RATING EXPERIENCE

Comparison of the MCS composite scores assigned by novice and expert raters during three viewing sessions (1, 2 and 3) of the athletes Trial 1 movement performance showed novices assigned higher MSC scores than experts ($F_{1,44}$ = 170.4, p<0.001, η^2=0.79) (Figure 2). The mean score across all sessions and raters was 14.9 for novices and 12.8 for experts, suggesting expert raters detected more errors in athlete performances. As seen in Figure 2, the pattern of MCS scores across the three viewing sessions also differed between the rater groups, as borne out by a significant interaction between groups and sessions ($F_{2,88}$ = 4.9, p<0.01, η^2=0.10). Post hoc tests showed in novices, only session 1v3 scores were significantly different (15.4 vs 14.5; p=0.007), with no significant change for session 1v2 (p=0.24) or session 2v3 (p=0.74) scores. In experts, the only significant change was session 1v2 (13.7 vs 12.1; p<0.001), with session 2v3 not being different (p=0.51).

INTRA-RATER RELIABILITY - NOVICES

The novice intra-rater reliability between session 1 and 2 in the single view of the performance of the movements in Trial 1 (Table 1, left half) was shown to have fair kappa scores across all movements, except for a slight score in the tuck jump (0.18) and moderate score in the right lunge (0.43). Between session 2 and 3, there was a general increase in kappa scores compared to session 1v2 (p<0.01), with an improvement to moderate in the overhead squat (0.33 to 0.59), left lunge (0.26 to 0.52) and prone hold (0.32 to 0.45). Moderate percentage agreement was observed for all movements between session 1 and 2, with the right single leg squat scoring excellent (80%). Again, there was a general increase in scores between session 2 and 3 compared to session 1v2 (p<0.0001), with the category changing to excellent for the overhead squat (81%) and left single leg squat (83%). In contrast, the ICCs for the MCS composite score indicated excellent reliability for both sessions 1v2 (0.85) and 2v3 (0.89).

Table 1. Intra-rater Reliability of Trial 1 – Novice Raters.

Movement	Single View						Multiple View					
	Rating session 1 v 2			Rating session 2 v 3			Rating session 1 v 2			Rating session 2 v 3		
	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%
Tuck Jump	0.48	0.18	67	0.40	0.22	78	0.41	0.24	61	0.51	0.33	66
Overhead Squat	0.71	0.33	73	0.80	0.59	81	0.72	0.45	71	0.66	0.37	68
Single leg squat left	0.29	0.22	77	0.70	0.27	83	0.55	0.38	77	0.48	0.39	82
Single leg squat right	0.27	0.25	80	0.78	0.29	83	0.52	0.37	78	0.46	0.36	80
Dip test left	0.50	0.30	62	0.65	0.39	68	0.44	0.23	64	0.52	0.25	62
Dip test right	0.59	0.33	63	0.63	0.39	67	0.62	0.36	69	0.53	0.31	64
Lunge left	0.54	0.26	73	0.73	0.52	78	0.72	0.48	76	0.65	0.44	73
Lunge right	0.58	0.43	76	0.70	0.51	79	0.74	0.51	79	0.71	0.49	76
Prone hold	0.57	0.32	52	0.80	0.45	65	0.83	0.45	63	0.77	0.48	64
Total/Mean(±SD) Score	0.85	0.29 (0.07)	69 (9)	0.89	*0.40 (0.13)	*76 (7)	0.95	0.39 (0.10)	71 (7)	0.88	0.38 (0.08)	71 (7)

*p<0.01 compared with session 1v2.

Table 2. Inter-rater Reliability of Trial 1 - Novice Raters.

Movement	Single View									Multiple View								
	Rating session 1			Rating session 2			Rating session 3			Rating session 1			Rating session 2			Rating session 3		
	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%
Tuck Jump	0.50	0.28	56	0.29	0.16	45	0.39	0.12	49	0.30	0.11	49	0.27	0.22	59	0.19	0.09	47
Overhead Squat	0.65	0.38	64	0.56	0.34	59	0.60	0.31	61	0.67	0.38	65	0.61	0.29	59	0.42	0.10	45
Single leg squat left	0.30	0.13	60	0.06	0.08	63	0.12	0.06	72	0.37	0.18	62	0.60	0.43	76	0.07	0.05	71
Single leg squat right	0.32	0.09	61	0.08	0.14	65	0.17	0.32	71	0.37	0.29	63	0.52	0.25	70	0.19	0.13	68
Dip test left	0.33	0.03	40	0.34	0.05	33	0.32	0.08	40	0.43	0.08	50	0.44	0.14	61	0.38	-0.04	36
Dip test right	0.35	0.07	44	0.50	0.11	41	0.44	0.16	49	0.45	0.09	49	0.59	0.24	62	0.28	-0.03	45
Lunge left	0.16	0.08	62	0.30	0.03	58	0.28	0.11	52	0.47	0.16	53	0.55	0.21	63	0.09	0.04	45
Lunge right	0.30	0.20	65	0.51	0.22	68	0.25	0.07	51	0.46	0.16	64	0.48	0.14	63	0.19	0.06	48
Prone hold	0.12	0.06	17	0.39	0.08	21	0.50	0.19	37	0.70	0.19	40	0.60	0.25	45	0.53	0.26	47
Total/ Mean (±SD) Score	0.75	0.15 (0.12)	52 (16)	0.85	0.13 (0.10)	50 (16)	0.73	0.16 (0.10)	54 (12)	0.84	0.18 (0.10)	55 (9)	0.91	0.24 (0.09)	62 (8)	0.70	*0.07 (0.09)	*50 (11)

*p<0.02 compared with session 2.

Table 3. Intra- and Inter-rater Reliability of Trial 1 - Expert Raters. *p<0.02 compared with session 1v2.

Movement	Intra-rater						Inter-rater								
	Rating session 1 v 2			Rating session 2 v 3			Rating session 1			Rating session 2			Rating session 3		
	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%	ICC	k	%
Tuck Jump	0.43	0.41	77	0.34	0.22	77	0.26	0.18	63	0.48	0.36	72	0.28	0.20	71
Overhead Squat	0.70	0.58	78	0.61	0.45	82	0.65	0.32	67	0.58	0.04	73	0.69	0.43	73
Single leg squat left	0.15	0.49	85	0.03	0.06	83	0.42	0.24	81	0.29	0.20	89	0.57	0.29	87
Single leg squat right	0.17	0.46	85	0.09	0.03	87	0.34	0.17	80	0.15	0.30	93	0.59	0.31	87
Dip test left	0.32	0.04	53	0.43	0.22	55	0.33	0.07	39	0.51	0.39	70	0.37	0.24	55
Dip test right	0.30	0.25	43	0.49	0.17	58	0.26	0.06	39	0.39	0.26	66	0.61	0.35	62
Lunge left	0.23	0.07	55	0.57	0.35	63	0.10	0.07	36	0.56	0.35	65	0.54	0.33	64
Lunge right	0.34	-0.10	58	0.53	0.28	68	0.19	0.05	35	0.39	0.25	62	0.52	0.31	63
Prone hold	0.46	0.24	45	0.50	0.23	57	0.25	-0.01	24	0.03	0.10	31	0.39	0.17	30
Total/Mean(±SD) Score	0.81	0.27 (0.23)	64 (17)	0.85	0.22 (0.13)	*70 (12)	0.71	0.13 (0.11)	52 (21)	0.88	0.25 (0.12)	##69 (18)	0.85	#0.29 (0.08)	##66 (17)

#p<0.02 compared with session 1, ##p<0.001 compared with session 1.

Multiple viewings of videos of the movements in trial 1 did not improve the intra-rater reliability of novices (Table 1, right half), either for kappa scores (p=0.41) or percentage agreement (p=0.62). Moreover, unlike the single view, there was no increase in kappa scores (p=0.99) or percentage agreement (p=0.99) for session 2v3 compared with session 1v2. As with the single view data, the reliability of the MCS composite score again indicated excellent reliability, with ICCs of 0.95 and 0.88 for session 1v2 and session 2v3 respectively.

INTER-RATER RELIABILITY - NOVICES

In viewing sessions 1, 2 and 3 of Trial 1 (Table 2, left half), the novice inter-rater reliability was poor, with kappa scores varying from slight to fair (0.03 – 0.38) and a similar pattern of poor to moderate (17% – 72%) percentage agreement across all movements. There were no significant changes in kappa (p=0.78) or percentage agreement (p=0.58) across sessions.

ICCs for the MCS composite score, however, indicated excellent reliability for session 2 (0.85) with fair/good reliability in sessions 1 (0.75) and 3 (0.73).

Multiple viewings of videos did not improve inter-rater reliability in novice raters (Table 2, right half). Kappa scores were slight to fair (0 - 0.38) throughout, with a moderate score, for the session 2 left single leg squat (0.43). Indeed, the reliability decreased significantly for session 3 compared to sessions 1 ($p<0.02$) and 2 ($p<0.001$). The percentage agreement scores were poor to moderate (40% - 76%) throughout, and decreased significantly in session 3 compared to session 2 ($p<0.02$). Intra-class correlations of MCS composite scores indicated excellent reliability in session 1 (0.84) and 2 (0.91), with fair/good reliability in session 3 (0.70).

INTRA-RATER RELIABILITY - EXPERTS

The expert intra-rater reliability (Table 3, left half), when comparing session 1v2 and session 2v3 in single views of the performance of the movements in Trial 1, varied between slight, fair or moderate kappa scores, with no significant change from sessions 1v2 to 2v3 (p=0.63). Percentage agreement scores were more consistent, with most MCS scores in the moderate range but excellent scores for single leg squats, and a significant increase from session 1v2 to session 2v3 ($p<0.02$). The ICCs for MCS composite scores again had excellent reliability (0.81 and 0.85). These intra-rater reliability scores were not significantly different (kappa: p=0.07; percentage agreement: p=0.35) from those for the novice raters reported in Table 1 (left half).

INTER-RATER RELIABILITY - EXPERTS

The expert inter-rater reliability (Table 3, right half), in viewing sessions 1, 2 and 3 of Trial 1, varied from slight to fair kappa scores, and one moderate score for the overhead squat in session 3. Scores increased from session 1 to sessions 2 and 3, with session 3 improvement being significant ($p<0.02$). Percentage agreement was poor to moderate, except for the single leg squats which all had excellent agreement. Here there was significant improvement from session 1 to sessions 2 and 3 ($p<0.02$). The ICCs for the MCS composite score displayed fair/good reliability for session 1 (0.71) and excellent reliability in sessions 2 (0.88) and 3 (0.85). These inter-rater reliability scores for sessions 2 and 3 were higher than those for novice raters reported in Table 2 (left half) but the difference was significant only for kappa ($p<0.05$) and not percentage agreement (p=0.20).

DISCUSSION

Strategies to reduce prevalence of sporting musculoskeletal injuries by identifying and improving movement competency of athletes have considerable appeal due to the detrimental social and economic effects of sporting injuries.[25] For such strategies to be successful, the identification of movement competency must be reliable,[7] but also less costly or time-consuming than laboratory processes.[26] This study highlights various factors that must be taken into account by coaches and clinicians when screening athletes.

The composite score, obtained from the sum of scores for the nine movements, whether rated by novices or experts, showed no evidence of improvement across their three performance trials of the MCS. This finding indicates, contrary to previous research by Hansen et al.,[27] that the athletes did not display a familiarization effect when performing a novel task over repeated attempts. The between-study difference here is likely due to differences between the tasks performed. However, given significant differences in composite score were observed between the novice and expert raters, it is suggested that a single rater should conduct repeated measures of the MCS to ensure reliable representation of the athlete's movement competency.

Analysis of assigned MCS composite score did indicate that novice raters tend to score athletes higher than expert raters, suggesting expert raters might be better able to identify errors within an athlete's movements. Furthermore, novice and expert raters showed evidence of a small decrease in assigned MCS scores across repeated viewing of the same movements, suggesting more accurate detection of movement errors in both groups on repeated viewings.

The results for both intra- and inter-rater reliability in this study showed a marked divergence between the consistently high ICCs between the composite scores and the highly variable kappa and percentage agreement scores for individual movements. Intra-rater reliability of composite scores was consistently excellent, with no effects of multiple or repeated viewings, and no differences between novices and experts. Similarly, the inter-rater reliability of composite scores was good to excellent, with no effects of multiple or repeated viewings, and no differences evident between novices and experts. These results for the reliability of the composite score in both experience conditions (i.e. novice and expert) highlight that this MCS (when considering its overall score) may be reliably replicated by both novices and experts in real-time, field-based environments in which the MCS would be typically employed.

In contrast, for the individual movements, the intra-rater reliability was only fair to moderate, showed no difference between novices and experts, showed no improvement with multiple viewings of the same video sessions (task simplifi-

Table 4. Kappa Analysis: Crosstabulation of two raters' scores for a left single leg squat.

		Rater 1		
	Movement Screen Score	1.00	2.00	(*n*)
Rater 2	1.00	48	1	49
	2.00	2	0	2
(*n*)		50	1	51

cation), but did increase with repeated viewings of the same movements (repeat assessments). The inter-rater reliability likewise was poor to moderate for the individual movements and showed no improvement with task simplification of multiple viewings of the same video sessions. It also increased with repeated viewings of the same movements, but only in experts, and showed some evidence of higher reliability in experts than novices.

These findings, that novice and expert raters (Tables 1-3) were not reliable when assessing the individual tasks in the MCS in isolation, was postulated to be due to the complex nature of evaluating multiple features at once, thought to interfere with information processing.[28] Yet this study showed that simplification of the task, by providing multiple viewings and reducing the number of features evaluated at each viewing, did not improve either intra- or inter-rater reliability. Rather, it appears that becoming familiar with the movements over repeated exposure and the errors that can present is the best strategy for ensuring reliability. This was true even for the experts.

It is possible that both the poor reliability for individual tasks in the MCS, and the lack of effect from reducing the complexity of assessment for raters, is confounded by the small scale on which individual tasks were scored.[29] Like the Functional Movement Screen™ (FMS™),[4,5] this study's individual movement scoring system required the movement errors observed to be counted and placed into one of only three categories. Using this three-category scoring system has been suggested to reduce reliability, validity, and discrimination compared to systems with 7 to 10 categories.[29] More categories could be employed by simply counting the errors instead.[30]

In contrast to the ratings of individual movements, the composite score displayed excellent reliability, as indicated by the ICCs for composite scores in both novice and expert raters. This discrepancy in reliability between individual movements and the composite score is likely due to the larger scale of the composite (0-27) compared to that (0-3) for individual movements.[29]

The reduction in statistical power due to the low number of categories was likely to be further confounded by the skewed distribution of the data for each individual movement task.[14] Within this study most athletes displayed numerous movement errors, leading to ratings of category 1 or 2 for most movements. This skewed data distribution reflects the overall poor movement competency of this adolescent athlete cohort, evident in the poor MCS composite scores (mean ± SD for all raters; Session 1, 16/27 ± 4.3, Session 2, 15/27 ± 4.4, Session 3, 15/27 ± 4.1). This data distribution contributed to poor kappa scores, due to the inability to differentiate between random and systematic agreements.[14] For example, as illustrated in Table 4, the single leg squat displayed unequal data distribution, contributing to its low kappa score despite high percentage agreement (Table 2, left side; Table 3, right side). Both raters gave a categorical score of 1 for 48 athletes and only scored a discrepancy for three athletes. It is therefore critical that future research ensure normal distribution of data when assessing the reliability of a MCS.

Several potential limitations of this study exist and must be considered when interpreting the results presented. The principal investigator ensured that all raters were familiar with the movement screening criteria, but because familiarization was an aspect to be analyzed for both raters and athletes through the movement screening process, no formal training was undertaken. It is possible that specific training for both novice and expert raters may have increased the reliability of individual tasks within the movement screen. Volunteer raters had many tasks and athletes to rate that could have led to reduced attention during some rating tasks due to the tedious nature of sessions. Difficulty in recruiting experts for this study led to expert raters rating only a single view session of data trial 1 over 3 sessions. Raters were defined based on their movement screening experience, not on their industry experience, which was not recorded within this study. The movement screening was recorded on two standard video cameras (frontal and sagittal views), meaning it was only possible to watch one view at a time, thereby increasing the time required to carry out each MCS. Since each rater was required to screen each athlete three times, depending on the viewing method, raters took approximately 20-40 hours to screen all participants. Watching the videos of the participants performing the movements may not reflect the real-time field-based assessment typically performed in real-world application. This study design enabled all raters to view the same data to determine rater familiarization and rater reliability of individual movements, which may have caused some confounding between the results for these two outcomes. A lack of evenly distributed individual movement scores within this study may have contributed to the lack of reliability assessed, due to an inability to distinguish between random and systematic agreements in statistical procedures.[14] This study only investigated sub-elite youth athletes and thus the findings of this study cannot be extrapolated to various skill levels, age, as well as sports to see if results can be replicated and generalised to the different population cohorts.

CONCLUSION

Overall results of the current study suggest that the MCS composite score can be reliably used to determine movement competency, but the individual movement scores should not be relied on. It is also recommended that a single rater should conduct any repeated measures of the MCS and the scaling range for individual movement screening scores be increased in future research to obtain more reliable individual movement scores. A familiarization session with MCS movements is not required for athletes when using the MCS composite score. It was identified that expert raters detected more errors than novices overall, however both novice and expert raters improved their detection of movement errors with repeated viewings of the same movement. Therefore, it is recommended that raters familiarize themselves with the MCS.

CONFLICTS OF INTEREST

The authors report no conflicts of interest.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Hunter Academy of Sport for providing access to the participants for this study and the contribution of the raters who generously gave their time to complete the movement screening.

Submitted: November 11, 2021 CDT, Accepted: March 24, 2022 CDT

REFERENCES

1. Mottram S, Comerford M. A new perspective on risk assessment. *Phys Ther Sport*. 2008;9(1):40-51.

2. Mann KJ, Edwards S, Drinkwater EJ, Bird SP. A lower limb assessment tool for athletes at risk of developing patellar tendinopathy. *Med Sci Sports Exerc*. 2013;45(3):527-533.

3. Parkkari J, Kujala UM, Kannus P. Is it possible to prevent sports injuries? Review of controlled clinical trials and recommendations for future work. *Sports Med*. 2001;31(14):985-995.

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. Myer GD, Ford KR, Hewett TE. Tuck jump assessment for reducing anterior cruciate ligament injury risk. *Athlet Ther Today*. 2008;13(5):39-44.

7. Padua DA, Boling MC, DiStefano LJ, Onate JA, Beutler AI, Marshall SW. Reliability of the landing error scoring system-real time, a clinical assessment tool of jump-landing biomechanics. *J Sport Rehabil*. 2011;20:145-156.

8. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Medica*. 2012;22(3):276-282.

9. Tinsley HE, Weiss DJ. Interrater reliability and agreement of subjective judgments. *J Couns Psychol*. 1975;22(4):358.

10. 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.

11. 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.

12. Gribble PA, Brigle J, Pietrosimone BG, Pfile KR, Webster KA. Intrarater reliability of the functional movement screen. *J Strength Cond Res*. 2013;27(4):978-981.

13. Gribble PA, Brigle J, Pietrosimone BG, Pfile KR, Webster KA. Intrarater reliability of the functional movement screen. *J Strength Cond Res*. 2013;27(4):978-981.

14. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257-268.

15. Teyhen DS, Shaffer SW, Lorenson CL, et al. The functional movement screen: A reliability study. *J Orthop Sports Phys Ther*. 2012;42(6):530-540.

16. 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.

17. Perrott MA, Pizzari T, Opar M, Cook J. Development of clinical rating criteria for tests of lumbopelvic stability. *Rehabil Res Pract*. 2011;2012:7.

18. Kritz M, Cronin J, Hume P. Using the Body Weight Forward Lunge to Screen an Athlete's Lunge Pattern. *Strength & Conditioning Journal*. 2009;31(6):15.

19. De Blaiser C, De Ridder R, Willems T, et al. Evaluating abdominal core muscle fatigue: Assessment of the validity and reliability of the prone bridging test. *Scand J Med Sci Spor*. 2018;28(2):391-399.

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

21. Birkimer JC, Brown JH. Back to basics: Percentage agreement measures are adequate, but there are easier ways. *J Appl Behav Anal*. 1979;12(4):535-543.

22. Baumgartner TA, Strong CH, Hensley LD. *Conducting and Reading Research in Health and Human Performance (3rd Ed.)*. McGraw-Hill; 2006.

23. Fleiss JL. Reliability of measurement. In: *The Design and Analysis of Clinical Experiments*. John Wiley and Sons, Inc; 1986:2-32.

24. Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychol Bull*. 1979;86(2):376.

25. Caine D, Maffulli N, Caine C. Epidemiology of injury in child and adolescent sports: injury rates, risk factors, and prevention. *Clinics in Sports Medicine*. 2008;27(1):19-50.

26. McLean S, Walker K, Ford K, Myer G, Hewett T, Van Den Bogert A. Evaluation of a two dimensional analysis method as a screening and evaluation tool for anterior cruciate ligament injury. *Brit J Sport Med*. 2005;39(6):355-362.

27. Hansen M, Dieckmann B, Jensen K, Jakobsen B. The reliability of balance tests performed on the kinesthetic ability trainer (KAT 2000). *Knee Surg Sport Tr A*. 2000;8(3):180-185.

28. Whiteside D, Deneweth JM, Pohorence MA, et al. Grading the functional movement screen: A comparison of manual (real-time) and objective methods. *J Strength Cond Res*. 2016;30(4):924-933.

29. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol*. 2000;104(1):1-15.

30. Perrott M, Pizzari T, Cook J. Assessment of lumbopelvic stability: Beyond a three-point rating scale. *J Sci Med Sport*. 2017;20:25-26.