Article

AMERICAN ORTHOPAEDIC FOOT & ANKLE SOCIETY®

Foot & Ankle International®
1–8

DOI: 10.1177/1071100718818573
journals.sagepub.com/home/fai

Validation of the Ankle Osteoarthritis Scale Instrument for Preoperative Evaluation of End-Stage Ankle Arthritis Patients Using Item Response Theory

Guiping Liu, PhD[1], Alexander C. Peterson, MSc[1], Kevin Wing, MD[2], Trafford Crump, PhD[3], Alastair Younger, MD[2], Murray Penner, MD[2], Andrea Veljkovic, MD[2], Hannah Foggin, BSc[1], and Jason M. Sutherland, PhD[1]

Abstract

Background: Significant ankle arthritis results in functional limitations and patient morbidity. There is a need to measure symptoms and the impact of interventions on patient's quality of life using valid and reliable patient-reported measurement instruments. The objective of this research was to validate the Ankle Osteoarthritis Scale instrument in the preoperative setting using factor analysis, item response theory, and differential item function methods.

Methods: This research is based on secondary analysis of patients scheduled for ankle arthrodesis or total ankle replacement in Vancouver, Canada. Participants completed the instrument between September 2014 and August 2017. Item response theory was used to estimate item difficulty and discrimination parameters, controlling for study participants' underlying level of ankle function. Differential item function was examined for sex, age group, and surgery. There were 88 participants.

Results: Modification indices suggested that item 10, "walking around the house," would better fit the pain domain rather than the disability domain. Items in the pain domain displayed a range of discrimination and difficulty. Items in the disability domain exhibited a range of discrimination, though the disability domain had low difficulty. Differential item functioning for sex, age group, and ankle arthrodesis or total ankle replacement appeared to be ignorable.

Conclusion: This evaluation of the Ankle Osteoarthritis Scale found the instrument to be a strong measure of the effect of pain and dysfunction among patients with end-stage ankle arthritis, even when removing items 7 and 8, supporting its prior use in numerous clinical studies.

Level of Evidence: Level II, prospective comparative study.

Keywords: Ankle Osteoarthritis Scale, end-stage ankle arthritis, item response theory, patient-reported outcome measure

End-stage ankle arthritis (ESAA) is a degenerative condition of the tibiotalar joint resulting from cartilage damage. Risk factors, such as obesity, age, and low muscle strength or neuromuscular control, contribute to the deterioration of the joint by directly damaging cartilage or altering the biomechanics of the ankle joint.[1,4,21,30] However, more than 80% of ankle arthritis is due to previous trauma.[23,31] Clinically significant ankle arthritis results in significant functional limitations and patient morbidity—painful and impaired mobility, rest pain, and diminished range of motion. ESAA

[1]Centre for Health Services and Policy Research, School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada
[2]Department of Orthopaedics, University of British Columbia, Vancouver, BC, Canada
[3]Department of Surgery, University of Calgary, Calgary, Alberta

Corresponding Author:
Jason M. Sutherland, PhD, Centre for Health Services and Policy Research, School of Population and Public Health, University of British Columbia, 201-2206 East Mall, Vancouver, BC V6T 1Z3, Canada.
Email: jason.sutherland@ubc.ca

patients report lower SF-36 mental, physical, and general health scores than patients with hip arthritis[16] and have a relatively long life expectancy with high activity demands.[8,36] If the symptoms progress despite nonoperative treatment, surgery is offered and is most often ankle arthrodesis (AA) or total ankle replacement (TAR).[3,4,22] Given the disability of ankle arthritis on patients' quality of life, there is a need to measure symptom improvement and the impact of operative interventions on quality of life using valid and reliable patient-reported measurement instruments.[4]

Published studies evaluating ESAA effects and treatment options have used the Ankle Osteoarthritis Scale (AOS) as a patient-reported outcome measure for quality of life (QoL). The AOS is a validated patient-reported outcome (PRO) measure for ankle arthritis patients.[12] The AOS measures 2 domains of foot and ankle-related quality of life: pain and disability. Each domain is measured with 9 items. The instrument has been validated for use in Canadian populations using classical test theory, demonstrating acceptable reliability, construct, and criterion validity, and has been favorably received by patients,[2,20,35] though analyses have shown that a smaller number of items from the AOS may describe variability among end-stage ankle arthritis patients' responses, simplifying the instrument.[35]

In spite of the relevant findings, there are 3 aspects of the AOS that have not been thoroughly studied that would further inform understanding of the 18 items of the instrument. First, no studies have yet investigated the potential for differential item functioning (DIF), which occurs when there are systematic differences in the measurement qualities of an instrument between 2 or more groups (eg, men and women or fusion and replacement). Second, although construct validity has been demonstrated through a simple correlation with functional disability, the study sample size was small (N=10) and the underlying factor structure of the instrument has not been examined using more sophisticated methods.[12] Finally, limited information is available on the psychometric characteristics of individual items/questions in the AOS. Modern item response theory (IRT), which assesses the relationship between each item, the underlying construct it measures, and patient characteristics, can be used to better understand how each item contributes to the total score.[9] This is relevant because some items in the AOS have floor or ceiling effects whereas other items have little variation between respondents that require investigation.[10,32,35]

In Canada, ankle arthritis causes a significant burden on patients and impact on provincial health spending and lost wages.[15] The incidence and prevalence of ankle arthritis has risen in Canada since 2000 and is predicted to continue rising.[21,30] To measure the impact of the condition on patients' health, disability, and disease burden, strong instruments are needed. The objective of this research was to validate the AOS instrument in the preoperative setting using DIF, factor analysis, and IRT methods. The information generated from this research could provide detailed information regarding the measurement properties of the AOS, and it would be important for clinicians and outcomes researchers. Moreover, the findings could also identify items that could be applied to computer adaptive testing.

Methods

This study is based on a secondary analysis of patients scheduled for elective AA or TAR. Briefly, the original data were collected for a study measuring the preoperative health of patients queued for elective surgery in Vancouver, Canada.[27] To be eligible for participation, patients had to speak or understand English, reside in the community and be at least 18 years of age. Patients scheduled for AA or TAR were contacted by phone to complete the AOS. Patients completed their preoperative surveys between September 2014 and August 2017. Participants' sex, age, self-reported comorbidities, and procedure (AA or TAR) were also collected. Participants were defined as those that agreed to take part in the study, and nonparticipants those that declined. Patients who responded but completed less than 50% of the items were excluded from the study.

The AOS is a disease-specific instrument for the clinical assessment of ankle function.[12] The instrument has a pain and disability domain, each composed of 9 items (18 items total). Each item is presented as a visual analog scale (VAS) in which participants rate their response to each item by placing a mark on a 100mm horizontal line. The pain domain ranges from "no pain" to "worse pain imaginable," and the difficulty domain from "no difficulty" to "so difficult unable." The location of the mark was measured and represents the patients' score. Items' responses in each domain were averaged to calculate a domain score, and the pain and disability score were averaged to calculate the instrument's overall score. The overall score ranges from 0 (indicating best ankle function) to 100 (worst ankle function).

There were 208 patients eligible for participation, of which 98 agreed to participate and returned their survey packet, resulting in a participation rate of 47%. Responders were 3 years older than nonresponders. No other differences were observed between participants and nonparticipants. Ten participants were excluded for completing less than half of the items of the AOS or whose surgery was inaccurately recorded, leaving 88 participants included in the study's analysis. Among these 88 participants, 41 participants were scheduled for surgery on their left side and 47 participants were scheduled for surgery on their right. Sixty-nine percent of the study participants were male and the average age was 59.5 years. Patients aged between 50 and 59 years reported higher AOS scores, as did women and patients with comorbidities. Patients with right-side surgeries reported higher AOS scores, and patients with fusion surgeries reported higher AOS scores. See Table 1 for summary statistics of demographic and clinical characteristics of participants.

Table 1. Summary of Demographic and Clinical Characteristics of the 88 Participants.

		AOS Pain		AOS Disability		Total AOS[a]	
Characteristic	N	Mean	SD	Mean	SD	Mean	SD
Total sample	88	53.7	19.2	62.1	19.8	58.4	18.4
Operative side							
Left	41	50.3	19.9	58.4	22.0	54.8	20.1
Right	47	56.7	18.3	65.3	17.2	61.5	16.4
Age group, y							
≤49	13	49.7	21.0	54.8	22.6	52.6	20.7
50-59	26	61.1	19.2	66.6	17.1	64.2	17.3
60-69	27	51.9	21.5	61.0	21.1	57.0	20.5
>70	22	49.4	13.0	62.4	19.1	56.7	14.9
Gender							
Male	61	52.1	20.4	60.5	19.9	56.8	19.1
Female	27	57.2	15.9	65.8	19.2	62.0	16.7
Number of comorbidities							
0	7	43.9	25.8	56.9	24.6	51.2	23.8
1	30	52.7	17.5	61.0	18.0	57.4	17.0
≥2	51	55.6	19.2	63.4	20.3	60.0	18.6
Fusion or replacement							
Replacement	34	55.4	18.9	64.1	17.0	60.3	16.5
Fusion	54	51.0	19.7	59.0	23.5	55.5	21.1

Abbreviation: AOS, Ankle Osteoarthritis Scale.
[a]The total AOS value is the average of the pain and disability domains.

Statistical Methods

Differences between participants' and nonparticipants' demographic characteristics were evaluated using chi-square tests. Mean domain (pain and disability), total scores with standard deviations across the entire sample and by demographic group were calculated, along with mean item response and standard deviation to assess floor and ceiling effects. Item-level missing data were also examined.

A DIF analysis was performed to determine the demographic characteristics of patients who interpreted or responded to a PRO questionnaire differently in a systematic way. For example, a DIF analysis can be used to determine whether observed differences in pain scores between men and women are due to differences in the type and severity of pain (ie, a "true" difference), or due to differences in how men and women respond to the items of a PRO (ie, DIF).[14] Failing to detect or adjust for DIF can result in biased or inaccurate PRO scores.[29]

This study examined DIF for sex, age group, and patients scheduled for 2 procedures—fusion and replacement using the multiple-indicator multiple-cause (MIMIC) model.[19,28] The MIMIC model began by fitting a confirmatory factor analysis (CFA) model assuming 2 latent variables (pain and disability) were appropriately measured by the AOS items. A frequency histogram of each item's responses was plotted to assess normality. Robust maximum likelihood estimation was to adjust for potential non-normality in item responses. Exogenous variables (eg, sex) were added to the model and tested against the model without the exogenous variables for statistically significant improvement in model fit using the chi-square difference test. Then, modification indices were sequentially examined to measure whether there were significant effects of the exogenous variable on item responses indicating the presence of DIF. If there was evidence of DIF by sex or age group or the 2 procedure groups, subsequent analyses were run separately for these groups. If there was no DIF, there were likely no systematic differences in participants' responses to the AOS and their data could be pooled. As the mean age of participants was 59.7 years, age was dichotomized into 2 categories, those younger than 60 years and those aged 60 and older, for evaluating DIF.

The CFA model also assessed the construct validity of the AOS: whether it measured pain and disability, represented by 2 underlying factors. To test whether this model adequately described the data, a chi-squared test of overall model fit, with a P value greater than .05 indicating acceptable fit, was used. In practice, the hypothesis of perfect model fit was often rejected, so the comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA) were also provided. A CFI and TLI greater than 0.90, and SRMR and RMSEA less than 0.08, indicated acceptable model fit.[17]

Item response theory (IRT) was then used to estimate item difficulty and discrimination parameters, controlling

for study participants' underlying level of ankle function. As the AOS items used a bounded, continuous visual analog scale for each item, the continuous response model (CRM) was selected. Although this model has not been used as frequently as other IRT methods, perhaps because software implementations have been limited until recently, it had an interpretation similar to those used for binary response data.[13,24-26,33,37,38] However, the CRM was appropriate for this analysis because the use of a continuous scale could provide more information (ie, more accurate estimation) about a latent construct than a graded, Likert-style response.[5,25,33]

The CRM estimated 2 parameters of interest: the item discrimination parameter a_j represented the strength of the association between item j and the latent variable it measured. The item difficulty parameter b_j was on the same scale as the latent variable and represented the level of underlying pain or disability at which a respondent would select the middle of the 100-mm scale.[13,33,37] Two CRM models would be fit, one for the pain domain and one for the disability domain.

Missing data is a common issue for PROs. Participants who responded to less than 50% of the items of the AOS were excluded from analysis, following practices adopted for low item response.[34] For the remaining participants, multiple imputation, assuming item response was missing at random,[18] was performed and all analyses were conducted using 100 imputed data sets. Rates of item nonresponse were also examined, noting that items 7 ("When you walked using shoe inserts or braces") and 8 ("When you stood wearing shoe inserts or braces") on the AOS were known to cause difficulty for respondents, because many respondents did not wear inserts or braces.[35] Sensitivity analysis was also performed to compare the results of multiple imputation with complete case analysis.

Data analysis was conducted using SAS 9.4 for data manipulation and factor analysis. R, version 3.3.3, package EstCRM, was used for the continuous IRT model,[37] and MPLUS version 7 was used for detecting Differential Item Functioning through MIMIC.

Results

Item Nonresponse and Mean Responses

Rates of missing data were low for most items (less than 7%), except for item 7 ("When you walked wearing shoe inserts or braces") and item 8 ("When you stood wearing shoe inserts or braces") from the pain domain. Approximately one-third of the sample omitted these items (36% and 35%, respectively), likely because they did not use inserts or braces. This finding is consistent with other research on this instrument.[35] Because the rate of missing data was so high for these 2 items, they were excluded from subsequent analysis.

Table 2. Summary Statistics of AOS Item Responses.

Pain Domain			
Item	Text	Mean	SD
1	At its worst?	71.3	19.1
2	Before you get up in the morning?	31.9	28.5
3	When you walked barefoot?	55.5	25.4
4	When you stood barefoot?	51.3	25.8
5	When you walked wearing shoes?	52.6	22.4
6	When you stood wearing shoes?	47.6	23.6
9	At the end of the day?	65.7	20.9
Disability Domain			
Item	Text	Mean	SD
10	Walking around the house?	44.6	22.8
11	Walking outside on uneven ground?	66.3	22.8
12	Walking 4 blocks or more?	72.3	24.6
13	Climbing stairs?	56.8	24.7
14	Descending stairs?	61.3	23.2
15	Standing on tip toes?	74.7	26.6
16	Getting of a chair	44.0	26.8
17	Climbing up or down curbs?	51.8	26.5
18	Walking fast or running?	87.1	18.9

Mean item responses ranged from 31.9 for item 2 ("before you get up in the morning") to 71.3 for item 1 ("at its worst") in the pain domain. The lowest mean item response in the disability domain was item 16 ("getting out of a chair") at 44.0 whereas the highest was item 18 ("walking fast or running") at 87.1. Approximately 20% of the sample marked the maximum score (100) for this item. See Table 2 for the mean and standard deviation of responses for each item.

Confirmatory Factor Analysis

The CFA model fit statistics produced mixed results. Although the chi-square model fit statistic of 144.9 with 86 degrees of freedom, and P value less than .001, indicating significant misfit, the CFI and TLI were both greater than 0.90 and the SRMR was less than 0.80. The RMSEA was 0.09. Modification indices and residual item correlations were investigated to determine any sources of misfit. Modification indices suggested that item 10 ("walking around the house") better fit the pain domain rather than the disability domain. Making this alteration reduced the chi-square statistic to 122.4, with associated P value .006, and reduced RMSEA to 0.07. The CFI and TLI increased slightly, and the SRMR was largely unaffected.

Results of this modified CFA model are provided in Table 3. Under this model, all factor loadings were statistically significant and ranged from 0.54 to 0.92. The latent pain and disability variables were statistically significantly and highly correlated ($r = 0.89$).

Table 3. Confirmatory Factor Analysis Results.

Model Results			
Statistic	Value		*P* Value
χ^2	122.5	86 DF	.006
CFI	0.974		
TLI	0.964		
RMSEA	0.068		
SRMR	0.054		
Domain correlation	0.886		<.001
Domain 1: Pain			
Item	Loading	SE	*P* Value
1	0.58	0.07	.00
2	0.54	0.08	.00
3	0.88	0.03	.00
4	0.77	0.04	.00
5	0.90	0.03	.00
6	0.83	0.04	.00
9	0.74	0.05	.00
10*	0.86	0.03	.00
Domain 2: Disability			
Item	Loading	SE	*P* Value
11	0.83	0.04	.00
12	0.69	0.06	.00
13	0.92	0.02	.00
14	0.89	0.03	.00
15	0.61	0.07	.00
16	0.78	0.05	.00
17	0.86	0.03	.00
18	0.62	0.07	.00

Abbreviations: CFI, comparative fit index; DF, degree of freedom; RMSEA, root mean square error of approximation; SE, standard error; SRMR, standardized root mean square residual; TLI, Tucker-Lewis index.

Table 4. Estimated Discrimination and Difficulty Parameters From the Continuous Response Model.

Domain 1: Pain				
	Discrimination Parameter (a)		Difficulty Parameter (b)	
Item	Estimate	SE	Estimate	SE
1	0.91	0.08	−1.26	0.15
2	0.61	0.08	1.03	0.22
3	1.38	0.11	−0.25	0.08
4	1.87	0.14	−0.02	0.06
5	2.31	0.18	−0.12	0.05
6	2.07	0.16	0.09	0.05
9	1.30	0.10	−0.76	0.09
10*	0.86	0.08	0.43	0.13
Domain 2: Disability				
	Discrimination Parameter (a)		Difficulty Parameter (b)	
Item	Estimate	SE	Estimate	SE
11	1.40	0.11	−0.67	0.09
12	1.12	0.09	−0.80	0.11
13	1.76	0.14	−0.14	0.06
14	1.98	0.15	−0.45	0.06
15	0.70	0.07	−1.32	0.20
16	1.00	0.08	0.29	0.11
17	1.68	0.13	−0.11	0.06
18	0.54	0.11	−2.50	0.53

Continuous IRT Model

The results of the CRM with item 10 assigned to the pain domain are provided in Table 4. The items showed a range of discrimination in the preoperative setting. Item 2 ("before you get up in the morning") and item 18 ("walking fast or running") had lower discrimination parameters than the other items, with discrimination parameters 0.61 and 0.54, respectively. Item 5 ("when you walked wearing shoes") and item 6 ("when you stood wearing shoes") had the highest discrimination parameters at 2.31 and 2.07, respectively. High discrimination is a desirable characteristic, because, in this sample, it indicates that these items are strongly associated with pain, and are good at differentiating patients with high levels of pain from those with low levels of pain.

The items in the pain domain also exhibited a range of difficulty. Item 1 ("at its worst") had the lowest difficulty estimate, at −1.26. This indicates that even patients with relatively low levels of pain would indicate some problems with pain on this item. Conversely, item 2 ("before you get up in the morning") had the highest difficulty estimate at 1.03. This means that only those individuals with high levels of pain were likely to report problems with pain before getting up in the morning, and that this would be less common among those with low levels of pain.

Items in the disability domain also exhibited a range of discrimination parameters. The items with the lowest discrimination included item 18 ("Walking fast or running") and item 15 ("Standing on tip toes"). The items with the highest discrimination included item 14 ("Descending stairs") and item 17 ("Climbing up or down curbs"). These findings suggest that among these preoperative patients, descending stairs or climbing up or down curbs were strong indicators of disability, while problems with walking/running or standing on tip toes may be less useful at discriminating between patients with ESAA waiting for surgery.

Most items in the disability domain had relatively low difficulty, which suggests most participants in this sample had at least some problems with these items. Item 18

Table 5. Differential Item Function Results.[a]

Exogenous Variable	AOS Item Identified	Effect Estimate	SE	*P* Value	Model χ^2 Difference (DF)	Model *P* Value
Sex (female)	1 ("At its worst")	−7.32	2.30	.015	5.75 (1)	.018
Age (<60 y)	5 ("When you walked wearing shoes")	−4.37	1.72	.011	6.29 (1)	.013
Procedure (fusion)	18 ("Walking fast or running?")	5.25	2.22	.018	5.43 (1)	.021

Abbreviation: AOS, Ankle Osteoarthritis Scale; DF, degree of freedom; SE, standard error.
[a]One degree of freedom for each exogenous variable.

("Walking fast or running") had the lowest difficulty estimate at −2.56. Almost everyone in this sample indicated problems with this item. The items with the highest difficulty were item 17 ("climbing up or down curbs") at −0.10 and item 16 ("getting out of a chair") at 0.34, indicating these items were more of a concern for patients with high levels of disability due to ESAA.

Differential Functioning by Age Category, Sex and Surgery

Following the MIMIC procedure, as shown in Table 5, there was also some evidence of DIF on the basis of sex (the net change of chi-square statistic is 5.75 with 1 degree of freedom change and *P* value .018) for item 1 ("at its worst"). Women underreported their pain by about 7.3 points, on average, compared to men with similar levels of underlying pain for this item. No other items demonstrated DIF by sex, suggesting a very limited sex-related measurement difference.

Including age category as an exogenous variable did statistically significantly improve the CFA model overall. The chi-squared statistic net change is 6.29 with 1 degree of freedom change and *P* value less than .013. Only item 5 ("when you walked wearing shoes") had evidence of DIF. Patients less than 60 years of age underreported their pain by about 4 points on average compared to patients older than 60.

The DIF for surgery (AA or TAR) found that patients scheduled for fusion surgery over-reported 5.3 points on item 18 ("Walking fast or running").

Although several items indicated some degree of DIF based on sex, age group, and procedure groups, the magnitude of the DIF was small (less than 10% of the 100-point scale), and would be partially evened out after averaging item scores across domains. The remaining analysis proceeded under the assumption of no differential functioning.

Sensitivity Analysis

There were no meaningful differences when using complete case analysis rather than multiple imputation, suggesting data was largely missing at random. There were no substantial differences in model fittings with CFA models when analyzing raw data and when analyzing the imputed data. The absolute difference in factor loading for item 2 was 0.035, the absolute differences for all other items were less than 0.02.

Discussion

This study examined the measurement characteristics of the AOS in a sample of patients awaiting AA or TAR in a major Canadian teaching hospital. This study observed a trend of higher scores for women and those with comorbidities, which is also consistent with previous research.[12] Interestingly, the higher scores in pain and disability on the AOS among women appears to be a real difference in this sample, and not the result of differential functioning. Only 1 item in the pain domain was identified as exhibiting DIF, and suggested that women underreported pain relative to men on that item. This suggests that the women in this sample were more strongly affected by ESAA than were men. Participants in this study had generally high levels of preoperative pain and disability as measured by the AOS, with mean scores in the range of 50 to 60 points, typical of patients planning to undergo AA or TAR.[6,11]

Item nonresponse was low for the AOS, with the exception of items 7 and 8 that ask about using shoe inserts or braces, which more than one-third of the sample omitted. Because nonresponse to these items has been documented previously, it is suggested that future administrations of the AOS omit these items.

Although the results of the CFA model were somewhat mixed, after moving item 10 to the pain domain, all of the model fit statistics except the chi-squared test met the recommended threshold demonstrating adequate model fit. In addition, factor loadings were generally strong and statistically significant. This study confirms the pain and disability construct validity of the AOS when excluding items 7 and 8, and moving item 10 to the pain domain. It is unclear why item 10 was more strongly associated with pain in this sample than with disability. Previous research proposed eliminating items that were highly correlated to each other and creating 2 redefined domains of measurement.[35] This study affirms that the AOS, as originally constructed, offers robust

information when one accounts for items 7, 8, and 10. Future research, especially those with larger samples or from different clinical settings, should further investigate these issues.

The CRM analysis provided important information demonstrating the generally positive measurement qualities of the AOS. Although the items exhibited a range of discrimination, most items were strongly associated with the latent variable they intended to measure. This means that most of the items of the AOS appear to be relevant to patients, were able to accurately measure pain and disability, and differentiate between patients with higher levels of pain and disability from those with lower levels.

As this study is based on preoperative patients with ESAA, item 18 ("Walking fast or running") may have been anticipated to have had low discrimination as most patients report a high level of pain. Analyses of postoperative data may reveal that item 18 has high discrimination, indicative of pain relief after surgery. Item 2 ("Before you get up in the morning"), however, may be less useful as it appears to perform poorly and be less clinically relevant. The difficulty estimates provide an indication of which items are most relevant only to patients with high levels of pain or disability, such as item 2 ("Before you get up in the morning"), item 10 ("Walking around the house"), and item 16 ("Getting out of a chair"). This finding is in line with the clinical progression of the disease. Those with more advanced ankle arthritis, such as participants scheduled for surgery in this study, may experience pain even at rest or difficulty with basic tasks.

While this study did uncover evidence of DIF by sex and age category, it appears that the effect would be small. Both age and sex only produced DIF on 1 item of 16, and the magnitude was only 4.4 to 7.3 points of 100. Surgery presents DIF on item 18, though the magnitude of the effects are likely ignorable. This suggests that the AOS is appropriate for measuring ankle pain and disability for men and women, for both younger and older patients and for patients waiting for fusion and replacement surgery. However, future research should continue to monitor for evidence of DIF, and determine whether any scoring adjustments or changes to the instrument are necessary.

There are several limitations to this study. First, the response rate was low, although not unusual for studies of this kind,[7] and the potential for selection bias cannot be eliminated. Including nonsurgically managed patients may have strengthened the generalizability of the study's findings. Second, this study excluded items 7 and 8 from analysis, and shifted item 10 to the pain domain in the CFA and IRT models. This means the results from this study for the pain domain are not comparable to studies that included these 2 items. As this sample was limited to patients with ESAA and scheduled for elective ankle fusion or replacement surgery in Vancouver, Canada, the results may not generalize to patients in other clinical contexts, such as patients undergoing nonoperative management of ankle osteoarthritis, those waiting for other procedures, or in the postoperative state, nor could this study untangle the role of surgeon in item's responses. Although the amount of missing data for included items was slight, it is possible that the study's treatment of missing data was not appropriate, possibly affecting the findings. Finally, patients in other countries with different models of access to care may also have a different response.

Conclusion

Our evaluation of the AOS found the instrument to be a strong measure of the effect of ankle arthritis on pain and dysfunction, even when removing items 7 and 8, supporting its prior use in numerous clinical studies. This study supports that 16 of the 18 AOS items have psychometric properties useful for baseline clinical assessment of patients with end stage ankle arthritis. Future research should investigate which items best capture functional change after operative management with fusion or replacement and whether a shortened instrument, such as the Ankle Arthritis Score, has the same positive measurement characteristics.

Acknowledgments

This study was funded by the Canadian Institutes for Health Research (CIHR) and in-kind support of Vancouver Coastal Health (VCH) Authority. The last author is a Scholar of the Michael Smith Foundation for Health Research (MSFHR). CIHR, VCH and MSFHR had no role in developing the methods, data analyses, interpreting the results or manuscript preparation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. ICMJE forms for all authors are available online.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Government of Canada; Canadian Institutes of Health Research; and Institute of Health Services and Policy Research.

ORCID iD

Jason M. Sutherland, PhD, https://orcid.org/0000-0002-1857-2432

References

1. Allen KD, Golightly YM. Epidemiology of osteoarthritis: state of the evidence. *Curr Opin Rheumatol*. 2015;27(3):276-283.
2. Angers M, Svotelis A, Balg F, Allard JP. Cross-cultural adaptation and validation of the Ankle Osteoarthritis Scale for

use in Frenchspeaking populations. *Can J Surg*. 2016;59(2): 123-127.
3. Barg A, Pagenstert G, Horisberger M, et al. Supramalleolar osteomies for degenerative joint disease of the ankle joint: indication, technique and results. *Int Orthop*. 2013;37(9):1683-1695.
4. Barg A, Pagenstert GI, Hügle T, et al. Ankle osteoarthritis: etiology, diagnostics, and classification. *Foot Ankle Clin*. 2013;18(3):412-426.
5. Bejar II. An application of the continuous response level model to personality measurement. *Appl Psychol Meas*. 1977;1(4):509-521.
6. Bouchard M, Amin A, Pinsker E, Khan R, Deda E, Daniels TR. The impact of obesity on the outcome of total ankle replacement. *J Bone Joint Surg Am*. 2015;97(11):904-910.
7. Brems C, Johnson ME, Warner T, Roberts LW. Survey return rates as a function of priority versus first-class mailing. *Psychol Rep*. 2016;99(2):496-501.
8. Conti S, Wong Y. Complications of total ankle replacement. *Clin Orthop Relat Res*. 2001;391:105-114.
9. Crocker L, Algina J. Introduction to classical and modern test theory. Belmont, CA: Wadsworth Thomson Learning; 1986.
10. Croft S, Wing KJ, Daniels TR, et al. Association of ankle arthritis score with need for revision surgery. *Foot Ankle Int*. 2017;38(9):939-943.
11. Daniels TR, Younger AS, Penner M, et al. Intermediate-term results of total ankle replacement and ankle arthrodesis. *J Bone Joint Surg Am*. 2014;96(2):135-142.
12. Domsic RT, Saltzman CL. Ankle Osteoarthritis Scale. *Foot ankle Int*. 1998;19(7):466-471.
13. Ferrando PJ. Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behav Res*. 2002;37(4):521-542.
14. Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores : true differences or differential item functioning? *Med Care*. 2003;41(7):III75-III86.
15. Gagné O, Veljkovic A, Glazebrook M, et al. Prospective cohort study on the employment status of working age patients after recovery from ankle arthritis surgery. *Foot Ankle Int*. 2018;39(6):657-663.
16. Glazebrook M, Daniels T, Younger A, et al. Comparison of health-related quality of life between patients with end-stage ankle and hip arthrosis. *J Bone Joint Surg Am*. 2008;90(3): 499-505.
17. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model*. 1999;6(1):1-55.
18. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. New York, NY: Wiley-Interscience; 2002.
19. Oort FJ. Simulation study of item bias detection with restricted factor analysis. *Struct Equ Model*. 1998;5(2):107-124.
20. Pinsker E, Inrig T, Daniels TR, Warmington K, Beaton DE. Reliability and validity of 6 measures of pain, function, and disability for ankle arthroplasty and arthrodesis. *Foot Ankle Int*. 2015;36(6):617-625.
21. Rahman MM, Cibere J, Goldsmith CH, Anis AH, Kopec JA. Osteoarthritis incidence and trends in administrative health records from British Columbia, Canada. *J Rheumatol*. 2014;41(6):1147-1154.
22. Robinson A, Keith T. Osteoarthritis of the ankle. *Orthop Trauma*. 2016;30(1):59-67.
23. Saltzman C, Salamon M, Blanchard G, et al. Epidemiology of ankle arthritis: report of a consecutive series of 639 patients from a tertiary orthopaedic center. *Iowa Orthop J*. 2005;25:44-46.
24. Samejima F. A use of the information function in tailored testing. *Appl Psychol Meas*. 1977;1(2):233-247.
25. Samejima F. Homogeneous case of the continious response model. *Psychometrika*. 1973;38(2):203-219.
26. Shojima K. A noniterative item parameter solution in each EM cycle of the continuous response model. *Educ Technol Res*. 2005;28(1-2):11-22.
27. Sutherland JM, Crump RT, Chan A, Liu G, Yue E, Bair M. Health of patients on the waiting list: opportunity to improve health in Canada? *Health Policy (New York)*. 2016;120(7):749-757.
28. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Med Care*. 2006;44(11)(suppl 3):S152-S170.
29. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res*. 2016;16(suppl 1):33-42.
30. Turkiewicz A, Petersson IF, Björk J, et al. Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. *Osteoarthritis Cartilage*. 2014;22(11):1826-1832.
31. Valderrabano V, Horisberger M, Russell I, Dougall H, Hintermann B. Etiology of ankle osteoarthritis. *Clin Orthop Relat Res*. 2009;467(7):1800-1806.
32. Veltman ES, Hofstad CJ, Witteveen AGH. Are current foot- and ankle outcome measures appropriate for the evaluation of treatment for osteoarthritis of the ankle? Evaluation of ceiling effects in foot- and ankle outcome measures. *Foot Ankle Surg*. 2017;23(3):168-172.
33. Wang T, Zeng L. Item parameter estimation for a continuous response model using an EM algorithm. *Appl Psychol Meas*. 1998;22:333-344.
34. Ware JE, Snow KK, Kosinski M, Gandek B. Scoring the SF-36. In: *SF-36 Health Survey Manual and Interpretation Guide*. Boston, MA: Nimrod Press; 1993:1-22.
35. Wing KJ, Chapinal N, Coe MP, et al. Measuring the operative treatment effect in end-stage ankle arthritis: are we asking the right questions? A COFAS multicenter study. *Foot Ankle Int*. 2017;38(10):1064-1069.
36. Wood P, Deakin S. Total ankle replacement. The results in 200 ankles. *J Bone Joint Surg Br*. 2003;85(3):334-341.
37. Zopluoglu C. EstCRM: an R package for Samejima's continuous IRT model. *Appl Psychol Meas*. 2012;36(2):149-150.
38. Zopluoglu C. A comparison of two estimation algorithms for Samejima's continuous IRT model. *Behav Res Methods*. 2013;45(1):54-64.