Generalizability Theory Analysis of Balance Error Scoring System Reliability in Healthy Young Adults

Steven P. Broglio, PhD, ATC; Weimo Zhu, PhD; Kay Sopiarz; Youngsik Park, MS

University of Illinois at Urbana-Champaign, Urbana, IL

Context: An assessment of postural control is commonly included in the clinical concussion evaluation. Previous investigators have demonstrated learning effects that may mask concussion-induced balance decrements.

Objective: To establish the test-retest reliability of the Balance Error Scoring System (BESS) and to provide recommendations that account for known learning effects.

Design: Test-retest generalizability study.

Setting: Balance research laboratory.

Patients or Other Participants: Young adults (n = 48) free from injuries and illnesses known to affect balance.

Intervention(s): Each participant completed 5 BESS trials on each of the assessment dates, which were separated by 50 days.

Main Outcome Measure(s): Total score of the BESS was used in a generalizability theory analysis to estimate the overall reliability of the BESS and that of each facet. A decision study was completed to estimate the number of days and trials needed to establish clinical reliability.

Results: The overall reliability of the BESS was G = 0.64. The test-retest reliability was improved when male (0.92) and female (0.91) participants were examined independently. Clinically acceptable reliability (greater than 0.80) was established when 3 BESS trials were administered in a single day or 2 trials were administered at different time points.

Conclusions: Learning effects have been noted in individuals with no previous exposure to the BESS. Our findings indicate that clinicians should consider interpreting the mean score from 3 BESS administrations on a given occasion for both normative data comparison and pretest and posttest design. The multiple assessment technique yields clinically reliable scores and provides the sports medicine practitioner with accurate data for clinical decision making.

Key Words: learning effects, concussions, postural control

Key Points

- Postural control is typically assessed as part of a concussion evaluation. However, learning effects may obscure balance deficits resulting from concussion.
- Clinically acceptable reliability (greater than 0.80) was attained when the Balance Error Scoring System was administered 3 times on a single day or 2 times on different days.

Current estimates[1] of sport-related concussion indicate that 1.6 to 3.8 million injuries occur in the United States on an annual basis. The true injury rate is likely much higher because more than 50% of interscholastic football athletes do not report their injuries to medical personnel.[2] Assessing an athlete with a suspected concussion is particularly difficult for the sports medicine professional as a result of the array of clinical outcomes associated with the injury and factors known to affect them. For example, the athlete's age[3] and sex[4] and the location and magnitude of impact[5] may all influence injury severity and recovery. To best control for these variables when identifying concussed athletes, sports medicine organizations advocate using a battery of tests that evaluate multiple aspects of cognitive functioning known to be affected by the injury.[6,7] Concussion has large negative effects on measures of self-reported symptoms, neurocognitive functioning, and postural control,[8] with each aspect of the test battery providing unique information that supports the clinical examination.[9]

One postural control test developed with the explicit intent of concussion assessment is the Balance Error Scoring System (BESS). The BESS was created to provide objective postural control information to the clinician without the need for expensive equipment or extensive training.[10] The test uses 6 testing conditions under which the administrator counts "errors" that represent poor postural control. Five of the 6 test conditions generated significant correlations with an instrumented balance assessment,[10] and on the BESS, concussed athletes followed a recovery pattern similar to that seen on the instrumented NeuroCom Sensory Organization Test.[11,12] Multiple administrations of the BESS, however, may result in improved balance performance (ie, fewer errors), which has been associated with learning effects.[13] As a result, the test-retest reliability (intraclass correlation coefficient [ICC]) of the BESS total error score has been reported at 0.70 in 9-year-old to 14-year-old youths.[14] This reliability is considered less than ideal for clinical applicability (ie, less than 0.80),[15] but the immature postural control mechanism of the young participants[16] and the athlete's sex may have influenced the findings.[16,17] Further, although the use of ICCs is a widely accepted measure of score reliability,[18] this statistical technique can only identify and quantify the total variance within the system. A generalizability study (G study), however, offers the advantage of describing the variance associated with each aspect of the assessment.[15] This technique is particularly useful when determining

which components of a test are adding to the total variance, and the technique establishes the level at which a measurement sample generalizes to all assessments.

Relative to the concussion assessment, a baseline measure is commonly administered at the beginning of the athletic season and is then used as a reference point, if needed, in diagnosing the injury during the season. When using this type of protocol, we must consider 2 forms of reliability. First, the reliability of the baseline measure must be established. Without this information it will be difficult to determine if the difference in the postmorbid measurement is due to the head injury or is simply a random error. This kind of reliability has been called *single test administration reliability* and is established by a single test administration.[19] Second, the stability of the trait (ie, balance) being measured has to be determined. Then a difference observed during the season can be identified as either a true change caused by injury or simply as growth or decline over time. This kind of reliability has been called *stability* and is established by a test-retest design. Using a G study, variance from both reliabilities (ie, administration reliability and stability) can be examined simultaneously. Evaluation of these and other components (ie, facets) permits the investigator to identify which factors provide measurement variance. Manipulating those components that induce the greatest variance into the system through a decision study (D study) can help to determine a score's reliability when it is modified by increasing or decreasing test components, test administrations, or other factors. By altering 1 or more components, greater reliability of the measurement can be established. A complete description of generalizability analysis has been provided elsewhere.[20–22]

Therefore, to better understand the reliability of the BESS in young adults, the purposes of this study were to determine the sources of variance within the BESS and to estimate the number of test administrations or trials needed to establish a clinically acceptable level of test-retest reliability. A secondary purpose of the study was to examine the effect of each participant's sex.

METHODS

A fully crossed design was used, in which all participants tested were crossed with all the conditions examined. A total of 48 young adults (age = 20.42 ± 2.08 years, height = 169.67 ± 9.60 cm, mass = 72.56 ± 12.77 kg) volunteered for this study. Before testing, participants read and signed a university-approved informed consent document and had their height and mass recorded. All volunteers indicated that they were unfamiliar with the BESS and that their balance was unaffected by lower extremity injuries, medical conditions, or medications known to affect postural control. After an explanation of the BESS, each participant was then administered the initial (baseline) BESS evaluation. The participants returned approximately 50 days later (posttest) for a follow-up evaluation and again indicated they were free from balance problems. During the baseline and posttest days, the participant completed 5 consecutive BESS assessments. Approximately 10 to 20 seconds separated BESS conditions, and BESS trials were separated by 2 to 3 minutes. The 50-day interval was selected to reflect the mean time between a baseline evaluation and postmorbid concussion assessment reported in an athletic environment.[23]
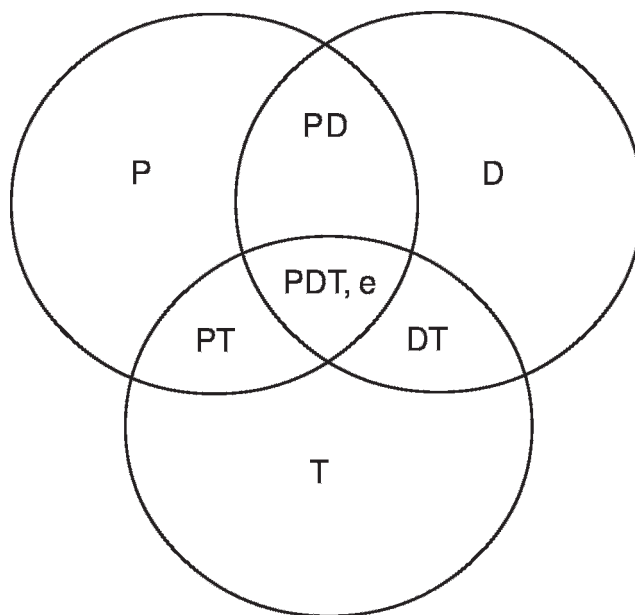
Figure 1. Variance components obtained by the fully crossed design of participants (P), days (D), and trials (T). Abbreviation: e, error.

The BESS was administered as previously described.[10] Briefly, the test requires the participant to complete 6 conditions consisting of 3 stances (double leg, single leg, tandem stance) on both firm and compliant surfaces (Balance Pad, Alcan Airex, Switzerland). Once the participant placed his or her hands on the iliac crests and closed his or her eyes, the investigator recorded countable errors during the 20-second trial. An error was indicated when the participant removed the hands from the iliac crests, opened the eyes, took a step, stumbled, abducted or flexed the hip more than 30°, lifted the forefoot or heel off the ground, or remained out of the test position for longer than 5 seconds. Before testing, all participants were informed of the errors and were asked to stand as motionless as possible once in the test position. The intertester reliability of the investigators was determined to be 0.92 during pilot testing and is similar to that in previous reports.[10]

We analyzed the data with a 2 × 10 (sex × time) repeated-measures analysis of variance to discern changes in BESS total error scores between the sexes and among the test points. When main effects were indicated, all possible pairwise comparisons were performed, with a Bonferroni adjustment for familywise error. Significance was noted when $P < .05$, and the analysis was completed using SPSS (version 14.0; SPSS Inc, Chicago, IL).

The data were then analyzed using GENOVA (version 3.1; The University of Iowa, Iowa City, IA), software for generalizability theory analysis.[21] A 3-facet (participant × day × trial) model was defined, and the facets of day (D) and trial (T) were set as random. Variance components obtained by this fully crossed design of participants, days, and trials are summarized in a Venn diagram (Figure 1). To determine the combination of day × trial that generated the most reliable measure, models included 1 and 2 assessment days, as well as up to 10 BESS trials in a given day. To examine the effect of sex, the analyses were conducted for all participants and for male and female participants only.
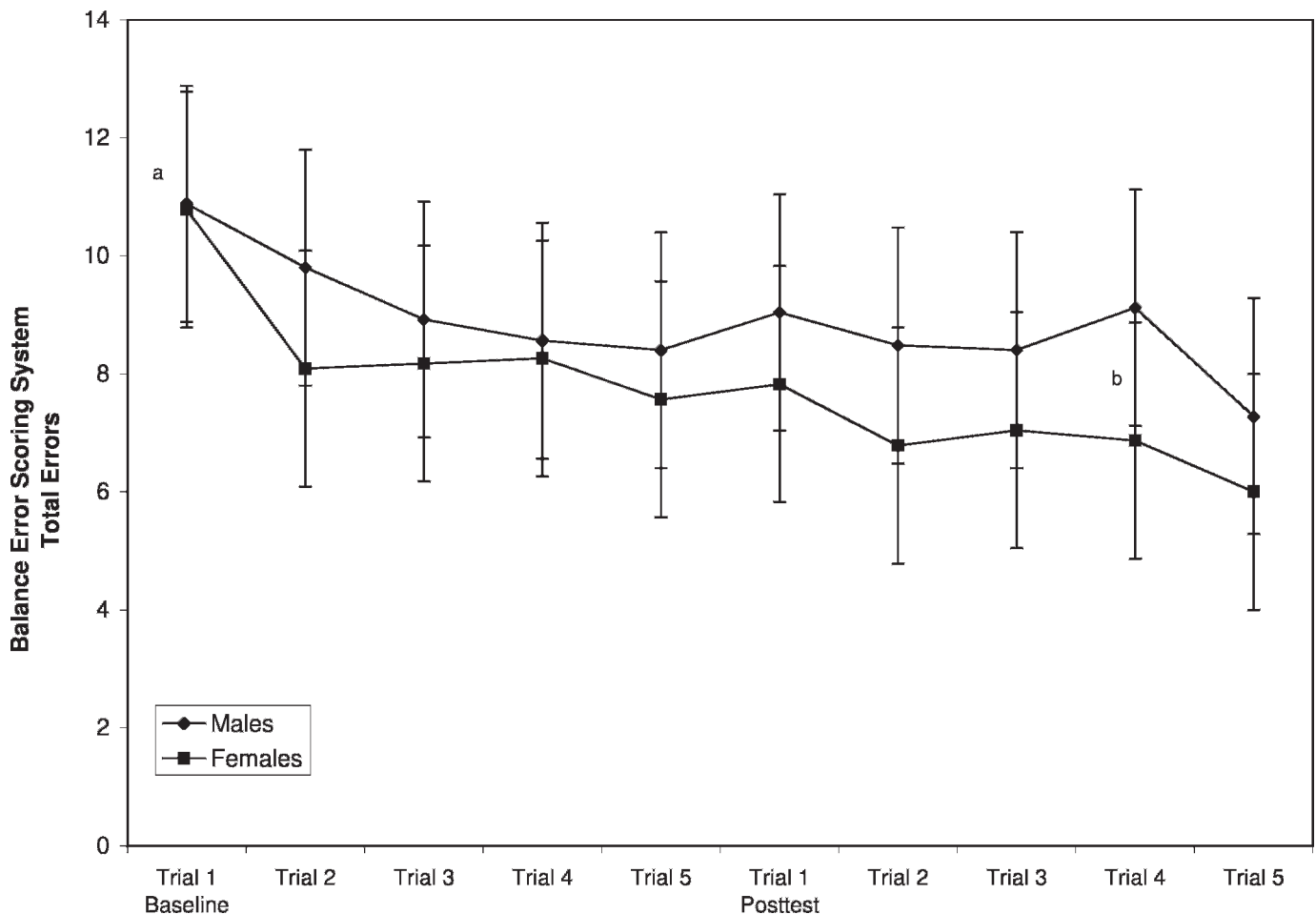
Figure 2. Balance Error Scoring System performance by sex at baseline and posttest for each test administration (trial). [a] Indicates difference between baseline and trial. [b] Indicates difference from all other 10 tests except for baseline, trial 5, and posttest, trials 2 and 3.

RESULTS

The study cohort consisted of 25 male participants (age = 20.88 ± 2.74 years, height = 176.23 ± 7.25 cm, mass = 79.90 ± 10.74 kg) and 23 female participants (age = 19.91 ± 0.73 years, height = 162.53 ± 6.16 cm, mass = 64.59 ± 9.75 kg). Participants were evaluated an average of 49.42 ± 1.29 days apart. Mean performance on each BESS trial for each group is provided in Figure 2. The sex $\times$ time interaction ($F_{9,414} = 1.21$, $P = .29$, $\eta^2_{partial} = 0.03$) and the sex main effect ($F_{1,46} = 1.90$, $P = .18$, $\eta^2_{partial} = 0.04$) were nonsignificant. The time main effect was significant ($F_{9,414} = 12.87$, $P < .01$). Post hoc analyses indicated that the participants' performance on the baseline trial 1 was greater than that on all other trials ($P < .05$) and that performance on posttest trial 5 was lower ($P < .05$) than that of all other performances except for baseline trial 5 and posttest trials 2 and 3.

The G study results, including the variance components and their contributions to the total variance, are shown in Table 1. When all participants were analyzed together, the G coefficient was estimated at 0.64. When the sources of variance were evaluated, participants, days, and trials each contributed approximately 10%. The interaction between any 2 facets was also small, but the greatest variance determinate was in the participant $\times$ day $\times$ trial interaction (56.09%). Because concussion assessments cannot be generalized but rather are specific to the individual, further analyses were conducted when the data were separated by participant sex.

The sex analyses revealed an increased G coefficient, with the reliabilities rising to 0.92 and 0.91 for males and females, respectively. The individual contribution of variance by day and trial continued to be small (less than 10%). The sex-separated participant variance, however, accounted for the largest source of variance. Of the total variance, males accounted for 51.15% and females for 59.03%. Interactions among participants, day, and trial were also small (Table 1).

The D study results show that reliability of the BESS improved as the number of trials administered increased (Table 2). A similar trend in trial performance was observed in both male and female participants. When 1 day of BESS testing was considered, 3 administrations provided acceptable reliability. Under these conditions, the G coefficient was 0.81 for the males and 0.79 for the females. When 2 days of testing were evaluated, with 2 BESS trials at each time point, reliability was sufficient (G coefficient: males = 0.85, females = 0.82).

DISCUSSION

The BESS assesses postural control to generate objective information that supports the clinical concussion evalua-

Table 1. Generalizability Study Variance and Percentage of Total Variance for Each Test Component (Participant, Day, Trial)

Effect	df	Facet Variance	Total Variance, %	Standard Error
All participants (n = 48)				
Participant (P)	47	2.52	9.27	0.90
Day (D)	1	2.76	10.16	2.98
Trial (T)	4	2.76	10.15	3.06
P × D	47	0.00	0	0.69
P × T	188	0.00	0	1.08
D × T	4	3.89	14.33	2.43
P × D × T	188	15.23	56.09	1.56
Total		27.15	100	
Males (n = 25)				
P	22	7.24	51.15	2.32
D	1	1.32	9.32	1.14
T	4	0.70	4.91	0.51
P × D	22	0.45	3.14	0.33
P × T	88	1.06	7.47	0.47
D × T	4	0.13	0.91	0.16
P × D × T	88	3.27	23.1	0.49
Total		14.16	100	
Females (n = 23)				
P	24	7.92	59.03	2.4
D	1	0.26	1.92	0.30
T	4	0.36	2.65	0.36
P × D	24	0.42	3.16	0.33
P × T	96	0.62	4.65	0.43
D × T	4	0.28	2.08	0.24
P × D × T	96	3.56	26.51	0.51
Total		13.42	100	

Table 2. Decision Study Summary[a]

Males (n = 25)			Females (n = 23)		
Day(s) of Testing	Trials per Day	Generalizability Coefficient	Day(s) of Testing	Trials per Day	Generalizability Coefficient
1	1	0.63	1	1	0.60
	2	0.76		2	0.74
	3	0.81[b]		3	0.79[b]
	4	0.84		4	0.83
	5	0.86		5	0.85
	6	0.88		6	0.86
	7	0.89		7	0.87
	8	0.89		8	0.88
	9	0.90		9	0.89
	10	0.90		10	0.89
2	1	0.75	2	1	0.71
	2	0.85[b]		2	0.82[b]
	3	0.89		3	0.87
	4	0.91		4	0.89
	5	0.92		5	0.91
	6	0.93		6	0.92
	7	0.93		7	0.92
	8	0.94		8	0.93
	9	0.94		9	0.93
	10	0.95		10	0.94

[a] Results are provided by 1 and 2 days of testing and single or multiple Balance Error Scoring System trials per day.
[b] Indicates the level at which strong reliability is present.

tion. Despite its use in the sports medicine community and despite support from medical organizations,[7] a thorough appraisal of the test's psychometric properties has yet to be conducted. Previous investigators have shown that when administered serially, the test-retest stability of the BESS may be less than ideal for clinical applicability.[13,24] Our findings support the notion that BESS performance changes as a result of learning effects associated with multiple exposures (Figure 2). The findings also provide a way to control for the balance improvements seen with multiple test administrations. Specifically, when the test is administered on a single occasion (eg, a postmorbid assessment only), the mean score from 3 administrations is necessary to obtain a stable measure of balance performance (Table 2). The final mean value may then be compared with normative data or an absolute criterion (if available) and used in clinical decision making. When the test is administered over 2 days and another source of variability is taken into consideration, the BESS only needs to be administered twice at each session. Similarly, the scores from both days should be averaged and interpreted. In doing so, the reliability of the test results increases to a level deemed acceptable for clinical use and interpretation (Table 2).

Improved performance between test administrations is likely associated with the development of balance strategies (ie, learning effects) related to multiple exposures to the novel task. A 4-point improvement was noted when preteen participants were evaluated 3 times over 5 days,[13] and adolescent participants improved by approximately 3

points using a similar assessment timeline.[24] The level of improvement seen here did not approach levels reported for younger participants, but our findings are similar to those reported in collegiate athletes.[25] Learning effects are not uncommon to postural control assessments and have been reported with other balance techniques used for concussion assessment. Peterson et al[26] reported a 10% improvement in the NeuroCom Sensory Organization Test composite balance score, which paralleled findings from other investigators[12,27] using identical assessment techniques. These balance improvements were also associated with multiple test exposures, but the authors did not provide an explanation or strategy to reduce or account for the learning effects. Recognizing and controlling for the reduced test stability plays a vital role in making an appropriate injury diagnosis and return-to-play decision. For example, if athletes with concussions demonstrate large learning effects, then subsequent improvements in balance may mask decrements linked to the concussive injury and result in an inaccurate diagnosis or a premature return-to-play decision.

Improvements in postural control have been hypothesized to occur through several mechanisms. Short-term improvements may be related to shifts in the emphasis placed on sensory information used in maintaining balance. For example, under normal stance conditions, vision plays an important role,[28] but when the eyes are closed and vision is removed from the postural control process, greater weight is placed on the vestibular and somatosensory mechanisms to maintain balance. Other short-term changes may be associated with the balance strategy an individual applies to complete a novel task. This can be seen by comparing a firm base of support, on which postural adjustments are made by slight anterior-posterior adjustments at the ankle,[29] with a compliant

foam surface. In the new stance condition, the ankle strategy becomes less effective and, thus, a different balance strategy must be adopted, such as increased use of the hip or knee joints. The long-term improvements in balance noted between the baseline and posttest assessments are likely related to adoption of the new balance strategies developed during the initial test exposures.[30]

Performance improvements on a novel balance task are inevitable and must be accounted for or controlled for in the clinical environment. Our findings indicate that the clinician should adopt a multiple-assessment model when using baseline and follow-up protocols. This technique will yield a stable and reliable balance assessment. However, although completion of the BESS takes less than 5 minutes, using this protocol increases the assessment time and may not be feasible for those administering baseline assessments to some large athletic teams. Depending on the setting, the clinician may elect to account for learning effects through statistical measures. The reliable change index has been successfully implemented with the BESS to account for the observed learning effects. One group[14] recommended that the postconcussion assessment on the BESS differ from baseline by at least 4 points to be considered clinically meaningful. A clear trend among our healthy participants was improved balance performance with an increased number of exposures (Figure 2), but the improvement did not exceed 4 points between any 2 administrations. Future authors should evaluate the sensitivity and specificity of the BESS using the multiple-administration format described here in a sample of concussed young adults.

Our findings also illustrate how to determine and understand the reliability of a clinical measure using a more advanced measurement method. Reliability is one of the most confusing psychometric concepts in constructing clinical measures. Often, the term *reliability* has been used to describe all related efforts in understanding a test-taker's performance consistency. To quantify this concept, statistical indexes such as the Pearson correlation or ICC are used to describe the "reliability" of the measure. For clinical measures like the BESS, single-test administration reliability and trait stability should be distinguished. Commonly used reliability coefficients, however, do not provide adequate information to address this important difference. Fortunately, G theory permits variance sources other than random error to be identified (Table 1), so that a specific type of reliability can be distinguished and quantified. More importantly, this information can help guide clinical practice. Specifically, the variance arising from the number of administrations was evaluated to determine the ideal number of trials needed for clinical utility. For example, increasing the number of trials in the BESS is more readily achieved and more practical than increasing the number of days. Thus, 3 trials on a given date are recommended for reliable assessment and decision making. Generalizability theory adds strength to investigations of reliability, warranting its use in future investigations.

CONCLUSIONS

Assessing balance as part of a concussion management protocol is supported by a variety of organizations.[6,7] Balance tests are sensitive to concussion's deleterious effects and add to the overall sensitivity of the assessment battery.[9] In many instances, sports medicine professionals commonly administer the tests, such as the BESS, during the preseason, with the intent of performing follow-up evaluations after a suspected concussion. When following this protocol, we find it likely that multiple test exposures may lead to improved balance, which may mask concussion-related balance decrements. Clinicians should be cognizant of this possibility. Applying G theory in this investigation was a useful approach in defining components of reliability and provided meaningful guidelines for clinical practice. Using this approach, clinicians choosing to implement the BESS should consider administering the test 3 times on a given occasion and using the mean score for interpretation and comparison to obtain the most reliable and clinically applicable scores.

REFERENCES

1. Langlois JA, Rutland-Brown W, Wald MM. The epidemiology and impact of traumatic brain injury: a brief overview. *J Head Trauma Rehabil*. 2006;21(5):375–378.
2. McCrea M, Hammeke T, Olsen G, Leo P, Guskiewicz K. Unreported concussion in high school football players: implications for prevention. *Clin J Sport Med*. 2004;14(1):13–17.
3. Field M, Collins MW, Lovell MR, Maroon JC. Does age play a role in recovery from sports-related concussion? A comparison of high school and collegiate athletes. *J Pediatr*. 2003;142(5):546–553.
4. Broshek DK, Kaushik T, Freeman JR, Erlanger DM, Webbe FM, Barth JT. Sex differences in outcome following sports-related concussion. *J Neurosurg*. 2005;102(5):856–863.
5. Guskiewicz KM, Mihalik JP, Shankar V, et al. Measurement of head impacts in collegiate football players: relationship between head impact biomechanics and acute clinical outcome after concussion. *Neurosurgery*. 2007;61(6):1244–1252.
6. Aubry M, Cantu R, Dvorak J, et al. Summary and agreement statement of the First International Conference on Concussion in Sport, Vienna 2001: recommendations for the improvement of safety and health of athletes who may suffer concussive injuries. *Br J Sports Med*. 2002;36(1):6–10.
7. Guskiewicz KM, Bruce SL, Cantu RC, et al. National Athletic Trainers' Association position statement: management of sport-related concussion. *J Athl Train*. 2004;39(3):280–297.
8. Broglio SP, Puetz TW. The effect of sport concussion on neurocognitive function, self-report symptoms, and postural control: a meta-analysis. *Sports Med*. 2008;38(1):53–67.
9. Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. *Neurosurgery*. 2007;60(6):1050–1057.
10. Riemann BL, Guskiewicz KM, Shields EW. Relationship between clinical and forceplate measures of postural stability. *J Sport Rehabil*. 1999;8(2):71–82.
11. 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.
12. Guskiewicz KM, Ross SE, Marshall SW. Postural stability and neuropsychological deficits after concussion in collegiate athletes. *J Athl Train*. 2001;36(3):263–273.
13. 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.
14. 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.
15. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. Norwalk, CT: Appleton & Lange; 1993: 63–65.

16. Steindl R, Kunz K, Schrott-Fischer A, Scholtz AW. Effect of age and sex on maturation of sensory systems and balance control. *Dev Med Child Neurol*. 2006;48(6):477–482.

17. Hirabayashi S, Iwasaki Y. Developmental perspective of sensory organization on postural control. *Brain Dev*. 1995;17(2):111–113.

18. Thompson B. *Score Reliability: Contemporary Thinking on Reliability Issues*. Newbury Park, CA: Sage; 2002:3–23.

19. Safrit MJ, Woods TM. *Introduction to Measurement in Physical Education and Exercise Science*. St Louis, MO: Mosby; 1995:159–169.

20. Cronbach LJ. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley; 1972:1–32.

21. Brennan RL. *Generalizability Theory*. New York, NY: Springer; 2001:1–19.

22. Morrow JR. Generalizability theory. In: Safrit MJ, & Woods TM, eds. *Measurement Concepts in Physical Education and Exercise Science*. Champaign, IL: Human Kinetics; 1989:73–96.

23. Lovell MR, Collins MW, Iverson GL, et al. Recovery from mild concussion in high school athletes. *J Neurosurg*. 2003;98(2):296–301.

24. Valovich TC, Perrin DH, Gansneder BM. Repeat administration elicits a practice effect with the Balance Error Scoring System but not with the standardized assessment of concussion in high school athletes. *J Athl Train*. 2003;38(1):51–56.

25. McCrea M, Guskiewicz KM, Marshall SW, et al. Acute effects and recovery time following concussion in collegiate football players: the NCAA Concussion Study. *JAMA*. 2003;290(19):2556–2563.

26. Peterson CL, Ferrara MS, Mrazik M, Piland SG, Elliot R. Evaluation of neuropsychological domain scores and postural stability following cerebral concussion in sports. *Clin J Sport Med*. 2003;13(4):230–237.

27. Guskiewicz KM, Riemann BL, Perrin DH, Nashner LM. Alternative approaches to the assessment of mild head injury in athletes. *Med Sci Sports Exerc*. 1997;29(suppl 7):S213–S221.

28. Wade MG, Jones G. The role of vision and spatial orientation in the maintenance of posture. *Phys Ther*. 1997;77(6):619–628.

29. Riemann BL, Guskiewicz KM. Contribution of the peripheral somatosensory system to balance and postural equilibrium. In: Lephart SM, & Fu FH, eds. *Proprioception and Neuromuscular Control in Joint Stability*. Champaign, IL: Human Kinetics; 2000:37–52.

30. Tjernstrom F, Fransson PA, Hafstrom A, Magnusson M. Adaptation of postural control to perturbations: a process that initiates long-term motor memory. *Gait Posture*. 2002;15(1):75–82.