Journal of Applied Biomechanics, 2019, 35, 32-36
https://doi.org/10.1123/jab.2018-0056

Human Kinetics
ORIGINAL RESEARCH

Validity of an Automated Balance Error Scoring System

Stephen M. Glass,[1] Alessandro Napoli,[2] Elizabeth D. Thompson,[3] Iyad Obeid,[2] and Carole A. Tucker[2,3]
[1]Ohio State University Wexner Medical Center; [2]Temple University College of Engineering;
[3]Temple University College of Public Health

The balance error scoring system (BESS) is a human-scored, field-based balance test used in cases of suspected concussion. Recently developed instrumented alternatives to human scoring carry substantial advantages over traditional testing, but thus far report relatively abstract outcomes that may not be useful to clinicians or coaches. In contrast, the automated assessment of postural stability (AAPS) is a computerized system that tabulates error events in accordance with the original description of the BESS. This study compared AAPS and human-based BESS scores. A total of 25 healthy adults performed the modified BESS. Tests were scored twice each by 3 human raters and the computerized system. Interrater (between human) and intermethod (AAPS vs human) agreement (interclass correlation coefficient$_{2,1}$) were calculated alongside Bland–Altman limits of agreement. Interrater analyses were significant ($P < .01$) and demonstrated good to excellent agreement. Intermethod agreement analyses were significant ($P < .01$), with agreement ranging from poor to excellent. Computerized scores were equivalent across rating occasions. Limits of agreement ranges for AAPS versus the human average exceeded the average limits of agreement ranges between human raters. Coaches and clinicians may consider a system such as AAPS to automate balance testing while maintaining the familiarity of human-based scoring, although scores should not yet be considered interchangeable with those of a human rater.

Keywords: kinematics, clinical biomechanics, motion analysis, motor behavior, sports medicine

Static postural control assessed during quiet standing is commonly used as an indicator of injury and recovery status in cases of suspected concussion or mild traumatic brain injury.[1,2] Quantitative assessment of postural control is perhaps best achieved through laboratory-grade instrumentation and testing protocols,[3,4] but such methods are seldom used for wide-scale or field-based testing due to their prohibitive cost and lack of accessibility. Where laboratory methods are impractical, field-based tests may be used in their place.

The balance error scoring system (BESS)[5] and its variants (eg, modified BESS[6] [mBESS]) are among the most familiar field-based tests of balance. Although these methods address the feasibility limitations associated with laboratory-based measurement, the quantity and quality of data that can be captured by humans are limited and inherently subjective. Such limitations pose a logistical barrier to the effective use of BESS tests as screening tools, particularly in high-volume settings where human resources are limited.[7,8] An automated scoring system relying on portable, low-cost instrumentation could overcome these limitations while maintaining applicability for field and clinical use.

To date, instrumented alternatives to standard BESS or mBESS[9–14] testing have produced novel, signal-based outcomes representing scales, which are (1) unfamiliar to most end users and (2) not directly comparable to the more common BESS scores. To address these issues, we developed the automated assessment of postural stability (AAPS). This system automates BESS administration and scoring using a low-cost, mass-produced sensor to quantify 3-dimensional kinematic motion.[15] Importantly, the outcomes generated by our system are the same as those made familiar by nearly 20 years of clinical and scientific use of the BESS[5]—specifically, error counts per testing condition.

The purpose of this research was to study the agreement of the AAPS system with standard clinician-based scoring of the BESS. To characterize AAPS performance in reference to the criterion of scoring by trained observers, we examined interrater (human vs human) and intersystem (AAPS vs human) agreement. We hypothesized that (1) interrater agreement of (human) BESS scoring would range from good (interclass correlation coefficient [ICC] = .6 to .75, the upper portion of Fleiss' "fair to good" range) to excellent (ICC > .75) as reported in previous work[16] and (2) intermethod agreement (AAPS vs human BESS scoring) would similarly range from good to excellent. Finally, we additionally sought to determine whether AAPS could be considered interchangeable with humans as a method for scoring BESS tests.

Methods

Subjects

A total of 25 healthy participants (13 females: 25.57 [3.13] y, 167.64 [6.72] cm, 67.13 [17.21] kg; 12 males: 24.77 [3.81] y, 180.34 [6.22] cm, 84.87 [14.89] kg) were recruited to participate in this study. The protocol was approved by the Temple University institutional review board. All participants provided written, informed consent prior to participating.

Procedures

The mBESS—excluding the double-leg stance trials—was administered and scored in accordance with previously published procedures.[6,17] This mBESS was specifically chosen considering the lack of information gained from the double-leg standing conditions with this population.[6,18] Testing was performed in a quiet room

Glass is with the Department of Otolaryngology, The Ohio State University, Columbus, OH. Napoli, Obeid, and Tucker are with the Department of Electrical and Computer Engineering, Temple University College of Engineering, Philadelphia, PA. Thompson and Tucker are with the Department of Physical Therapy, Temple University College of Public Health, Philadelphia, PA. Glass (stephen.glass@osumc.edu) is corresponding author.

with minimal distractions. Participants completed each of the conditions in the following order: (1) single-leg stance/stable surface (Single-Stable), (2) tandem stance/stable surface (Tandem-Stable), (3) single-leg stance/foam surface (Single-Foam), and (4) tandem stance/foam surface (Tandem-Foam). All trials were 20 seconds in duration and were performed with hands on hips and eyes closed facing the sensor. Single-leg trials were performed on the nondominant limb, while the dominant limb was held in 20° of flexion at both the hip and knee. Tandem trials were performed with feet positioned in-line (heel-to-toe) with the nondominant limb placed in the back.

Participants were provided standard instructions to maintain each of the testing positions to the best of their ability for each trial and to reassume the testing position as quickly as possible should they commit an error. The standard following errors were used in scoring: (1) removing the hands from the hips, (2) opening the eyes, (3) step, stumble, or fall, (4) flexion/abduction of the hip in excess of 30°, (5) lifting/moving of the feet, and (6) remaining out of the testing position for more than 5 seconds.[5]

Instrumentation

All testing was conducted in view of a Microsoft Kinect 2.0™ (Microsoft Corp., Redmond, WA) gaming device secured to a tripod and placed 1.37 m above the ground and facing the participant at a distance of 3.00 m. The Kinect 2.0™ integrates depth, color, and infrared data streams to render 3-dimensional joint positions using Microsoft's proprietary algorithm at a maximum frame rate of 30 Hz. The exact sampling frequency cannot be controlled directly; however, data collection for a given trial was terminated if the frame rate dropped below 12 Hz.

The Kinect™ sensor was interfaced with a Dell Latitude PC (Dell, Inc. Round Rock, TX) (Windows 10 64-bit, 2.6-GHz Core i7 processor, 8-GB RAM, and 500-GB solid-state drive) through a customized C# application developed using the Microsoft Kinect SDK 2.0 libraries.[15] Each raw BESS trial video was 26 seconds in duration. A 20-second interval for testing was identified beginning with the time stamp of the first frame in which the participant met all the conditions of the required stance position (eg, eyes closed, hands on hips, vertical alignment, and proper foot placement). This frame, and the corresponding time, was identified manually for each video by S.M.G. Once these frames were identified, a truncated video and .csv time series file were generated for human and computer scoring, respectively.

Outcomes

The Kinect™ video data were evaluated twice by each of 3 human raters who were experienced with BESS test administration: 2 physical therapists (each with >10 y of experience, including BESS experience deriving from research, instruction, and musculoskeletal screening) and 1 personal trainer (>10 y of experience, with >5 y of experience scoring and analyzing the BESS for research). Prior to the initial video viewing, the 3 raters conferred to review the BESS scoring criteria. Rating sessions were separated by a minimum of 2 weeks and raters were blinded to (1) each other's scores and (2) their own scores from the first rating sessions.

All trials were also scored twice by the AAPS system. The AAPS-based automated error detection used Kinect® joint position and eye data, exported as .csv files, to determine whether movements related to any error categories exceeded a baseline threshold established over a 1-second pretest observation window. Thresholds were defined by 3-dimensional distances between the left and right centers of the wrist, elbow, ankle, and knee joints; frontal/sagittal trunk and hip angles; and left and right forefoot segments. (For further technical details regarding automated error tabulation, please see our previous work.[15])

Statistics

We calculated interrater (between human) and intermethod (human vs AAPS) correlation coefficients ($ICC_{2,1}$ for absolute agreement)[19] using the psych[20] package in R-3.4.1 (The R Foundation, Vienna, Austria). For intermethod ICC, the average score among the 3 raters was used for the human component of the analysis. ICCs were interpreted using the following guidelines modified from Fleiss[16]: .00 to .40 (poor), .40 to .59 (fair), .60 to .74 (good), and .75 to 1.00 (excellent). Finally, we calculated Bland–Altman 95% limits of agreement[21] (LOA) for AAPS versus the human average and compared it to the average of 3 between-human LOA ranges (1 vs 2, 2 vs 3, and 1 vs 3) to assess whether AAPS can be used interchangeably with human BESS rating. For the latter, human scores were averaged across day prior to determining LOA.

Results

With 2 exceptions, BESS score agreement between human raters was excellent (ICC > .75). The 2 exceptions were Single-Foam on day 1 and Tandem-Foam on day 2, for which agreement was good in both cases. All interrater agreement analyses reached the threshold for statistical significance (Table 1).

Table 1 Interrater and Intermethod ICCs

Condition	ICC (CI_{95})	F_{df}	*P* value	LOA
Between human day 1				
SS	.86 (.76 to .93)	$20.49_{24,\ 48}$	<.001*	−2.16 to 1.73
TS	.77 (.56 to .89)	$14.38_{23,\ 46}$	<.001*	−1.24 to 0.86
SF	.78 (.61 to .89)	$14.01_{23,\ 46}$	<.001*	−3.19 to 2.13
TF	.65 (.44 to .81)	$7.12_{24,\ 48}$	<.001*	−3.75 to 2.79
Total	.83 (.68 to .92)	$19.79_{22,\ 44}$	<.001*	−6.91 to 4.12
Between human day 2				
SS	.96 (.92 to .98)	$68.80_{24,\ 48}$	<.001*	−1.14 to 0.87
TS	.86 (.74 to .93)	$19.11_{23,\ 46}$	<.001*	−0.94 to 0.77
SF	.73 (.54 to .86)	$8.89_{23,\ 46}$	<.001*	−3.10 to 2.77
TF	.79 (.64 to .90)	$12.46_{23,\ 46}$	<.001*	−2.58 to 2.15
Total	.86 (.75 to .94)	$20.48_{21,\ 42}$	<.001*	−5.24 to 3.97
AAPS vs human average				
SS	.81 (.63 to .91)	$9.98_{24,\ 24}$	<.001*	−2.54 to 1.96
TS	.44 (.06 to .71)	$2.52_{24,\ 24}$	<.01*	−1.89 to 1.67
SF	.38 (−.02 to .68)	$2.21_{23,\ 23}$	<.03*	−4.21 to 4.94
TF	.72 (.47 to .87)	$6.11_{24,\ 24}$	<.001*	−3.10 to 2.74
Total	.74 (.47 to .88)	$6.48_{21,\ 21}$	<.001*	−7.23 to 6.77

Abbreviations: CI, confidence interval; ICC, interclass correlation coefficient; LOA, limits of agreement; SF, Single-Foam; SS, Single-Solid; TF, Tandem-Foam; TS, Tandem-Solid. Note: BESS score agreement between human raters for day 1 (top) and day 2 (middle). Agreement for AAPS versus human raters is shown on the bottom, where human rater scores are collapsed across day and rater.
*Indicates statistical significance.

Intermethod agreement was excellent for Single-Solid, good for Tandem-Foam and the total BESS score, fair for Tandem-Stable, and poor for Single-Foam. All intermethod agreement analyses reached the threshold for statistical significance (Table 1).

Finally, LOA ranges for AAPS versus the human average exceeded the average LOA ranges between human raters. Disparities in LOA ranges were greatest in the Single-Foam condition and least in the Tandem-Stable condition (Table 1). For visualization, we present Bland–Altman plots for AAPS versus the human average alongside the 2 most consistent human raters in our data set for the stable conditions (Figure 1) and foam conditions (Figure 2).

Discussion

The purpose of this research was to study the agreement of an automated BESS scoring system (AAPS) with the existing standard of human observation. We found similar interrater agreement for BESS item and total scores as has been reported previously.[5,22] By comparison, our intermethod analyses suggest that agreement between human-derived and AAPS-derived BESS scores was lower than that between human raters, with 2 of the conditions (Single-Foam and Tandem-Stable) failing to reach the hypothesized "good" level of AAPS-Human agreement (ICC). Finally, although we are encouraged by these initial findings and the qualitative similarity of the aggregate data (Supplementary Data [available online]), 95% LOA ranges suggest further work is required before our system can be considered interchangeable with a human rater for purposes of registering BESS error behaviors in clinical settings.

Efficient collection of high-quality data has been identified as a major concern for organizations conducting large-scale medical or preparticipation exams.[7,8,23] Although not a replacement for a trained human BESS scorer, AAPS is 100% consistent in arriving at error scores and catalogs behavior in far greater detail than is possible through human observation. Thus, a system such as ours could be used to automate balance testing procedures in large-scale organizations, potentially increasing testing efficiency, while at the same time addressing limitations related to interrater or intrarater measurement consistency.[24]

In addition to automating the BESS, the present system does so with less abstraction than other methods in which kinematic features are used as a correlate of BESS scores.[10,13] That is, AAPS quantifies balance performance in the original BESS unit of error counts as opposed to a novel metric. The intention with this strategy was to present the clinician or researcher with information that is not only intuitively meaningful but which is readily compared with previously published normative data.[25] Creating such a system, however, requires finding a balance between the relative strengths of human and machine observation.

The BESS was originally designed to provide a tool that could be used by humans to assess balance in nonlaboratory settings.[5] As a human-friendly tool, there are some areas in which the scoring criteria favor human intuition. For example, humans may be able to separate a single prolonged error from multiple distinct errors using intuition that cannot currently be coded. This bias in favor of human intuition comes at the cost of limitations on human focus and multitasking. The BESS scoring criteria do not require the

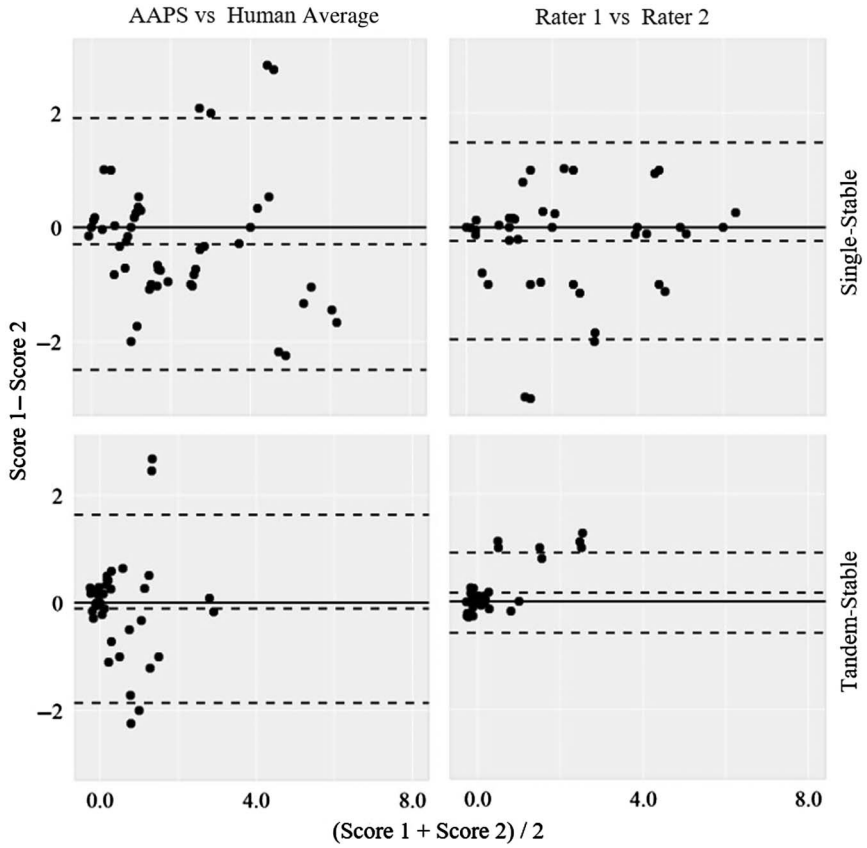

Figure 1 — Bland–Altman limits of agreement plots in stable stance conditions for (1) AAPS versus the human average collapsed across day and rater (left), and (2) for rater 1 versus rater 2 (right), which represents the highest level of between-human agreement observed in our data set. Here, rater 1 and rater 2 scores are averaged across day. AAPS indicates automated assessment of postural stability.

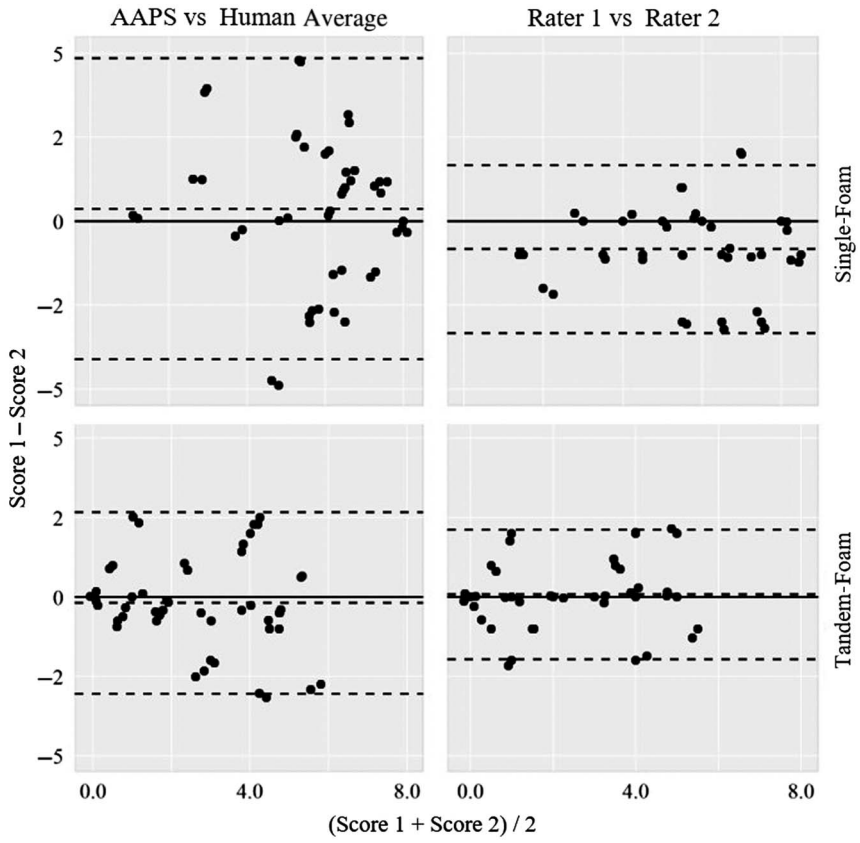

Figure 2 — Bland–Altman limits of agreement plots in Foam stance conditions for (1) AAPS versus the human average collapsed across day and rater (left), and (2) for rater 1 versus rater 2 (right), which represents the highest level of between-human agreement observed in our data set. Here, rater 1 and rater 2 scores are averaged across day. AAPS indicates automated assessment of postural stability.

observer to log simultaneous error events, error event types, or times of error occurrence. Arguably, this is not because such information is unimportant, but rather because it is not feasibly monitored by a human. So, although there may be limits to how precisely an automated system can distinguish errors from non-errors, the user has access to information that is more consistent and more descriptive (eg, error type, time of occurrence, and magnitude) than human judgment.

The results of this study are limited to the sample of balance participants used for our analyses. Further work will be required to cross-validate the error detection algorithm with different samples. Another limitation concerns the content of the errors detected. Because BESS trial scores from human raters present only the total error count, we cannot determine the extent to which our algorithm registers the same error events observed by a human rater. This limitation is currently being addressed with the development of a real-time annotation program, which will aid human raters in producing labeled trials.

In conclusion, the AAPS system provides a low-cost method of quantifying balance performance, requiring only a laptop and a Kinect™ for administration. (Future releases will support use with other RGB-D sensors.) It is currently available for research use. AAPS can be applied to automate balance test administration in high-volume screening settings or others in which human rating is inefficient and/or cost-prohibitive, although error scores from the current algorithm are not yet interchangeable with BESS scores from a trained human rater. Forthcoming work from our laboratory will describe the system's additional features, as well as robustness to sources of noise that might be encountered during field use.

Acknowledgment

The authors have no conflicts of interest to disclose.

References

1. Guskiewicz KM. Postural stability assessment following concussion: one piece of the puzzle. *Clin J Sport Med*. 2001;11(3):182–189. PubMed ID: 11495323 doi:10.1097/00042752-200107000-00009
2. Guskiewicz KM, Ross SE, Marshall SW. Postural stability and neuropsychological deficits after concussion in collegiate athletes. *J Athl Train*. 2001;36(3):263–273. PubMed ID: 12937495
3. Peterka RJ, Black FO. Age-related changes in human posture control: sensory organization tests. *J Vestib Res*. 1989;1(1):73–85.
4. Nashner LM, Black FO, Wall C 3rd. Adaptation to altered support and visual conditions during stance: patients with vestibular deficits. *J Neurosci*. 1982;2(5):536–544. PubMed ID: 6978930 doi:10.1523/JNEUROSCI.02-05-00536.1982
5. Riemann BL, Guskiewicz KM, Shields EW. Relationship between clinical and forceplate measures of postural stability. *J Sport Rehabil*. 1999;8(2):71–82. doi:10.1123/jsr.8.2.71
6. Hunt TN, Ferrara MS, Bornstein RA, Baumgartner TA. The reliability of the modified balance error scoring system. *Clin J Sport*

Med. 2009;19(6):471–475. PubMed ID: 19898074 doi:10.1097/JSM.0b013e3181c12c7b
7. Teyhen D, Bergeron MF, Deuster P, et al. Consortium for health and military performance and American College of Sports Medicine Summit: utility of functional movement assessment in identifying musculoskeletal injury risk. *Curr Sports Med Rep*. 2014;13(1):52–63. PubMed ID: 24412892 doi:10.1249/JSR.0000000000000023
8. Teyhen DS, Shaffer SW, Umlauf JA, et al. Automation to improve efficiency of field expedient injury prediction screening. *J Strength Cond Res*. 2012;26(suppl 2):S61–S72. doi:10.1519/JSC.0b013e31825d80e6
9. Goble DJ, Manyak KA, Abdenour TE, Rauh MJ, Baweja HS. An initial evaluation of the btracks balance plate and sports balance software for concussion diagnosis. *Int J Sports Phys Ther*. 2016;11(2):149–155. PubMed ID: 27104048
10. Alberts JL, Thota A, Hirsch J, et al. Quantification of the balance error scoring system with mobile technology. *Med Sci Sports Exerc*. 2015;47(10):2233–2240. PubMed ID: 26378948 doi:10.1249/MSS.0000000000000656
11. Alsalaheen BA, Haines J, Yorke A, Stockdale K, Broglio SP. Reliability and concurrent validity of instrumented balance error scoring system using a portable force plate system. *Phys Sportsmed*. 2015;43(3):221–226. PubMed ID: 26109242 doi:10.1080/00913847.2015.1040717
12. Amick RZ, Jansen SD, Chaparro A, Hakansson NA, Patterson JA, Jorgensen MJ. Comparison of the SWAY balance mobile application to the abbreviated balance error scoring system. *Athl Train Sports Health Care*. 2015;7(3):89–96. doi:10.3928/19425864-20150422-04
13. Brown HJ, Siegmund GP, Guskiewicz KM, Van Den Doel K, Cretu E, Blouin JS. Development and validation of an objective balance error scoring system. *Med Sci Sports Exerc*. 2014;46(8):1610–1616. PubMed ID: 24500539 doi:10.1249/MSS.0000000000000263
14. Alberts JL, Hirsch JR, Koop MM, et al. Using accelerometer and gyroscopic measures to quantify postural stability. *J Athl Train*. 2015;50(6):578–588. PubMed ID: 25844853 doi:10.4085/1062-6050-50.2.01
15. Napoli A, Glass SM, Tucker C, Obeid I. The automated assessment of postural stability: balance detection algorithm. *Ann Biomed Eng*. 2017;45(12):2784–2793. PubMed ID: 28856486 doi:10.1007/s10439-017-1911-8
16. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York, NY: Wiley; 1999.
17. Riemann BL, Guskiewicz KM. Effects of mild head injury on postural stability as measured through clinical balance testing. *J Athl Train*. 2000;35(1):19–25. PubMed ID: 16558603
18. Valovich McLeod TC, Perrin DH, Guskiewicz KM, Shultz SJ, Diamond R, Gansneder BM. Serial administration of clinical concussion assessments and learning effects in healthy young athletes. *Clin J Sport Med*. 2004;14(5):287–295. PubMed ID: 15377968 doi:10.1097/00042752-200409000-00007
19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428. PubMed ID: 18839484 doi:10.1037/0033-2909.86.2.420
20. Revelle W. *Procedures for Personality and Psychological Research*. Evanston, IL: Northwestern University; 2015.
21. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–160. PubMed ID: 10501650 doi:10.1177/096228029900800204
22. Valovich McLeod TC, Barr WB, McCrea M, Guskiewicz KM. Psychometric and measurement properties of concussion assessment tools in youth sports. *J Athl Train*. 2006;41(4):399–408. PubMed ID: 17273465
23. Teyhen DS, Shaffer SW, Butler RJ, et al. Application of athletic movement tests that predict injury risk in a military population: development of normative data. *Mil Med*. 2016;181(10):1324–1334. PubMed ID: 27753571 doi:10.7205/MILMED-D-15-00297
24. Finnoff JT, Peterson VJ, Hollman JH, Smith J. Intrarater and interrater reliability of the Balance Error Scoring System (BESS). *PM R*. 2009;1(1):50–54. doi:10.1016/j.pmrj.2008.06.002
25. Iverson GL, Koehle MS. Normative data for the balance error scoring system in adults. *Rehabil Res Pract*. 2013;2013:846418. PubMed ID: 23577257