JB&JS OPEN ACCESS

Efficiency and Accuracy of Computerized Adaptive Testing for the Oswestry Disability Index and Neck Disability Index

Tracy Y. Zhu, PhD, Otho R. Plummer, PhD, Audrey Hunt, BS, and Alexander Joeris, MD, MSc

Investigation performed at the AO Innovation Translation Center, AO Foundation, Davos, Switzerland, and Universal Research Solutions, LLC, Columbia, Missouri

Background: This study aimed to determine the efficiency and accuracy of computerized adaptive testing (CAT) models of the Oswestry Disability Index (ODI) and Neck Disability Index (NDI).

Methods: The study involved simulation using retrospectively collected real-world data. Previously developed CAT models of the ODI and NDI were applied to the responses from 52,551 and 18,196 patients with spinal conditions, respectively. Efficiency was evaluated by the reduction in the number of questions administered. Accuracy was evaluated by comparing means and standard deviations, calculating Pearson r and intraclass correlation coefficient (ICC) values, plotting the frequency distributions of CAT and full questionnaire scores, plotting the frequency distributions of differences between paired scores, and Bland-Altman plotting. Score changes, calculated as the postoperative ODI or NDI scores minus the preoperative scores, were compared between the CAT and full versions in patients for whom both preoperative and postoperative ODI or NDI questionnaires were available.

Results: CAT models of the ODI and NDI required an average of 4.47 and 4.03 fewer questions per patient, respectively. The mean CAT ODI score was 0.7 point lower than the full ODI score (35.4 ± 19.0 versus 36.1 ± 19.3), and the mean CAT NDI score was 1.0 point lower than the full NDI score (34.7 ± 19.3 versus 33.8 ± 18.5). The Pearson r was 0.97 for both the ODI and NDI, and the ICC was 0.97 for both. The frequency distributions of the CAT and full scores showed marked overlap for the ODI and NDI. Differences between paired scores were less than the minimum clinically important difference in 98.9% of cases for the ODI and 98.5% for the NDI. Bland-Altman plots showed no proportional bias. The ODI and NDI score changes could be calculated in a subgroup of 6,044 and 4,775 patients, respectively; the distributions of the ODI and NDI score changes were near identical between the CAT and full versions.

Conclusions: CAT models were able to reduce the question burden of the ODI and NDI. Scores obtained from the CAT models were faithful to those from the full questionnaires, both on the population level and on the individual patient level.

Level of Evidence: Prognostic Level III. See Instructions for Authors for a complete description of levels of evidence.

Low back pain and neck pain are among the most common musculoskeletal problems associated with substantial economic burdens and risks of disability[1-4]. The assessment of pain and disability is essential for planning and monitoring interventions for such pain. Patient-reported outcome measures (PROMs) are standardized validated questionnaires that measure patients' perceptions of their health status, function, and well-being. The Oswestry Disability Index (ODI) and Neck Disability Index (NDI) are 2 widely used self-administered PROMs for patients with spinal conditions. The ODI contains 10 items addressing the intensity of back pain and its impact on various daily activities; each has 6 answer options ranging from no disability (scored as 0) to severe disability (scored as 5)[5]. The NDI consists of 5 items derived from the ODI and 5 new items to assess the impact of neck pain on patients' daily activities[6]. Both questionnaires have been culturally adapted in multiple languages and demonstrated fair to good psychometric properties for various spinal conditions[6-14].

There is growing interest in integrating PROMs into research and routine clinical practice, which is driven by the evidence that they

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSOA/A469).

can improve the care of individual patients by providing patient-centered information, facilitating patient-clinician communication, and monitoring the effects of treatment[15]. However, routine collection of PROMs could be time-consuming for patients and healthcare providers. Computerized adaptive testing (CAT) is an emerging solution. CAT is a dynamic computerized process in which questions are selected from a set of items on the basis of the respondent's level of the trait that is being measured, such as pain or function[16]. Because the questions that CAT presents to the respondents are thus tailored on the basis of their previous responses, different respondents answer different questions. By selecting the most informative, respondent-tailored, specific question items out of the full question item bank, the same amount of information can be obtained with fewer items, ensuring the accuracy of the assessment while minimizing the time and scoring burden.

CAT has been increasingly adopted in the medical field and provides an efficient alternative to standardized full questionnaires[16-19]. OBERD software (Universal Research Solutions, www.oberd.com) is health intelligence software for collection of patient outcome data. We have previously used OBERD software to develop and validate several CAT models of common PROMs in orthopaedics[20-24]. The objective of this study was to determine whether the CAT models of the ODI and NDI could improve efficiency compared with their original full questionnaires while maintaining the accuracy of scores.

Materials and Methods

Study Design and Patients

The study involved simulation using retrospectively collected real-world data. The "test sets" used responses to and scores on the ODI and NDI questionnaires retrieved from a database of patients with spinal conditions from several large orthopaedic clinics in the United States at which PROMs were a standard component of patient care and follow-up. The patients completed the full version of the ODI or NDI using the OBERD software, either remotely via the internet or using a tablet device (iPad, Apple) during their clinic visits. Their responses and scores are retained in an online database for use by the medical staff. This database offered an exceptional opportunity to test the CAT versions of these instruments.

There were no restrictions with respect to patient age, sex, height, weight, diagnoses, visit types (new patient or follow-up, preoperative or postoperative), treatments, and procedures. Non-English-speaking patients and those who did not respond to all questions were excluded. Ethical approval and informed consent were not required as patients of the clinics gave consent to subsequent use of their data for retrospective research in a de-identified and aggregated way.

CAT Models

The CAT models of the ODI and NDI were developed within OBERD through prior training using "training sets" of de-identified responses to the questionnaires obtained from 17,808 patients for the ODI and 23,636 patients for the NDI. The "training sets" and "test sets" were randomly selected from the same database of patients with spinal conditions using the same eligibility criteria. There were no cases in common between the training and test sets. Details of the CAT models have been described elsewhere[22,23]. The CAT models were retrospectively applied to each patient in the test sets; responses were supplied from the patient's stored questionnaires, resulting in a CAT score for the ODI or NDI. The patients did not complete the CAT versions of the questionnaires. The CAT ODI or NDI score for each patient was paired with his or her full ODI or NDI score.

Data Analysis

The efficiency of the CAT models of the ODI and NDI was evaluated by the reduction in the number of questions administered.

To aid in interpretation, the accuracy of the CAT models was evaluated in the context of the minimal clinically important difference (MCID) for the ODI and NDI. The MCID is the smallest change in score perceived by the patient as important. Previous literature reported various MCIDs for the ODI and NDI, largely depending on the patient characteristics, diagnoses, treatments, and calculation methods[8,25-34]. For this study, we used the MCIDs from the study by Hung et al., which included a more generalizable patient population with a wide range of spinal conditions and treatment types[26]. Hung et al. used various methods to calculate MCIDs, and the present study used the ones calculated using the mean change scores (an anchor-based method) at the 6-month follow-up: 18.5 points for the ODI and 12.8 points for the NDI[26]. The percentages of paired-score differences (full score minus CAT score for the same patient) that fell outside of the range of ±1 MCID were calculated.

Additional analyses were performed to determine the accuracy of the CAT scores. First, the mean and standard deviation (SD) were compared between the CAT and full scores. Second, the Pearson correlation coefficient (r) was calculated to determine the strength of a linear relationship between the 2 scores, and the intraclass correlation coefficient (ICC) was calculated to evaluate the similarity between the 2 scores. Third, the frequency distributions of the CAT and full scores were plotted and compared to examine whether their distributions were similar. Fourth, the frequency distribution of the paired-score differences was plotted to examine the similarity between the 2 scores at the individual patient level. Fifth, Bland-Altman plots were generated to assess the agreement between the CAT and full scores[35]. In the Bland-Altman plot, the paired-score difference was plotted against the mean of the 2 paired scores. Finally, in subgroups of patients who underwent an operation for which both preoperative and postoperative ODI or NDI patient questionnaires were available (6,044 patients for the ODI and 4,775 for the NDI), score changes (postoperative ODI or NDI score minus preoperative score) were calculated separately for the CAT and full versions and then compared between them to determine whether the CAT models could identify changes in patient condition, reflected in the changes in the total score, that were consistent with the full questionnaires.

Data analyses were performed using R (version 3.4.3; R Foundation for Statistical Computing), Python (version 3.4.5; Python Software Foundation), and spreadsheets.

TABLE I Patient Age and Sex Distribution

Variable	Oswestry Disability Index	Neck Disability Index
No. of patients	52,551	18,196
Age *(yr)*		
Mean ± std. dev.	56.45 ± 15	55.55 ± 13.7
Median	57	56
Min., max.	10, 102	9, 99
Age group *(no. [%])*		
Min.-30 years	2,956 (5.6)	795 (4.4)
31-45 years	9,408 (17.9)	3,192 (17.5)
46-60 years	17,990 (34.2)	7,737 (42.5)
61-75 years	17,453 (33.2)	5,219 (28.7)
76-max. years	4,744 (9.0)	1,253 (6.9)
Sex *(no. [%])*		
Women	27,056 (51.5)	10,130 (55.7)
Men	25,467 (48.5)	8,061 (44.3)
Unknown	28 (0.05)	5 (0.03)

Source of Funding
This work was partially funded by the AO Foundation, an independent medically guided not-for-profit organization, and Tracy Y. Zhu and Alexander Joeris are employees of the AO Foundation. Otho R. Plummer and Audrey Hunt are employees of Universal Research Solutions, LLC, which is the developer of the OBERD software used in the study.

Results

Patient Characteristics

The test sets included 52,551 ODI and 18,196 NDI questionnaires gathered between 2014 and 2020. The mean and SD, median, minimum, and maximum values for age were similar between the patients who completed the ODI and those who completed the NDI, while the percentage distribution among age groups varied slightly (Table I). For both the ODI and NDI, women represented slightly more than half of the sample.

Efficiency

The CAT models determined that the most effective first question to ask was "How does your back pain affect your social life?" for the ODI and "How does your neck pain affect your recreation?" for the NDI. The CAT models of the ODI and NDI required an average of 5.53 and 5.97 questions, respectively, to be administered per patient. The CAT model of the ODI required 58.2% of patients to answer 5 questions, 30.9% to answer 6 questions, and 10.9% to answer 7 questions; the percentages for the CAT model of the NDI were 32.3%, 38.3%, and 29.4%, respectively. The overall reduction in the question burden was 44.7% for the ODI and 40.3% for the NDI.

Accuracy

There were minimal differences between the mean CAT ODI and NDI scores and the mean full ODI and NDI scores, and their SDs were also similar (Table II). The CAT ODI score was an average of 0.7 point (SD, 4.4 points) lower than the full ODI score. The CAT NDI score was an average of 1.0 point (SD: 5.0 points) lower than the full NDI score. The Pearson r was 0.97 for both the ODI and NDI, indicating near-perfect correlations between the CAT and full scores. These correlations were independent of the number of questions administered (Fig. 1). The ICC was 0.97 for both the ODI and NDI.

The distributions of the CAT scores were extremely similar to those of the full scores for both the ODI and NDI (Fig. 2). The paired-score differences for the ODI ranged from −22.9 to 28.2 points, and their distribution clustered around zero and was relatively symmetric (Fig. 3-A). Only 1.1% of the paired-score differences for the ODI were outside of the range of ±1 MCID. The paired-score differences for the NDI ranged from −18.4 to 35.2, and 1.5% were outside of the range of ±1 MCID (Fig. 3-B). Bland-Altman plots showed no proportional bias, as the differences were clustered around 0 with no consistent pattern (Fig. 4).

Comparisons of Score Changes Between CAT and Full Versions

There were minimal differences in mean score changes between the CAT and full versions (Table III). The correlation coefficient (r) between the score changes for the 2 versions was 0.97 for the ODI and 0.96 for the NDI. The distributions of the score changes for the CAT and full versions were near identical for both the ODI and NDI (Fig. 5). The between-version differences in ODI and NDI score changes had symmetric distributions with means of approximately 0.

TABLE II Summary Statistics for the Oswestry Disability Index and Neck Disability Index*

Questionnaire	No. of Patients	Mean ± Standard Deviation		Pearson r	ICC
		CAT Version	Full Version		
Oswestry Disability Index	52,551	35.4 ± 19.0	36.1 ± 19.3	0.97	0.97
Neck Disability Index	18,196	34.7 ± 19.3	33.8 ± 18.5	0.97	0.97

*CAT = computerized adaptive testing, ICC = intraclass correlation coefficient.

JBJS Open Access • 2023:e22.00036. **openaccess.jbjs.org** 4

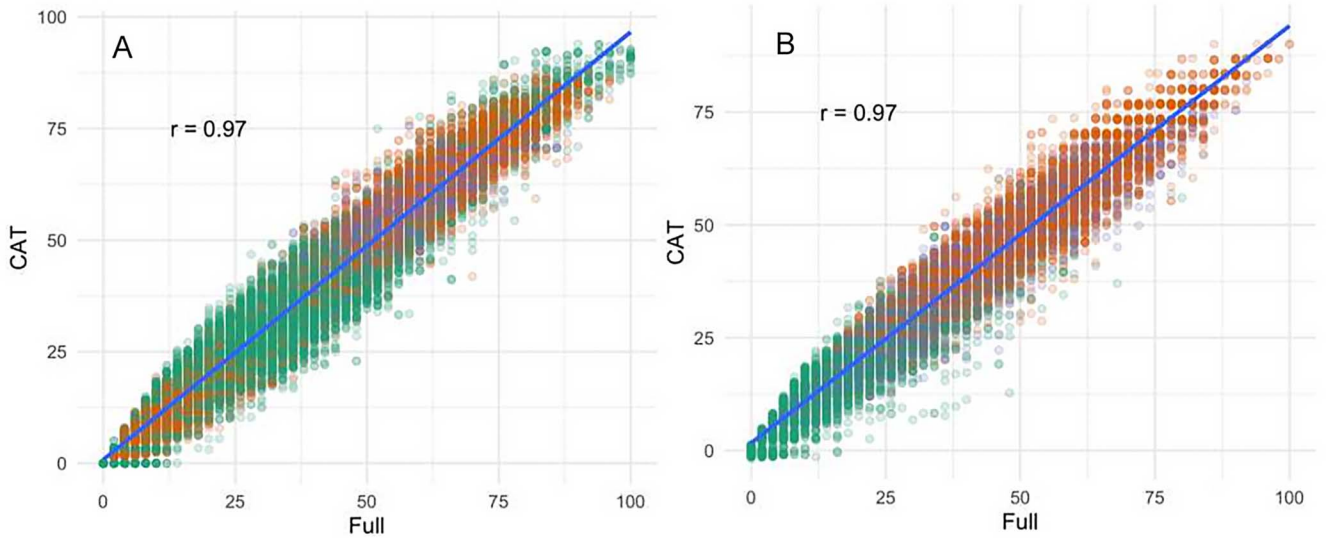

Fig. 1

Scatterplots of the full ODI score versus the CAT ODI score (**Fig. 1-A**) and the full NDI score versus the CAT NDI score (**Fig. 1-B**); darker dots represent more patients with that combination of full score and CAT score. The number of questions asked by the CAT model is indicated by the colors of the dots. Green dots: 5 questions asked; orange dots: 6 questions asked; purple dots: 7 questions asked. ODI = Oswestry Disability Index, NDI = Neck Disability Index, CAT = computerized adaptive testing.

Discussion

Several studies have investigated the validity of using CAT models to measure PROMs in the field of orthopaedics[17,18,36-40]. The CAT models in these studies relied on a bank of questionnaire items calibrated to an item response theory (IRT) model[16]. The CAT models in the present study differed from those used in these previous studies in that they relied on machine learning algorithms rather than IRT and utilized the items from a legacy questionnaire rather than from a calibrated item bank[20-24]. A weighted decision tree scheme was used for picking the next question, with the goal of generating a total score that was interchangeable with the total score obtained from the full questionnaire while reducing the question burden. This approach did not require developing difficulty rankings for individual item or other traits of IRT. The CAT scores from the models inherit the psychometric properties of the full scores. This avoids the assumption that the CAT score is additively obtained from the individual question responses; therefore, evaluation of the validity of the CAT models was focused on their efficiency and accuracy.

Reducing patient and staff burden is the key to effective integration of PROMs into the pre-existing clinical routine. For both questionnaires, the CAT models reduced the number of questions administered by >40%, to an average of <6 questions per patient. The reduction in question burden appeared to be greater for the ODI CAT model, with nearly 60% of patients being required to answer only 5 questions. This reduction may seem trivial; however, considering the amount of information

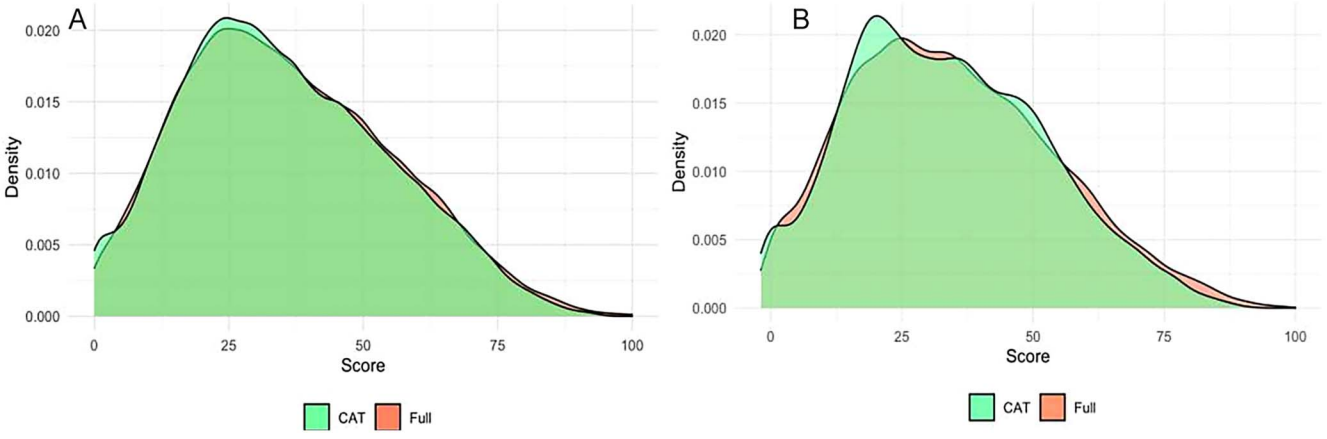

Fig. 2

Distributions of the full and CAT version scores for the Oswestry Disability Index (**Fig. 2-A**) and Neck Disability Index (**Fig. 2-B**). CAT = computerized adaptive testing.

JBJS Open Access • 2023:e22.00036. **openaccess.jbjs.org** 5

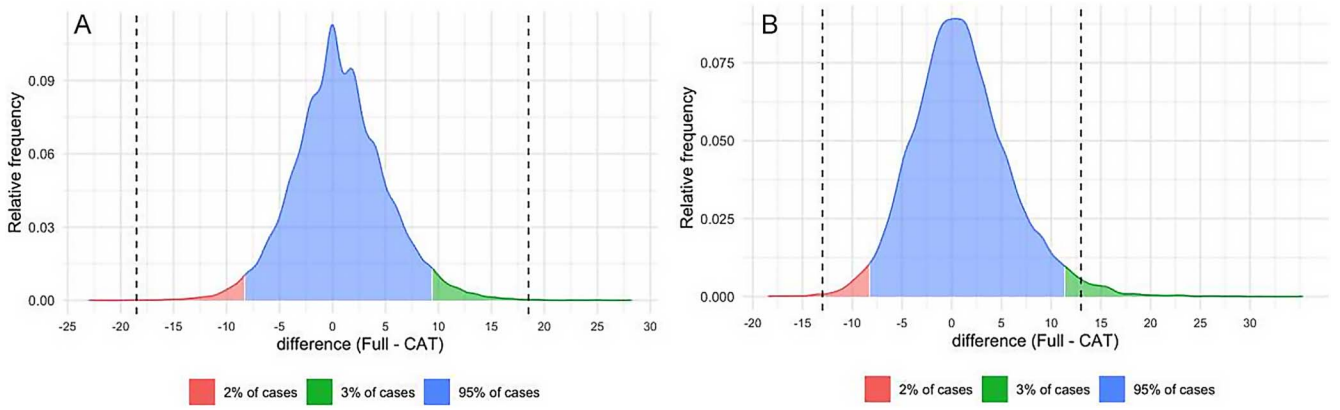

Fig. 3

Frequency distributions of the score difference between the full and CAT versions of the Oswestry Disability Index (**Fig. 3-A**) and Neck Disability Index (**Fig. 3-B**) for each patient. The vertical dashed lines show minimally clinically important differences. CAT = computerized adaptive testing.

that a patient frequently must provide at each visit, any effort to streamline the data collection and minimize patients' cognitive burden will potentially improve their experience and compliance while speeding up their preparation for the interactions with the surgeon, particularly when the patient volume is large. If a computer system is in place, or envisioned, at the clinic for patient data collection and analyses, the incremental costs of implementing such a CAT system could be nominal and would very likely be offset by the reduction in staff burden and the improved efficiency of the clinical workflow. On the other hand, the costs of implementing the CAT system at a clinic from scratch are largely independent of the patient volume. Cost savings are expected to increase with increasing size of the practice, and increasing size results in increasing staff burden for PROM collection.

The CAT scores had a marked resemblance to the full scores. On the population level, this was reflected by the similarity of the means and SDs as well as the nearly identical distributions. The high degree of overlap between the distributions of the CAT and full scores was a particularly important demonstration of the resemblance, given that the distributions were not normal. The ICCs for the ODI and NDI were both >0.90, indicating that most of the variance in the data was between patients rather than between full and CAT scores[41]. The accuracy of the CAT scores was evaluated in the context of the MCID. The MCID is particularly important for PROMs, for which the clinical importance of a given change may not be obvious to clinicians[42]. On the individual patient level, the absolute difference between the full and CAT scores was less than the MCID in nearly 99% of cases for both the ODI and NDI. The paired-score differences were symmetrically distributed and clustered around zero. Paired-score differences did not depend on the magnitude of the scores. The distributions of the score changes from before to after the operation were also nearly identical between the full and CAT versions, suggesting that the CAT versions accurately

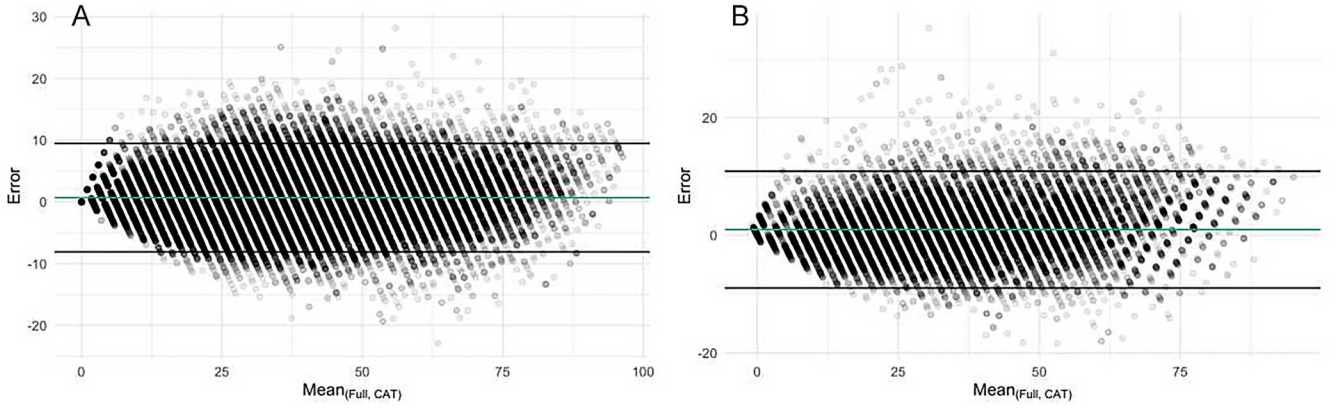

Fig. 4

Bland-Altman plots for the Oswestry Disability Index (**Fig. 4-A**) and Neck Disability Index (**Fig. 4-B**). The score difference between the full and CAT versions for each patient is plotted against the mean of the 2 scores. Each dot (data point) represents a combination of the score difference and the mean of the 2 scores. Darker dots represent more patients with that combination (of the score difference and the mean of the 2 scores). The green line indicates the mean difference, and the interval between the 2 black lines indicates the 95% confidence interval of the mean. There is no trend of greater means at more negative or more positive differences, indicating that there is no proportional bias. CAT = computerized adaptive testing.

JBJS Open Access • 2023:e22.00036. **openaccess.jbjs.org** 6

TABLE III Score Changes for the Oswestry Disability Index and Neck Disability Index*

Variable	No. of Patients	Mean ± SD	Range	Median (IQR)
Oswestry Disability Index				
Score change, full version	6,044	−24.0 ± 21.4	−96.0, 56.0	−24.0 (−38.0, −8.0)
Score change, CAT version	6,044	−23.4 ± 21.4	−87.8, 53.1	−22.7 (−37.7, −8.1)
Difference in score change: full minus CAT version	6,044	−0.6 ± 5.1	−22.2, 23.1	−0.5 (−3.8, 2.7)
Neck Disability Index				
Score change, full version	4,775	−14.4 ± 19.0	−78.0, 64.0	−14.0 (−28.0, −2.0)
Score change, CAT version	4,775	−13.9 ± 19.4	−78.2, 64.2	−13.0 (−26.8, −1.3)
Difference in score change: full minus CAT version	4,775	−0.5 ± 5.7	−24.6, 21.3	−0.3 (−4.1, 3.1)

*From preoperative to postoperative visit. SD = standard deviation, IQR = interquartile range, CAT = computerized adaptive testing.

reflected the changes in patient health status evaluated with the full versions. In summary, these findings show that the CAT scores were faithful representations of the full scores, both on the population level and on the individual patient level.

The greatest strength of the study was the large sample size. Our CAT models were based on machine learning, whose performance relies heavily on the size and diversity of the available training data. More importantly, the large test set ensured a more accurate evaluation of model performance. In addition, the study included patients with a broad spectrum of characteristics, such as age, diagnosis, and treatment, which maximized the generalizability of our results.

Nonetheless, a few methodological limitations should be considered. First, as the study involved a simulation based on real-world data, the performance of the models was tested using previously stored responses rather than with live patients; therefore,

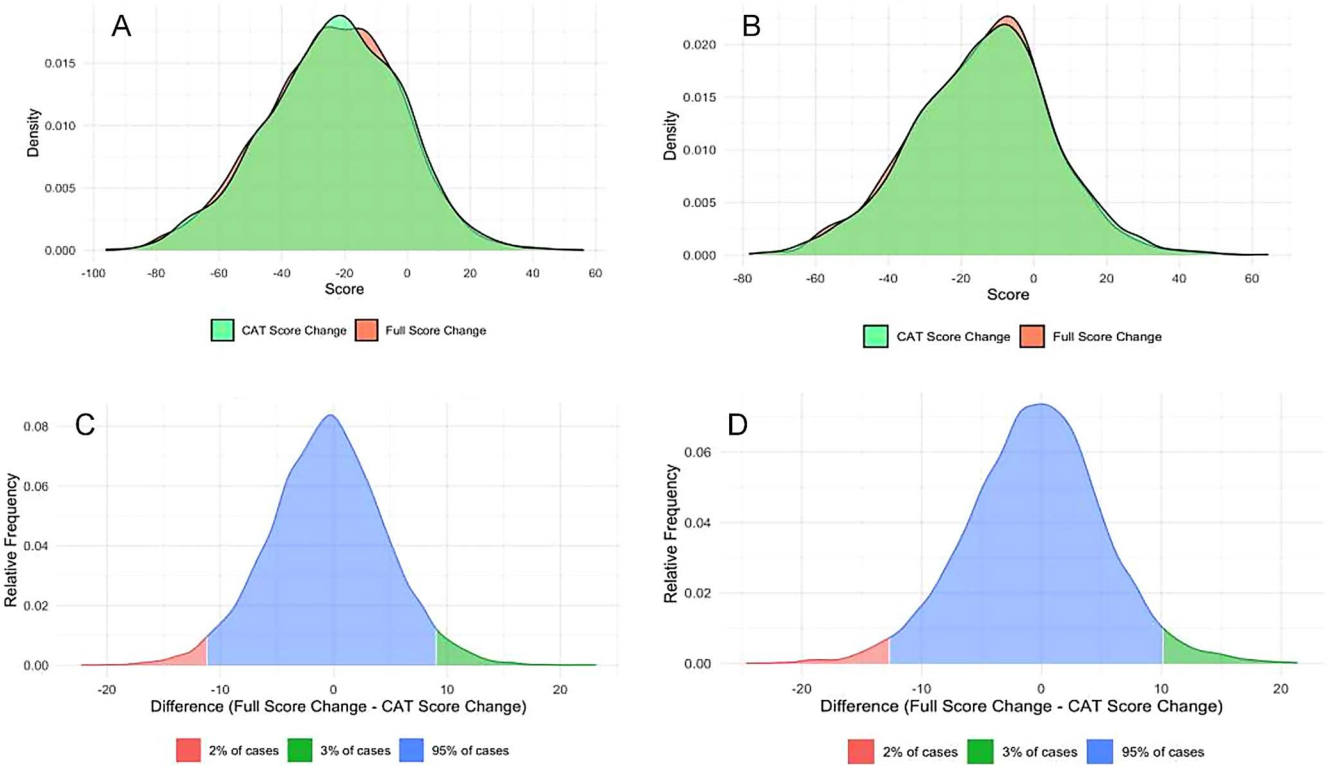

Fig. 5

Distributions of the full and CAT version score changes, calculated as the postoperative score minus the preoperative score, for the Oswestry Disability Index (**Fig. 5-A**) and Neck Disability Index (**Fig. 5-B**), and distributions of the difference in score change between the full and CAT versions for the Oswestry Disability Index (**Fig. 5-C**) and Neck Disability Index (**Fig. 5-D**). CAT = computerized adaptive testing.

the study could not evaluate the actual time saved in completing the questionnaire and the reduction in the staff burden. However, the number of questions saved could be a reasonable proxy for efficiency. It is also worth noting that the patient's experience of completing the CAT model questionnaire resembles that of completing the full questionnaire because, in both situations, the tablet device delivers 1 item at a time to the screen, and the patient is unaware of the version of the PROM. Second, only the first question in the CAT model is chosen by the algorithm as the most informative and it is always the same; thus, tracking the other questions would be problematic in any CAT model. In view of the large number of possible combinations of asked and unasked questions, no attempt was made to predict the responses to unasked questions; instead, the study specifically addressed the situation in which the total score was the desired information. Third, complete generalizability of the results across specific patient characteristics cannot be guaranteed because factors such as diagnosis, demographics, and disease characteristics were not examined. However, the large size of our sample suggests that CAT is generally applicable across the many conditions to which the ODI and NDI are applied. The diagnosis or other population characteristics may be included as independent variables in future work testing more focused populations, possibly providing additional gains in accuracy. Finally, because the number and sequence of questions differ from patient to patient with CAT models, the possibility exists that errors would be introduced if the question order is important. Such an impact must be small, in view of the accuracy of the CAT models, but it has not been entirely ruled out or quantified by this study.

In conclusion, scores obtained from CAT models developed with machine learning methods were faithful to those obtained from the full questionnaires on both the population and individual patient levels. CAT models of the ODI and NDI were shown to be reliable and compatible alternatives to the full versions of these important questionnaires to collect patient-reported outcomes in patients with spinal conditions. ■

Tracy Y. Zhu, PhD[1]
Otho R. Plummer, PhD[2]
Audrey Hunt, BS[2]
Alexander Joeris, MD, MSc[1]

[1]Clinical Science, AO Innovation Translation Center, AO Foundation, Davos, Switzerland

[2]Universal Research Solutions, LLC, Columbia, Missouri

Email for corresponding author: alexander.joeris@aofoundation.org

References

1. Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. Eur Spine J. 2006 Jun;15(6):834-48.
2. Hart LG, Deyo RA, Cherkin DC. Physician office visits for low back pain. Frequency, clinical evaluation, and treatment patterns from a U.S. national survey. Spine (Phila Pa 1976). 1995 Jan 1;20(1):11-9.
3. Hoy D, Bain C, Williams G, March L, Brooks P, Blyth F, Woolf A, Vos T, Buchbinder R. A systematic review of the global prevalence of low back pain. Arthritis Rheum. 2012 Jun;64(6):2028-37.
4. Strine TW, Hootman JM. US national prevalence and correlates of low back and neck pain among adults. Arthritis Rheum. 2007 May 15;57(4):656-65.
5. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976). 2000 Nov 15;25(22):2940-52, discussion 2952.
6. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991 Sep;14(7):409-15.
7. Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. Spine J. 2017 Mar;17(3):321-7.
8. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil. 2008 Jan;89(1):69-74.
9. Comins J, Brodersen J, Wedderkopp N, Lassen MR, Shakir H, Specht K, Brorson S, Christensen KB. Psychometric Validation of the Danish Version of the Oswestry Disability Index in Patients With Chronic Low Back Pain. Spine (Phila Pa 1976). 2020 Aug 15;45(16):1143-50.
10. Hung M, Cheng C, Hon SD, Franklin JD, Lawrence BD, Neese A, Grover CB, Brodke DS. Challenging the norm: further psychometric investigation of the Neck Disability Index. Spine J. 2015 Nov 1;15(11):2440-5.
11 Lochhead L, MacMillan P. Rasch analysis of the Oswestry Disability Index. In: Proceedings of the 14th International Congress on Circumpolar Health, July 11-16, 2009, Yellowknife, Canada. International Journal of Circumpolar Health. 2010; 69(sup7):1-598. Paper no. 210.
12. Lochhead LE, MacMillan PD. Psychometric properties of the Oswestry Disability Index: Rasch analysis of responses in a work-disabled population. Work. 2013 Jan 1; 46(1):67-76.
13. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, Goldsmith CH. Measurement properties of the Neck Disability Index: a systematic review. J Orthop Sports Phys Ther. 2009 May;39(5):400-17.
14. Saltychev M, Mattie R, McCormick Z, Bärlund E, Laimi K. Psychometric properties of the Oswestry Disability Index. Int J Rehabil Res. 2017 Sep;40(3):202-8.
15. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. Patient reported outcome measures in practice. BMJ. 2015 Feb 10;350:g7818.
16. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol. 2007 Jun; 34(6):1426-31.
17. Elhan AH, Oztuna D, Kutlay S, Küçükdeveci AA, Tennant A. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. BMC Musculoskelet Disord. 2008 Dec 18;9:166.
18. Hung M, Hon SD, Cheng C, Franklin JD, Aoki SK, Anderson MB, Kapron AL, Peters CL, Pelt CE. Psychometric Evaluation of the Lower Extremity Computerized Adaptive Test, the Modified Harris Hip Score, and the Hip Outcome Score. Orthop J Sports Med. 2014 Dec 19;2(12):2325967114562191.
19. Kopec JA, Badii M, McKenna M, Lima VD, Sayre EC, Dvorak M. Computerized adaptive testing in back pain: validation of the CAT-5D-QOL. Spine (Phila Pa 1976). 2008 May 20;33(12):1384-90.
20. Banerjee S, Plummer O, Abboud JA, Deirmengian GK, Levicoff EA, Courtney PM. Accuracy and Validity of Computer Adaptive Testing for Outcome Assessment in Patients Undergoing Total Hip Arthroplasty. J Arthroplasty. 2020 Mar; 35(3):756-61.
21. Kane LT, Abboud JA, Plummer OR, Beredjiklian PT. Improving Efficiency of Patient-Reported Outcome Collection: Application of Computerized Adaptive Testing to DASH and QuickDASH Outcome Scores. J Hand Surg Am. 2021 Apr;46(4):278-86.
22. Kane LT, Namdari S, Plummer OR, Beredjiklian P, Vaccaro A, Abboud JA. Use of Computerized Adaptive Testing to Develop More Concise Patient-Reported Outcome Measures. JB JS Open Access. 2020 Mar 12;5(1):e0052.
23. O'Neil JT, Plummer OR, Raikin SM. Application of Computerized Adaptive Testing to the Foot and Ankle Ability Measure. Foot Ankle Int. 2021 Jan;42(1):2-7.
24. Plummer OR, Abboud JA, Bell JE, Murthi AM, Romeo AA, Singh P, Zmistowski BM. A concise shoulder outcome measure: application of computerized adaptive testing to the American Shoulder and Elbow Surgeons Shoulder Assessment. J Shoulder Elbow Surg. 2019 Jul;28(7):1273-80.
25. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. Spine J. 2008 Nov-Dec;8(6):968-74.

26. Hung M, Saltzman CL, Kendall R, Bounsanga J, Voss MW, Lawrence B, Spiker R, Brodke D. What Are the MCIDs for PROMIS, NDI, and ODI Instruments Among Patients With Spinal Conditions? Clin Orthop Relat Res. 2018 Oct;476(10):2027-36.

27. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976). 2007 Dec 15;32(26):3047-51.

28. Carreon LY, Glassman SD, Campbell MJ, Anderson PA. Neck Disability Index, Short Form-36 Physical Component Summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. Spine J. 2010 Jun;10(6):469-74.

29. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. Spine (Phila Pa 1976). 2006 Mar 1;31(5):598-602.

30. Hägg O, Fritzell P, Nordwall A; Swedish Lumbar Spine Study Group. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J. 2003 Feb;12(1):12-20.

31. Johnsen LG, Hellum C, Nygaard OP, Storheim K, Brox JI, Rossvoll I, Leivseth G, Grotle M. Comparison of the SF6D, the EQ5D, and the Oswestry Disability Index in patients with chronic low back pain and degenerative disc disease. BMC Musculoskelet Disord. 2013 Apr 26;14:148.

32. Jorritsma W, Dijkstra PU, de Vries GE, Geertzen JH, Reneman MF. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J. 2012 Dec;21(12):2550-7.

33. Parker SL, Adogwa O, Mendenhall SK, Shau DN, Anderson WN, Cheng JS, Devin CJ, McGirt MJ. Determination of minimum clinically important difference (MCID) in pain, disability, and quality of life after revision fusion for symptomatic pseudoarthrosis. Spine J. 2012 Dec;12(12):1122-8.

34. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. Spine J. 2009 Oct;9(10):802-8.

35. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10.

36. Brodke DS, Goz V, Voss MW, Lawrence BD, Spiker WR, Hung M. PROMIS PF CAT Outperforms the ODI and SF-36 Physical Function Domain in Spine Patients. Spine (Phila Pa 1976). 2017 Jun 15;42(12):921-9.

37. Hart DL, Mioduski JE, Stratford PW. Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. J Clin Epidemiol. 2005 Jun;58(6):629-38.

38. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. Qual Life Res. 2008 Oct;17(8):1081-91.

39. Hung M, Baumhauer JF, Latt LD, Saltzman CL, SooHoo NF, Hunt KJ; National Orthopaedic Foot & Ankle Outcomes Research Network. Validation of PROMIS ® Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. Clin Orthop Relat Res. 2013 Nov;471(11):3466-74.

40. Hung M, Saltzman CL, Voss MW, Bounsanga J, Kendall R, Spiker R, Lawrence B, Brodke D. Responsiveness of the Patient-Reported Outcomes Measurement Information System (PROMIS), Neck Disability Index (NDI) and Oswestry Disability Index (ODI) instruments in patients with spinal disorders. Spine J. 2019 Jan;19(1):34-40.

41. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155-63.

42. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014 Oct 1;312(13):1342-3.