Psychometric Evaluation of the Lower Extremity Computerized Adaptive Test, the Modified Harris Hip Score, and the Hip Outcome Score

Man Hung,*[†‡§||] PhD, Shirley D. Hon,[†¶] BS, Christine Cheng,[†#] BS, Jeremy D. Franklin,[†**] MA, Stephen K. Aoki,[†] MD, Mike B. Anderson,[†] MS, Ashley L. Kapron,[†] PhD, Christopher L. Peters,[†] MD, and Christopher E. Pelt,[†] MD

Investigation performed at the Department of Orthopaedics, University of Utah, Salt Lake City, Utah, USA

Background: The applicability and validity of many patient-reported outcome measures in the high-functioning population are not well understood.

Purpose: To compare the psychometric properties of the modified Harris Hip Score (mHHS), the Hip Outcome Score activities of daily living subscale (HOS-ADL) and sports (HOS-sports), and the Lower Extremity Computerized Adaptive Test (LE CAT). The hypotheses was that all instruments would perform well but that the LE CAT would show superiority psychometrically because a combination of CAT and a large item bank allows for a high degree of measurement precision.

Study Design: Cohort study (diagnosis); Level of evidence, 2.

Methods: Data were collected from 472 advanced-age, active participants from the Huntsman World Senior Games in 2012. Validity evidences were examined through item fit, dimensionality, monotonicity, local independence, differential item functioning, person raw score to measure correlation, and instrument coverage (ie, ceiling and floor effects), and reliability evidences were examined through Cronbach alpha and person separation index.

Results: All instruments demonstrated good item fit, unidimensionality, monotonicity, local independence, and person raw score to measure correlations. The HOS-ADL had high ceiling effects of 36.02%, and the mHHS had ceiling effects of 27.54%. The LE CAT had ceiling effects of 8.47%, and the HOS-sports had no ceiling effects. None of the instruments had any floor effects. The mHHS had a very low Cronbach alpha of 0.41 and an extremely low person separation index of 0.08. Reliabilities for the LE CAT were excellent and for the HOS-ADL and HOS-sports were good.

Conclusion: The LE CAT showed better psychometric properties overall than the HOS-ADL, HOS-sports, and mHHS for the senior population. The mHHS demonstrated pronounced ceiling effects and poor reliabilities that should be of concern. The high ceiling effects for the HOS-ADL were also of concern. The LE CAT was superior in all psychometric aspects examined in this study. Future research should investigate the LE CAT for wider use in different populations.

Keywords: Rasch modeling; LE CAT; mHHS; HOS; PROMIS; psychometrics

The perspective of the patient is becoming increasingly important in health care decisions. Patient-reported outcomes (PROs) provide a complementary component to the clinical measures that physicians have traditionally used to assess the conditions and improvements of patients.[24] However, many of the PRO instruments have not been sufficiently validated, and their applicability to various populations is unknown.

In orthopaedics, high-functioning patients remain a challenging group to measure. The modified Harris Hip Score (mHHS) is a commonly used joint-specific outcomes measure for hip osteoarthritis, arthroscopic, and arthroplasty procedures.[1,3,25,26] The Hip Outcome Score (HOS) was developed as an evaluative self-report instrument to assess the outcomes of arthroscopic hip surgery.[19-21] However, the applicability and validity of the mHHS and HOS in the high-functioning population are not well understood.

The mHHS and HOS have both undergone validity and reliability testing with mixed results. Although the mHHS is commonly used, there is limited research on its validity

The Orthopaedic Journal of Sports Medicine, 2(12), 2325967114562191
DOI: 10.1177/2325967114562191
© The Author(s) 2014

This open-access article is published and distributed under the Creative Commons Attribution - NonCommercial - No Derivatives License (http://creativecommons.org/licenses/by-nc-nd/3.0/), which permits the noncommercial use, distribution, and reproduction of the article in any medium, provided the original author and source are credited. You may not alter, transform, or build upon this article without the permission of the Author(s). For reprints and permission queries, please visit SAGE's Web site at http://www.sagepub.com/journalsPermissions.nav.

compared with other measures,[25] and its reliability has not been established.[1,16] Kemp et al[17] reviewed a number of hip PRO instruments and found both the HOS and mHHS had excellent test-retest reliability and content validity. Furthermore, both were able to detect differences between patients that received arthroscopic surgery and control groups. The mHHS demonstrated good responsiveness as well.[17] While neither the mHHS nor the HOS activities of daily living subscale (HOS-ADL) had any floor effects, the HOS-ADL did have ceiling effects.[17] Another study found that the mHHS was of moderate quality and recommended the use of the HOS in conjunction with the Nonarthritic Hip Score because there was no evidence for the use of a single PRO instrument.[34] Other studies concluded the HOS was the most reliable and valid PRO instrument for patients undergoing arthroscopy despite psychometric investigations, which were not the goal of the research.[31,33] In 2007, Martin and Philippon[20] found the HOS-ADL and HOS sports (HOS-sports) subscales had a high correlation to the Short Form–36 physical subscale. Yet they found the HOS-ADL and HOS-sports scores were significantly different based on current activity level, surgical outcome, and age.[20] Naal et al[23] showed that neither the 2-factor structure nor the unidimensionality of each of the HOS subscales was supported. Safran and Hariri[30] suggested that the HOS may not be as applicable for older patients, even though many clinicians and researchers have used it with older patients.

Recently developed instruments using advance methodologies such as item response theory (IRT) and computerized adaptive testing (CAT) have emerged seeking to improve on legacy instruments such as the HOS and mHHS. CAT, utilizing IRT,[29] has been used in the educational field to optimize test administration for decades by reducing test length, time constraints, data entry errors, respondent anxiety, fatigue, and administration cost while maintaining measurement efficacy.[5,22] By tailoring questions based on respondents' abilities, CAT can reduce the time burden.[22] The questions that CAT presents to the respondents are individualized. If a respondent answers a question indicating that they cannot walk 1 block, CAT would not pull a question to test if the respondent can walk 1 mile, thus substantially cutting down irrelevant items and time for test administration. This is very important in the clinical setting as CAT enables precise assessment without lowering clinicians' productivity.

In the past decade, the National Institutes of Health has funded the establishment of the Patient-Reported Outcomes Information System (PROMIS) utilizing IRT and CAT. One of the PROMIS initiatives was to develop validated PRO item banks freely available for public use.[4,28] In recent studies, the PROMIS physical function (PF) instruments have demonstrated advantages compared with legacy, that is, commonly used PRO instruments.[10-13,15] They have been validated in various orthopaedic patient populations, including foot and ankle, spine, and trauma patients.[8,10-13,15] However, the PROMIS PF instruments have been shown to have item bias between patients who have lower extremity versus upper extremity problems. To address this issue, researchers at the University of Utah developed a 79-item lower extremity (LE) CAT item bank from the larger PROMIS PF item bank to target patients with lower extremity disorders.[10-14] Preliminary results suggested that the LE CAT performs well in the orthopaedic patient population[20,21,27]; however, as with the HOS and mHHS, the LE CAT has never been studied in the high-performing older population. Given that the HOS, mHHS, and LE CAT were all developed to measure the physical function trait, comparison of these instruments would be very informative. Currently, there is insufficient knowledge whether the LE CAT, HOS-ADL, HOS-sports, and mHHS are adequate for assessing athletes and high-performing individuals.

It is of critical importance that an instrument is able to measure the function of healthy or high-performing individuals. One main goal in any medical treatment is to help patients return to normal health conditions. If an instrument is able to measure a person's functioning status while he or she is sick but not able to measure well when he or she recovers or returns to normal, then the value of that instrument would be questionable. Furthermore, if an instrument is not able to measure well when people return to normal, healthy conditions, we will not know whether the treatment or intervention is effective. For benchmarking purposes, it is also necessary for an instrument to be sensitive to the normal, healthy, or high-performing population.

Given the high performance potential and advanced age of senior athletes participating in the Huntsman World Senior Games, we set to evaluate the psychometric performance of the LE CAT compared with the mHHS, HOS-ADL, and HOS-sports—legacy instruments that are sometimes used to evaluate hip function and outcomes in similarly high-performing individuals in clinical practices. This study aimed to evaluate validity and reliability

[*]Address correspondence to Man Hung, PhD, Department of Orthopaedics, University of Utah, 590 Wakara Way, Salt Lake City, UT 84108, USA (e-mail: man.hung@hsc.utah.edu).

[†] Department of Orthopaedics, University of Utah, Salt Lake City, Utah, USA.

[‡] Division of Public Health, University of Utah, Salt Lake City, Utah, USA.

[§] Division of Epidemiology, University of Utah, Salt Lake City, Utah, USA.

[‖] Huntsman Cancer Institute, University Medical Center, Salt Lake City, Utah, USA.

[¶] Department of Computer & Electrical Engineering, University of Utah, Salt Lake City, Utah, USA.

[#] College of Pharmacy, Roseman University of Health Sciences, South Jordan, Utah, USA.

[**]Department of Education, Culture & Society, University of Utah, Salt Lake City, Utah, USA.

One or more of the authors has declared the following potential conflict of interest or source of funding: This investigation was supported by the L.S. Perry Foundation, University of Utah Department of Orthopaedics, Center for Outcomes Research and Assessment, with funding in part from the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (award number U01AR067138). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

evidences of all 4 instruments. We hypothesized that all 4 measures would perform well but that the LE CAT would show superiority psychometrically because a combination of CAT and a large item bank allows for a high degree of measurement precision.

METHODS

Data Collection

After obtaining approval from our institutional review board, we conducted a prospective cross-sectional study by administering the LE CAT, HOS-ADL, HOS-sports, and mHHS to athletes participating in the Huntsman World Senior Games in October 2012. The Huntsman World Senior Games is an international competition for athletes aged 50 years and older. There are certain competitions, such as the partner dance, that allowed participants to be younger than 50 years as long as the average age of the partners is at least 50 years. Twenty-seven events (see Appendix 1) are included in the games, ranging from traditional team games and individual races to target shooting and minimal exertion/recreational activities. After informed consent, participants provided demographic information including age, sex, race, and ethnicity. Those who did not participate in the Senior Games were excluded. The PRO instruments were administered on computer tablets via the PROMIS assessment center website (www.assessmentcenter.net). The following data were collected: participant demographics (ie, age, sex, race, and ethnicity) and patient responses.

PRO Instruments

The HOS contains 19 items in the HOS-ADL subscale and 9 items in the HOS-sports subscale.[24] With the suggestion from the scoring guidelines, only 17 of the 19 items in the HOS-ADL were scored and used for all analyses.[18] The response options of the HOS items range from 0, indicating "extreme difficulty," to 4, indicating "no difficulty at all." Derived from the Harris Hip Score, the mHHS has 8 items that cover 8 areas: pain, limp, support, distance walked, stairs, shoes/socks, sitting, and public transportation.[6,24] The mHHS is scored on a 100-point scale, with each answer receiving a specific amount of points. The LE CAT includes a bank of 79 items that can be drawn from CAT algorithms.[10-14] Item responses from the LE CAT bank are based on a 5-point rating scale. Appendices 2 through 5 show all of the items and response options in these instruments.

Analytic Approach

Sample and instrument characteristics were examined using mean, standard deviation, proportion, and correlation as appropriate. Psychometric evaluation of the 4 instruments was carried out using the Rasch partial credit model. The Rasch partial credit model is a formal measurement model for evaluation of items that contain unique rating scale structures[27] and has been used in modern instrument development, refinement, and evaluation.[7,32]

In this study, we evaluated the psychometric performance of the LE CAT, HOS-ADL, HOS-sports, and mHHS via multiple important indicators of validity and reliability. Specifically, we examined validity through item fit, dimensionality, monotonicity, local independence, differential item functioning, person raw score to measure correlation and instrument coverage, and examined reliability through Cronbach alpha and person separation index. Table 1 presents a list of these validity and reliability indicators and a brief guide for interpretation.

RESULTS

Sample and Instrument Descriptive

The final sample size for the study was 472 consecutive participants. The majority of the sample was male (n = 266; 56.4%), white (n = 442; 93.6%), and not Latino/Hispanic (n = 447; 96.8%) (Table 2). The average age of the participants was 67 years (SD, 8 years; range, 47-91 years).

Table 3 presents descriptive statistics of the outcomes instruments studied. On average, 9 items (range, 4-12) from the LE CAT item bank were administered to the participants. All 8 items from the mHHS, the 19 items from the HOS-ADL, and the 9 items from the HOS-sports were administered. The Pearson product-moment correlations for all 4 instruments were calculated. The correlation between HOS-ADL and mHHS was high ($r = 0.725$), and the correlation between the HOS-sports and the mHHS was almost equally as high ($r = 0.708$). The HOS-sports and the HOS-ADL exhibited a high correlation ($r = 0.846$). The LE CAT was moderately correlated with the HOS-ADL ($r = 0.583$), the HOS-sports ($r = 0.574$), and the mHHS ($r = 0.419$).

Validities

Item Fit. Items from all 3 instruments demonstrated good fit to the model (Table 4). The LE CAT demonstrated an average outfit mean square (MNSQ) statistic of 0.79. The MNSQ statistic is a measure of item fit to the Rasch Partial Credit model and ranges from negative infinity to positive infinity, with values close to 1 as the best fit. The average outfit MNSQ for the HOS-ADL was 1.02, the HOS-sports was 0.91, and for the mHHS was 0.92.

Dimensionality. After accounting for the first dimension, the unexplained variances of the residuals were 1.5% for the LE CAT, 5.4% for the HOS-ADL, 7.4% for the HOS-sports, and 5.2% for the mHHS. The LE CAT was clearly unidimensional while the HOS-ADL and the mHHS were marginally unidimensional. The HOS-sports had the highest percentage of unexplained variance in the first dimension.

Monotonicity. None of the instruments had any items with disordered thresholds, implying that item response categories worked as intended.

TABLE 1
Multiple Indicators of Validities and Reliabilities Examined[a]

Psychometric Property	Description/Interpretation
Validities	
Item fit	Validity evidence of the 3 instruments (the LE CAT, HOS, and mHHS) was gathered through multiple perspectives. We initially examined whether the data fit the Rasch partial credit model. We utilized the outfit mean square (MNSQ) statistic to measure fit of the data to the Rasch partial credit model. An MNSQ that is <1.5 indicates that the data fit the Rasch model well.[2,9,35] If the data do not fit the Rasch partial credit model, it would not be appropriate to proceed to further analyses using this model, as the instrument likely does not conform to the axioms of quantitative measurement.[27]
Dimensionality	The dimensionality of each of the instruments was investigated to determine if each instrument was unidimensional (measuring a single dimension, eg, construct, idea, phenomenon, factor) or multidimensional. Principal component analyses of residuals were conducted to determine the dimensionality of each instrument. After controlling for the first dimension, if the unexplained variance of the residuals in the first dimension was <5%, the instrument was viewed as unidimensional.[35]
Monotonicity	Monotonicity refers to the circumstance that item response categories are working as intended in increasing or decreasing hierarchical order. An item lacks monotonicity if the response categories are not correctly ordered (eg, 0 = never, 1 = always, 2 = sometimes). Response categories not in correct orders are also referred as disordered thresholds. A valid working instrument should not contain any items with disordered thresholds.
Local independence	Local independence occurs when the response to one item is independent of the response to another item, after taking into account the first dimension. When local independence is violated, the response to one item determines the response to another item. Local independence was determined by investigating the item residual correlations (residuals are part of the data that are not explained by the first dimension). We considered items with residual correlations >0.8 as substantially departing from local independence.
Differential item functioning (DIF)	DIF measures item bias. A properly constructed instrument should not vary greatly when administered to various subgroups within a population (eg, sex, age, ethnicity, race, socioeconomic status), at different time points, or when employing assorted modes of instrument administration. DIF was assessed on an item by item basis using Mantel-Haenszel chi-square test. We examined age (<65 years or ≥65 years) and sex (male or female) DIF in this study and considered items with Mantel-Haenszel chi-square test $P < .05$ as having significant DIF.
Raw score to measure correlation	Person raw scores for each of the 3 instruments are on an ordinal scale. Generally, the raw scores are not useful for parametric statistics unless they are in an interval scale. Interval scale scores are called measures. A low correlation between raw scores and measures indicates that it is not appropriate to use common statistical procedures such as sum, mean, standard deviation, and *t* test. We considered raw scores to measure correlation <0.4 as low and >0.8 as high.
Instrument coverage	Instrument coverage, or targeting, is the extent to which items in an instrument adequately measure the entire range of the sample's trait levels (eg, ability levels, functioning levels, pain levels). If the items are not able to sufficiently cover people's upper levels or lower levels of the trait, the instrument is said to have ceiling effects or floor effects, respectively. Instruments with high ceiling or floor effects are not useful for longitudinal or comparative effectiveness studies as they lack the ability to detect changes. Coverage is computed by taking the item and person score distributions (both in interval scale measures) and calculating the percentage of persons on the upper (ceiling) and the lower (floor) ends of the person score distribution that are not aligned with the item score distribution. Instruments >15% ceiling or floor are considered as problematic.
Reliabilities	
Internal consistency	Internal consistency reliability is the extent to which all of the items within an instrument measure the same construct. We examined internal consistency of the instruments using the Cronbach alpha. Cronbach alpha ranges from 0 to 1, with a value of ≥0.70 generally regarded as adequate.
Person separation	We also calculated the person separation index (PSI) of the LE CAT, HOS-ADL, HOS-sports, and mHHS. The PSI is similar to the conventional Cronbach alpha except that there is no upper bound to the PSI; the PSI is on a ratio scale and ranges from 0 to infinity. In other words, as opposed to Cronbach alpha, the PSI has no ceiling in measuring reliability. The higher the PSI, the more reliable the instrument.[35] An instrument with PSI of <1 is undesirable, as it is insensitive enough to distinguish the sample into at least 2 strata (such as high and low functioning abilities), and thus more items should be added to the instrument.

[a]HOS-ADL, Hip Outcome Score–activities of daily living subscale; HOS-sports, Hip Outcome Score–sports subscale; LE CAT, Lower Extremity Computerized Adaptive Test; mHHS, modified Harris Hip Score.

TABLE 2
Participants' Demographic Characteristics (N = 472)[a]

Age, y, mean ± SD (range)	67.0 ± 8.3 (47-91)
<65	195 (41.3)
≥65	277 (58.7)
Sex	
Male	266 (56.4)
Female	206 (43.6)
Race	
White	442 (93.6)
Black	9 (1.9)
Asian	8 (1.7)
Other	11 (2.3)
Missing	2 (0.4)
Ethnicity	
Not Hispanic or Latino	447 (96.8)
Hispanic or Latino	15 (3.2)
Missing	10 (2.1)

[a] Values are expressed as n (%) unless otherwise indicated.

TABLE 3
Descriptive Statistics of the LE CAT, HOS, and mHHS[a]

		HOS		
	LE CAT[b]	ADL	Sports	mHHS
Mean	71.25	62.49	30.47	86.09
SD	10.12	7.71	6.6	8.34
Median	75.2	65	32	91
IQR	61.60-81.10	60.00-68.00	27.00-36.00	86.00-91.00

[a] ADL, activities of daily living subscale; HOS, Hip Outcome Score; IQR, interquartile range; LE CAT, Lower Extremity Computerized Adaptive Test; mHHS, modified Harris Hip Score; Sports, sports subscale.
[b] The LE CAT was expressed in *T*-score.

TABLE 4
Summary of Psychometric Analyses
for the LE CAT, HOS, and mHHS

		HOS		
	LE CAT	ADL	Sports	mHHS
Validities				
Item fit: outfit MNSQ	0.79	1.02	0.91	0.92
Dimensionality–first dimension: unexplained variance of residual, %	1.5	5.4	7.4[b]	5.2
Monotonicity: disordered thresholds, n	0	0	0	0
Local independence: residual correlation >0.8, n	0	0	0	0
Differential item functioning				
Sex, n	0	3	0	2
Age, n	2	2	4[b]	2
Person raw score to measure: correlation	0.94	0.86	0.83	0.84
Instrument coverage				
Ceiling effect, %	8.47	36.02[b]	0	27.54[b]
Floor effect, %	0	0	0	0
Reliabilities				
PSI	2.75	1.28	1.34	0.08[b]
Cronbach α	1	0.97	0.97	0.41[b]

[a] ADL, activities of daily living subscale; HOS, Hip Outcome Score; LE CAT, Lower Extremity Computerized Adaptive Test; mHHS, modified Harris Hip Score; MNSQ; mean square; PSI, person separation index; Sports, sports subscale.
[b] Area of concern.

Local Independence. None of the instruments had item residual correlations greater than 0.8. This means that all 3 instruments were locally independent and answers to 1 item did not determine answers to the other items.

Differential Item Functioning (DIF). We found no significant sex DIF for the LE CAT. The HOS-ADL contained 3 items with a significant sex DIF. The 3 items were "rolling over in bed" (chi-square [χ^2] = 4.6269; P = .0315), "walking 15 minutes or greater" (χ^2 = 5.3037; P = .0213), and "light to moderate work (standing, walking)" (χ^2 = 3.9762; P = .0461). Specifically, the item "rolling over in bed" was less difficult for females to endorse than males, but the items "walking 15 minutes or greater" and "light to moderate work (standing, walking)" were less difficult for males to endorse than females. The mHHS showed significant sex DIF for 2 items: "ability to put on your shoes and socks" (χ^2 = 5.3387; P = .0209) and "ability to climb stairs" (χ^2 = 4.8863; P = .0271). It was less difficult for males to endorse the item "ability to put on your shoes and socks" than females; the reverse was true for the item "ability to climb stairs."

In terms of DIF across age, we compared participants who were younger (<65 years) versus older (≥65 years) for the instruments. The LE CAT, HOS-ADL, and mHHS each had 2 items with significant age DIF. Those items are "bending, kneeling, or stooping" (χ^2 = 5.8533; P = .0155) and "ability to run 100 yards" (χ^2 = 2.8980; P = .0483) for the LE CAT, "walking up steep hills" (χ^2 = 4.585, P = –.0275) and "going up one (1) flight of stairs" (χ^2 = 6.1141; P = .0134) for the HOS-ADL, and "your limp" (χ^2 = 4.5927; P = .0321) and "ability to sit in a chair" (χ^2 = 15.0618; P = .001) for mHHS. The item "your limp" was less difficult for older individuals than younger ones; the reverse was true for "ability to sit in a chair." Younger individuals rated the item "walking up steep hills" as less difficult than older individuals, but the opposite was true for "going up one (1) flight of stairs." For both the items "bending, kneeling, or stooping" and "ability to run 100 yards," younger individuals found them to be less difficult to endorse than older individuals. The HOS-sports had 4 items with significant age DIF. Those items are "1 mile" (χ^2 = 22.8928; P = .0000), "cutting" (χ^2 = 4.5986; P = .0320), "stop" (χ^2 = 20.9531; P = .0000), and "swing" (χ^2 = 3.9963; P = .0456). Older participants found the item "1 mile" to be less difficult to endorse than younger individuals. For the items "cutting," "stop," and "swing," older individuals found them to be more difficult to endorse than did younger individuals.

Raw Score to Measure Correlation. The person raw score to measure correlations were high for all instruments

(LE CAT, 0.94; HOS-ADL, 0.86; HOS-sports, 0.83; and mHHS, 0.84).

Instrument Coverage. The LE CAT, HOS-ADL, HOS-sports, and mHHS exhibited no floor effects. The ceiling effects were high for the HOS-ADL and the mHHS (36.02% and 27.54%, respectively) and acceptable for the LE CAT (8.47%). The HOS-sports exhibited no ceiling effects.

Reliabilities

Internal Consistency. The LE CAT, HOS-ADL, and HOS-sports had a high Cronbach alpha of 1.00, 0.97, and 0.97, respectively. The Cronbach alpha for the mHHS was 0.41, indicating poor internal consistency reliability.

Person Separation Index (PSI). With the highest PSI (2.75), the LE CAT is capable of distinguishing at least 3 strata of participants. The mHHS had an extremely low PSI of 0.08, indicating the mHHS could not discriminate various performing participants in the sample. The HOS-ADL and HOS-sports had acceptable PSI of 1.28 and 1.34, respectively.

DISCUSSION

We evaluated the psychometric properties of the LE CAT, HOS-ADL, HOS-sports, and mHHS to better understand how the instruments perform. This study showed that the HOS-ADL, HOS-sports, and mHHS instruments exhibited questionable psychometric properties, especially after reviewing their ceiling effects, unidimensionality, and reliability indicators. Specifically, the HOS-sports subscale had very high unexplained variance and high proportion of items that performed differently across age groups. Additionally, the mHHS manifested an extremely poor PSI and Cronbach alpha. The study, however, did reveal that the LE CAT is a much better performing instrument for assessing the hip and joints for the high-functioning senior population from a large body of validity and reliability evidences.

After confirming that outcomes instruments each fit the Rasch model well, we proceeded with the Rasch analysis to examine the instruments' validities and reliabilities. The LE CAT, HOS-ADL, and mHHS provided evidence of unidimensionality, with the HOS-ADL and mHHS being marginally unidimensional and the LE CAT clearly unidimensional. The HOS-sports subscale demonstrated the furthest departure from unidimensionality. Considering that the HOS-ADL, HOS-sports, and mHHS are specific hip and joint outcomes instruments, we would have expected them to have a less unexplained residual variance because these instruments are supposed to be targeted to a specific region of the body. Surprisingly, the LE CAT showed the lowest unexplained residual variance and was the best among the 4 measures studied.

When investigating item bias, we found differences in male and female responses for the HOS-ADL and the mHHS. The LE CAT did not have any items with sex bias. Overall, we found that 17.6% of items in the HOS-ADL had sex bias and 25% of items in the mHHS had sex bias. The LE CAT, HOS-ADL, and mHHS had 2 items with age bias, which corresponded to 2.5% items in the LE CAT item bank, 11.8% of items in the HOS-ADL, and 25% of items in the mHHS. The HOS-sports had 4 items with age bias, which corresponds to a very large 44% of items in the subscale. The proportion of items with age bias in the LE CAT item bank was minimal. The proportion of items with sex and age bias in the HOS-ADL, HOS-sports, and the mHHS could be of potential concern, especially for the mHHS and HOS-sports. Further modification of these items or separate scoring is needed.

The person raw scores to measure correlations were satisfactory for all instruments indicating that their raw scores are acceptable for common statistical analyses. All instruments had a raw score to measure correlation greater than 0.8, with the LE CAT again being the best at 0.94. As a result of these high correlations, it may be possible to use the raw scores of the instruments to perform common statistical procedures.

Participants that took the LE CAT and the HOS-sports were better targeted by all items, but participants that took the HOS-ADL and mHHS were not nearly as well covered, especially when considering high-functioning participants. The LE CAT, HOS-ADL, HOS-sports, and mHHS showed no floor effects, but the HOS-ADL and mHHS had serious ceiling effects. The HOS-sports subscale was the only instrument that demonstrated no ceiling and floor effects, and hence, it was applicable to the high-functioning population. Our findings were similar to previous studies that found ceiling effects of the HOS-ADL.[17] The ceiling effects were very high for instruments that are supposed to assess an all-encompassing hip and joint population. The ceiling effects are particularly worrisome because of the population that was being assessed. While the population was athletes, they are also seniors with a mean age of 67 years and a minimum age of 47 years. Since the athletes are participating in highly competitive senior games, we might assume that the participants were in better than average health than their senior peers. Unfortunately, the HOS-ADL and mHHS instruments could not capture those that were really high-performing seniors. We are left to question whether these instruments are adequate for active seniors. Previous research has shown that the mHHS has not been adequately evaluated.[1,16,25] As a result, they could not recommend using the mHHS to assess an active patient population that has had hip arthroscopy.[18] The LE CAT demonstrated much lower ceiling effects that would likely be considered more reasonable and applicable for assessing hip and joint patients and more generally, patients with lower extremity disorders. The HOS-sports demonstrated no ceiling effects, and this was expected as it is an instrument designed for populations that are higher performing and functioning than their peers.

Finally, the LE CAT, HOS-ADL, and HOS-sports demonstrated good internal reliability, but the mHHS did not. The mHHS had a low Cronbach alpha and person separation, indicating that its reliabilities were poor and it could not distinguish between different performing participants. In fact, with such low reliabilities, the mHHS may not be very useful for assessment of outcomes. The HOS-ADL and HOS-sports, on the other hand, did have better reliabilities than the mHHS. More items can be added to the HOS-ADL in future instrument refinement. Overall, the LE CAT performed the best in all fronts.

Limitations

This study, like many studies, has limitations. First, this study was conducted with an older, highly active population. Thus, the findings of this study might not be applicable to all older adults because they may not be as athletic as our participants nor may it be applicable to a younger, active population. Additionally, the population was overwhelmingly identified as white, which is not representative of demographics in the United States.

Second, we did not evaluate responsiveness to change. Being a cross-sectional study, we only captured a single point in time and did not measure how participants might have improved over time. Additional studies are needed for all instruments to assess longitudinal changes in different populations, especially the younger populations that exhibit sex, race, and ethnic diversity.

When examining the overall results of this study, we found that the LE CAT is the best performing, well-rounded instrument among the 4. Findings from previous studies and this study should indicate to clinicians and researchers that the HOS-ADL, HOS-sports, and mHHS will require additional scrutiny and psychometric testing to identify which population is best served by each instrument. It may be the case that each instrument should only be used for a very specific hip and joint population.

CONCLUSION

Among a senior, athletic population, we evaluated the psychometric properties of the most commonly used hip and joint assessments along with a promising instrument that is increasingly being used to assess lower extremities. The LE CAT exhibited better overall psychometric performance than did the legacy instruments—the HOS-ADL, HOS-sports, and mHHS. Additional modification for the HOS-ADL, HOS-sports, and mHHS are strongly recommended prior to further use in clinical settings. While the LE CAT can certainly benefit from further refinement and an addition of more items to close the ceiling gap, as it currently stands, the LE CAT is clearly more superior than the HOS-ADL, the HOS-sports, and the mHHS in all psychometric aspects examined.

REFERENCES

1. Aprato A, Jayasekera N, Villar RN. Does the modified Harris hip score reflect patient satisfaction after hip arthroscopy? *Am J Sports Med*. 2012;40:2557-2560.
2. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New York, NY: Routledge; 2012.
3. Byrd JW, Jones KS. Prospective analysis of hip arthroscopy with 2-year follow-up. *Arthroscopy*. 2000;16:578-587.
4. Cella D, Riley W, Stone A, et al; PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010;63:1179-1194.
5. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv Res*. 2005;40:1694-1711.
6. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. *J Bone Joint Surg Am*. 1969;51:737-755.
7. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol*. 2007;6:1094-1105.
8. Hung M, Baumhauer JF, Latt LD, Saltzman CL, SooHoo NF, Hunt KJ. Validation of PROMIS (R) Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clin Orthop Relat Res*. 2013;471:3466-3474.
9. Hung M, Carter M, Hayden C, et al. Psychometric assessment of the Patient Activation Measure Short Form (PAM-13) in rural settings. *Qual Life Res*. 2013;22:521-529.
10. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res*. 2011;29:947-953.
11. Hung M, Clegg DO, Greene T, Weir C, Saltzman CL. A lower extremity physical function computerized adaptive testing instrument for orthopaedic patients. *Foot Ankle Int*. 2012;33:326-335.
12. Hung M, Franklin JD, Hon SD, Cheng C, Conrad J, Saltzman CL. Time for a paradigm shift with computerized adaptive testing of general physical function outcomes measurements. *Foot Ankle Int*. 2014;35:1-7.
13. Hung M, Hon SD, Franklin JD, et al. Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. *Spine (Phila Pa 1976)*. 2014;39:158-163.
14. Hung M, Nickisch F, Beals TC, Greene T, Clegg DO, Saltzman CL. New paradigm for patient-reported outcomes assessment in foot & ankle research: computerized adaptive testing. *Foot Ankle Int*. 2012;33:621-626.
15. Hung M, Stuart AR, Higgins TF, Saltzman CL, Kubiak EN. Computerized adaptive testing using the PROMIS Physical Function item bank reduces test burden with less ceiling effects compared to the short musculoskeletal function assessment in orthopaedic trauma patients. *J Orthop Trauma*. 2014;28:439-443.
16. Kemp JL, Collins NJ, Makdissi M, Schache AG, Machotka Z, Crossley K. Hip arthroscopy for intra-articular pathology: a systematic review of outcomes with and without femoral osteoplasty. *Br J Sports Med*. 2012;46:632-643.
17. Kemp JL, Collins NJ, Roos EM, Crossley KM. Psychometric properties of patient-reported outcome measures for hip arthroscopic surgery. *Am J Sports Med*. 2013;41:2065-2073.
18. Martin RL. Hip arthroscopy and outcome assessment. *Oper Tech Orthop*. 2005;15:290-296.
19. Martin RL, Kelly BT, Philippon MJ. Evidence of validity for the hip outcome score. *Arthroscopy*. 2006;22:1304-1311.
20. Martin RL, Philippon MJ. Evidence of validity for the hip outcome score in hip arthroscopy. *Arthroscopy*. 2007;23:822-826.
21. Martin RL, Philippon MJ. Evidence of reliability and responsiveness for the hip outcome score. *Arthroscopy*. 2008;24:676-682.
22. McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Ann Intern Med*. 2003;139:403-409.
23. Naal FD, Impellizzeri FM, von Eisenhart-Rothe R, Mannion AF, Leunig M. Reproducibility, validity, and responsiveness of the hip outcome score in patients with end-stage hip osteoarthritis. *Arthritis Care Res (Hoboken)*. 2012;64:1770-1775.
24. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health*. 2007;10(suppl 2):S125-S137.

25. Philippon MJ, Schenker ML, Briggs KK, Kuppersmith DA, Maxwell RB, Stubbs AJ. Revision hip arthroscopy. *Am J Sports Med*. 2007; 35:1918-1921.
26. Potter BK, Freedman BA, Andersen RC, Bojescul JA, Kuklo TR, Murphy KP. Correlation of Short Form-36 and disability status with outcomes of arthroscopic acetabular labral debridement. *Am J Sports Med*. 2005;33:864-870.
27. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press; 1960.
28. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res*. 1997;6:595-600.
29. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol*. 2008; 61:17-33.
30. Safran MR, Hariri S. Hip arthroscopy assessment tools and outcomes. *Oper Tech Orthop*. 2010;20:264-277.
31. Schenker ML, Martin R, Weiland DE, Philippon MJ. Current trends in hip arthroscopy: a review of injury diagnosis, techniques, and outcome scoring. *Curr Opin Orthop*. 2005;16:89-94.
32. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health*. 2004;7(suppl 1):S22-S26.
33. Thorborg K, Roos EM, Bartels EM, Petersen J, Holmich P. Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: a systematic review. *Br J Sports Med*. 2010;44:1186-1196.
34. Tijssen M, van Cingel R, van Melick N, de Visser E. Patient-reported outcome questionnaires for hip arthroscopy: a systematic review of the psychometric evidence. *BMC Musculoskelet Disord*. 2011;12:117.
35. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: Mesa Press; 1982.

APPENDIX 1

Activities Hosted for the 2012 Huntsman World Senior Games

Archery
Badminton
Basketball
Bowling
Bridge
Cowboy action shoot
Cycling
Golf
Horseshoes
Lawn bowling
Mountain biking
Pickleball
Racewalking
Racquetball
Road races
Shotgun sports
Shuffleboard
Small bore/airgun benchrest
Soccer
Softball
Square dancing
Swimming
Table tennis
Tennis
Track & field
Triathlon
Volleyball
Walking tours

APPENDIX 2

Lower Extremity (LE) Physical Function Computerized Adaptive Test (CAT) Item Bank

Item No. ID[a]	Item[b]
1. PFA1	Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?
2. PFA3	Does your health now limit you in bending, kneeling, or stooping?
3. PFA4	Does your health now limit you in doing heavy work around the house like scrubbing floors, or lifting or moving heavy furniture?
4. PFA5	Does your health now limit you in lifting or carrying groceries?
5. PFA6	Does your health now limit you in bathing or dressing yourself?
6. PFA7	How much do physical health problems now limit your usual physical activities (such as walking or climbing stairs)?
7. PFA8	Are you able to move a chair from one room to another?
8. PFA9	Are you able to bend down and pick up clothing from the floor?
9. PFA10	Are you able to stand for 1 hour?
10. PFA11	Are you able to do chores such as vacuuming or yard work?
11. PFA12	Are you able to push open a heavy door?
12. PFA13	Are you able to exercise for an hour?
13. PFA14	Are you able to carry a heavy object (over 10 pounds)?
14. PFA15	Are you able to stand up from an armless straight chair?
15. PFA19	Are you able to run or jog for 2 miles?
16. PFA21	Are you able to go up and down stairs at a normal pace?
17. PFA23	Are you able to go for a walk of at least 15 minutes?
18. PFA25	Are you able to do yard work like raking leaves, weeding, or pushing a lawn mower?
19. PFA29	Are you able to pull heavy objects (10 pounds) toward yourself?
20. PFA30	Are you able to step up and down curbs?

(continued)

APPENDIX 2 (continued)

Item No. ID[a]	Item[b]	Item No. ID[a]	Item[b]
21. PFA31	Are you able to get up off the floor from lying on your back without help?	51. PFB49	Does your health now limit you in going for a short walk (less than 15 minutes)?
22. PFA32	Are you able to stand with your knees straight?	52. PFB50	How much difficulty do you have doing your daily physical activities, because of your health?
23. PFA33	Are you able to exercise hard for half an hour?	53. PFB51	Does your health now limit you in participating in active sports such as swimming, tennis, or basketball?
24. PFA37	Are you able to stand for short periods of time?	54. PFB54	Does your health now limit you in going OUTSIDE the home, for example, to shop or visit a doctor's office?
25. PFA39	Are you able to run at a fast pace for 2 miles?	55. PFC6	Are you able to walk a block on flat ground?
26. PFA41	Are you able to squat and get up?	56. PFC7	Are you able to run 5 miles?
27. PFA42	Are you able to carry a laundry basket up a flight of stairs?	57. PFC10	Does your health now limit you in climbing several flights of stairs?
28. PFA45	Are you able to get out of bed into a chair?	58. PFC12	Does your health now limit you in doing 2 hours of physical labor?
29. PFA49	Are you able to bend or twist your back?	59. PFC13	Are you able to run 100 yards?
30. PFA51	Are you able to sit on the edge of a bed?	60. PFC20	Does your health now limit you in walking 100 yards?
31. PFA53	Are you able to run errands and shop?	61. PFC29	Are you able to walk up and down 2 steps?
32. PFA56	Are you able to get in and out of a car?	62. PFC32	Are you able to climb up 5 flights of stairs?
33. PFB1	Does your health now limit you in doing moderate work around the house like vacuuming, sweeping floors or carrying in groceries?	63. PFC33	Are you able to run 10 miles?
34. PFB3	Does your health now limit you in putting a trash bag outside?	64. PFC34	Does your health now limit you in walking several hundred yards?
35. PFB5	Does your health now limit you in hiking a couple of miles on uneven surfaces, including hills?	65. PFC35	Does your health now limit you in doing 8 hours of physical labor?
36. PFB7	Does your health now limit you in doing strenuous activities such as backpacking, skiing, playing tennis, bicycling, or jogging?	66. PFC36	Does your health now limit you in walking more than 1 mile?
37. PFB8	Are you able to carry 2 bags filled with groceries 100 yards?	67. PFC37	Does your health now limit you in climbing 1 flight of stairs?
38. PFB9	Are you able to jump up and down?	68. PFC38	Are you able to walk at a normal speed?
39. PFB10	Are you able to climb up 5 steps?	69. PFC39	Are you able to stand without losing your balance for several minutes?
40. PFB11	Are you able to wash dishes, pots, and utensils by hand while standing at a sink?	70. PFC40	Are you able to kneel on the floor?
41. PFB12	Are you able to make a bed, including spreading and tucking in bed sheets?	71. PFC41	Are you able to sit down in and stand up from a low, soft couch?
42. PFB13	Are you able to carry a shopping bag or briefcase?	72. PFC45	Are you able to get on and off the toilet?
43. PFB14	Are you able to take a tub bath?	73. PFC46	Are you able to transfer from a bed to a chair and back?
44. PFB24	Are you able to run a short distance, such as to catch a bus?	74. PFC47	Are you able to be out of bed most of the day?
45. PFB32	Are you able to stand unsupported for 10 minutes?	75. PFC49	Are you able to water a house plant?
46. PFB40	Are you able to stand up on tiptoes?	76. PFC52	Are you able to turn from side to side in bed?
47. PFB42	Are you able to stand unsupported for 30 minutes?	77. PFC53	Are you able to get in and out of bed?
48. PFB43	Does your health now limit you in taking care of your personal needs (dress, comb hair, toilet, eat, bathe)?	78. PFC54	Does your health now limit you in getting in and out of the bathtub?
49. PFB44	Does your health now limit you in doing moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?	79. PFC56	Does your health now limit you in walking about the house?
50. PFB48	Does your health now limit you in taking a shower?		

[a] Identifier from the PROMIS item bank.

[b] Response options for questions 1-6, 33-36, 48-51, 53, 54, 57, 58, 60, 64-67, 78-79: 1 = cannot do, 2 = quite a lot, 3 = somewhat, 4 = very little, and 5 = not at all. Response options for questions 7-32, 37-47, 52, 55, 56, 59, 61-63, 68-77: 1 = unable to do, 2 = with much difficulty, 3 = with some difficulty, 4 = with a little difficulty, 5 = without any difficulty.

APPENDIX 3

Hip Outcome Score (HOS): Activities of Daily Living[a]

Because of your hip how much difficulty do you have with:

Item No.	Item
1. HOS_sta	Standing for 15 minutes
2. HOS_car	Getting into and out of an average car
3. HOS_put[b]	Putting on socks and shoes
4. HOS_uphi	Walking up steep hills
5. HOS_down	Walking down steep hills
6. HOS_upst	Going up 1 flight of stairs
7. HOS_dnst	Going down 1 flight of stairs
8. HOS_cur	Stepping up and down curbs
9. HOS_squ	Deep squatting
10. HOS_bat	Getting into and out of a bath tub
11. HOS_sit[b]	Sitting for 15 minutes
12. HOS_wki	Walking initially
13. HOS_wal	Walking approximately 10 minutes
14. HOS_wk15	Walking 15 minutes or greater
15. HOS_twi	Twisting/pivoting on involved leg
16. HOS_bed	Rolling over in bed
17. HOS_work	Light to moderate work (standing, walking)
18. HOS_hea	Heavy work (pushing/pulling, climbing, carrying)
19. HOS_rec	Recreational activities

[a] Response options for questions: 0 = unable to do, 1 = extreme difficultly, 2 = moderate difficulty, 3 = slight difficulty, 4 = no difficulty at all, N/A = not applicable.

[b] Per scoring guide, these are filler items not used for scoring.

APPENDIX 4

Hip Outcome Score (HOS): Sports[a]

Because of your hip how much difficulty do you have with:

Item No.	Item
1. HOS_s1mi	Running 1 mile
2. HOS_sjum	Jumping
3. HOS_sswg	Swinging objects like a golf club
4. HOS_slan	Landing
5. HOS_sstp	Starting and stopping quickly
6. HOS_scut	Cutting/lateral movements
7. HOS_slow	Low-impact activities like fast walking
8. HOS_stec	Ability to perform activity with your normal technique
9. HOS_sdes	Ability to participate in your desired sport as long as you would like

[a] Response options for questions: 0 = unable to do, 1 = extreme difficultly, 2 = moderate difficulty, 3 = slight difficulty, 4 = no difficulty at all, N/A = not applicable.

APPENDIX 5

The Modified Harris Hip Score (mHHS) Instrument

Answer the following categories as they relate to your hip:

Item No.	Item
1. mHHS_pai	Please describe any pain in your hip None/ignores (44 points) Slight, occasional, no compromise in activity (40 points) Mild, no effect on ordinary activity, pain after activity, uses aspirin (30 points) Moderate, tolerable, makes concessions, occasional codeine (20 points) Marked, serious limitations (10 points) Totally disabled (0 points)
2. mHHS_lim	Select the answer that best describes your limp None (11 points) Slight (8 points) Moderate (5 points) Severe (0 points) Unable to walk (0 points)
3. mHHS_sup	What is the amount and type of support that you use? None (11 points) Cane, long walks (7 points) Cane, full time (5 points) Crutch (4 points) 2 canes (2 points) 2 crutches (1 points) Unable to walk (0 points)
4. mHHS_dis	Select the answer that best describes how far you can walk Unlimited (11 points) 6 blocks (8 points) 2-3 blocks (5 points) Indoors only (2 points) Bed and chair (0 points)
5. mHHS_sta	Please select the answer that best describes your ability to climb stairs Normally (4 points) Normally with banister (2 points) Any method (1 point) Not able (0 points)
6. mHHS_sho	Please select the answer that best describes your ability to put on your shoes and socks With ease (4 points) With difficulty (2 points) Unable (0 points)
7. mHHS_sit	Please select the answer that best describes your ability to sit in a chair Any chair, 1 hour (5 points) High chair, half hour (3 points) Unable to sit, half hour, any chair (0 points)
8. mHHS_bus	Please select the answer that best describes your ability to use public transportation Able to enter public transportation (1 point) Unable to use public transportation (0 points)