Choosing the best match of the journal's keywords to your work will likely result in better match to reviewers with expertise in your interest area. It may also help to speed the review process.:	
Optional Terms: You have the option to add additional key words that are not on the journal's list to more specifically categorize your submission.:	Cochin Hand Function Scale

Using Optimal Test Assembly Methods for Shortening Patient-Reported Outcome Measures: Development and Validation of the Cochin Hand Function Scale-6 – A Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Alexander W. Levis, BSc[1,2]; Daphna Harel, PhD[3,4]; Linda Kwakkenbos, PhD[1,5,6]; Marie-Eve Carrier, MSc[1]; Luc Mouthon, MD, PhD[7,8]; Serge Poiraudeau, MD, PhD [7,9,10]; Susan J. Bartlett, PhD[11,12]; Dinesh Khanna, MD, MS [13]; Vanessa L. Malcarne, PhD[14,15]; Maureen Sauve, BA[16,17]; Cornelia H.M. van den Ende, PhD[18]; Janet L. Poole, PhD, OTR/L[19]; Anne A. Schouffoer, MD, PhD[20]; Joep Welling[21]; Brett D. Thombs, PhD[1,2,5,12,22-24]; SPIN Investigators[25]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; [2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Québec, Canada; [3]PRIISM Applied Statistics Center, New York University, New York, New York, USA; [4]Department of Humanities and Social Sciences in the Professions, New York University, New York, New York, USA [5]Department of Psychiatry, McGill University, Montréal, Québec, Canada; [6]Behavioural Science Institute, Clinical Psychology, Radboud University, Nijmegen, the Netherlands; [7]Université Paris Descartes, Assistance Publique-Hôpitaux de Paris, Paris, France; [8]Service de Médecine Interne, Hôpital Cochin, Paris, France; [9]Service de Médecine Physique et Réadaptation, Hôpital Cochin, Paris, France; [10]IFR Handicap INSERM, Paris, France; [11]McGill University Health Center, Montréal, Québec, Canada; [12]Department of Medicine, McGill University, Montréal, Québec, Canada; [13]University of Michigan Scleroderma Program, Ann Arbor, MI; [14]Department of Psychology, San Diego State University, San Diego, California, USA; [15]San Diego State

University/University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California, USA; [16]Scleroderma Society of Ontario, Hamilton, Ontario, Canada; [17]Scleroderma Society of Canada, Hamilton, Ontario, Canada; [18]Department of Rheumatology, Sint Maartenskliniek, The Netherlands; [19]Occupational Therapy Graduate Program, University of New Mexico, USA; [20]Leiden University Medical Center, Leiden, The Netherlands; [21]NVLE Dutch patient organization for systemic autoimmune diseases, Utrecht, The Netherlands; Departments of [22]Educational and Counselling Psychology; and [23]Psychology; and [24]School of Nursing, McGill University, Montreal, Quebec, Canada; [25]SPIN Investigators: Murray Baron, McGill University, Montréal, Québec, Canada; Dan Furst, University of California, Los Angeles, California, USA; Karen Gottesman, Scleroderma Foundation, Los Angeles, California, USA; Frank van den Hoogen, Radboud University Medical Center and Sint Maartenskliniek, Nijmegen, The Netherlands; Maureen D. Mayes, University of Texas Health Science Center, Houston, Texas, USA; Warren R. Nielson, St. Joseph's Health Care, London, Ontario, Canada; Robert Riggs, Scleroderma Foundation, Danvers, Massachusetts, USA; Fredrick Wigley, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Shervin Assassi, University of Texas Health Science Center, Houston, Texas, USA; Isabelle Boutron, Université Paris Descartes, and Assistance Publique-Hôpitaux de Paris, Paris, France; Angela Costa Maia, University of Minho, Braga, Portugal; Ghassan El-Baalbaki, Université du Québec à Montréal, Montréal, Québec, Canada; Carolyn Ells, McGill University, Montréal, Québec, Canada; Kim Fligelstone, Scleroderma Society, London, UK; Catherine Fortune, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Tracy Frech, University of Utah, Salt Lake City, Utah, USA; Dominique Godard, Association des Sclérodermiques de France, Sorel-Moussel, France; Marie Hudson, Jewish General Hospital and McGill University, Montréal, Québec, Canada; Ann

Impens, Midwestern University, Downers Grove, Illinois, USA; Yeona Jang, McGill University, Montréal, Québec, Canada; Sindhu R. Johnson, Toronto Scleroderma Program, Mount Sinai Hospital, Toronto Western Hospital, and University of Toronto, Toronto, Ontario, Canada; Ann Tyrell Kennedy, Federation of European Scleroderma Associations, Dublin, Ireland; Maggie Larche, McMaster University, Hamilton, Ontario, Canada; Catarina Leite, University of Minho, Braga, Portugal; Carlo Marra, Memorial University, St. John's, Newfoundland, Canada; Karen Nielsen, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Janet Pope, University of Western Ontario, London, Ontario, Canada; Alexandra Portales, Asociación Española de Esclerodermia, Madrid, Spain; Tatiana Sofia Rodriguez Reyna, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; Russell J. Steele, McGill University, Montréal, Québec, Canada; Maria E. Suarez-Almazor, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; Durhane Wong-Rieger, Canadian Organization for Rare Disorders, Toronto, Ontario, Canada; Alexandra Albert, Université Laval, Québec, Québec, Canada; Guylaine Arsenault, Sherbrooke University, Sherbrooke, Québec, Canada; Lyne Bissonnette, Sherbrooke University, Sherbrooke, Québec, Canada; Gilles Boire, Sherbrooke University, Sherbrooke, Québec, Canada; Alessandra Bruns, Sherbrooke University, Sherbrooke, Québec, Canada; Patricia Carreira, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Lorinda Chung, Stanford University, Stanford, California, USA; Pierre Dagenais, Sherbrooke University, Sherbrooke, Québec, Canada; Christopher Denton, Royal Free London Hospital, London, UK; Robyn Domsic, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; James V. Dunne, St. Paul's Hospital and University of British Columbia, Vancouver, British Columbia, Canada; Paul Fortin, Université Laval, Québec, Québec, Canada; Anna Gill, Royal Free London Hospital, London, UK; Jessica Gordon, Hospital for Special

Surgery, New York City, New York, USA; Genevieve Gyger, Jewish General Hospital and McGill University, Montréal, Québec, Canada; Ariane L. Herrick, University of Manchester, Salford Royal NHS Foundation Trust, Manchester, UK; Monique Hinchcliff, Northwestern University, Chicago, Illinois, USA; Alena Ikic, Université Laval, Québec, Québec, Canada; Niall Jones, University of Alberta, Edmonton, Alberta, Canada; Artur Jose de B. Fernandes, Sherbrooke University, Sherbrooke, Québec, Canada; Suzanne Kafaja, University of California, Los Angeles, California, USA; Nader Khalidi, McMaster University, Hamilton, Ontario, Canada; Benjamin Korman, Northwestern University, Chicago, Illinois, USA; Patrick Liang, Sherbrooke University, Sherbrooke, Québec, Canada; Joanne Manning, Salford Royal NHS Foundation Trust, Salford, UK; Ariel Masetto, Sherbrooke University, Sherbrooke, Québec, Canada; David Robinson, University of Manitoba, Winnipeg, Manitoba, Canada; Sophie Roux, Sherbrooke University, Sherbrooke, Québec, Canada; Elena Schiopu, University of Michigan, Ann Arbor, Michigan, USA; Doug Smith, University of Ottawa, Ottawa, Ontario, Canada; Robert Spiera, Hospital for Special Surgery, New York, New York, USA; Virginia Steen, Georgetown University, Washington, DC, USA; Evelyn Sutton, Dalhousie University, Halifax, Nova Scotia, Canada; Carter Thorne, Southlake Regional Health Centre, Newmarket, Ontario, Canada; John Varga, Northwestern University, Chicago, Illinois, USA; Pearce Wilcox, St. Paul's Hospital and University of British Columbia, Vancouver, British Columbia, Canada; Vanessa C. Delisle, Jewish General Hospital and McGill University, Montréal, Québec, Canada; Claire Fedoruk, Jewish General Hospital, Montréal, Québec, Canada; Rina S. Fox, San Diego State University and University of California, San Diego, San Diego, California, USA; Shadi Gholizadeh, San Diego State University and University of California, San Diego, San Diego, California, USA; Lisa R. Jewett, Jewish General Hospital and McGill University, Montréal, Québec, Canada;

Brooke Levis, Jewish General Hospital and McGill University, Montréal, Québec, Canada; Katherine Milette, Jewish General Hospital and McGill University, Montréal, Québec, Canada; Sarah D. Mills, San Diego State University and University of California, San Diego, San Diego, California, USA; Mia R. Pepin, Jewish General Hospital, Montréal, Québec, Canada; Jennifer Persmann, Université du Québec à Montréal, Montréal, Québec, Canada.

Financial Support: SPIN is funded by a Canadian Institutes of Health Research (CIHR) Emerging Team Grant for Rare Diseases (PI, Thombs; TR3-119192). In addition to CIHR funding, SPIN has received institutional contributions from the Lady Davis Institute for Medical Research of the Jewish General Hospital, Montréal, Canada and from McGill University, Montréal, Canada. SPIN has also received support from the Scleroderma Society of Ontario, the Scleroderma Society of Canada, and Sclérodermie Québec. Mr. Levis was supported by a CIHR Institute of Musculoskeletal Health and Arthritis Studentship in Musculoskeletal Health and Arthritis. Dr. Harel was supported by NYU start-up research grants and SPIN. Dr. Kwakkenbos was supported by a CIHR Banting Postdoctoral Fellowship. Dr. Thombs was supported by an Investigator Salary Award from the Arthritis Society.

Word Count: 3,777

Address for Correspondence:

Daphna Harel, PhD; 246 Greene Street, 3rd floor, New York NY, 10003; Tel (212) 992-6701; E-mail: daphna.harel@nyu.edu

ABSTRACT

Objective: The objective was to develop and validate a short form of the Cochin Hand Function Scale (CHFS), which measures hand disability, for use in systemic sclerosis, using objective criteria and reproducible techniques.

Methods: Responses on the 18-item CHFS were obtained from English-speaking patients enrolled in the Scleroderma Patient-centered Intervention Network (SPIN) Cohort. CHFS unidimensionality was verified using confirmatory factor analysis, and an item response theory model was fit to CHFS items. Optimal test assembly (OTA) methods identified a maximally precise short form for each possible form length between 1 and 17 items. The final short form selected was the form with the least number of items that maintained statistically equivalent convergent validity, compared to the full-length CHFS, with the Health Assessment Questionnaire - Disability Index (HAQ-DI) and Physical Function domain of PROMIS-29.

Results: There were 601 patients included. A 6-item short form of the CHFS (CHFS-6) was selected. The CHFS-6 had a Cronbach's alpha of 0.93. Correlations of the CHFS-6 summed score with HAQ-DI (r = 0.79) and PROMIS-29 Physical Function (r = -0.54) were statistically equivalent to the CHFS (r = 0.81, r = -0.56). The correlation with the full CHFS was high (r = 0.98).

Conclusion: The OTA procedure generated a valid short form of the CHFS with minimal loss of information compared to the full-length form. The OTA method used was based on objective, pre-specified criteria, but should be further studied for viability as a general procedure for shortening patient reported outcome measures in health research.

Significance and Innovation:

- Valid and reliable short forms of patient-reported outcome (PRO) measures can reduce patient burden and increase research capacity.
- There are no standard methods for developing short forms to maximize information.
- This study showed that optimal test assembly (OTA) and equivalence testing methods may be used to create short forms objectively and reproducibly.
- The Cochin Hand Function Scale-6 was developed using OTA and equivalency methods and is a brief, valid short form for measuring hand disability in systemic sclerosis.

Patient-reported outcomes (PROs) assess patient health, well-being, and treatment response based on patient perspectives (1,2). In rheumatic diseases, PROs, such as quality-of-life and functional ability, are as important to many patients as survival (3). Inclusion of PROs has become central in many clinical trials and cohort-based observational studies (4). Thus, efficient measurement of PROs is essential to limit both cost of patient cohorts and burden to patients who may be asked to respond to many different scales.

In rare diseases, including systemic sclerosis (SSc), cohorts designed to collect medical and PRO data from large numbers of patients require collaborations that span countries, languages, and clinical settings. SSc is a chronic, multisystem autoimmune disorder characterized by fibrotic changes to the skin, joints and tendons, as well as vascular injury (5,6). Large SSc multinational cohorts include the European League Against Rheumatism Scleroderma Trials and Research cohort (7) and the Scleroderma Patient-centered Intervention Network (SPIN) Cohort (8), among others (9).

Digital ulcers, contractures and deformities of the hand, which lead to decreased flexion, limited extension and reduced thumb abduction, play a major role in functional disability among patients with SSc (10-12). Disability related to impaired hand function may be present in close to 90% of SSc patients (12,13). The Cochin Hand Function Scale (CHFS) (14) was developed to measure functional ability of the hand among patients with rheumatic diseases and has been validated (15,16) and used extensively in SSc (8,17-19). The CHFS consists of 18 items that, when patient- or clinician-completed, assess the ability to perform daily hand-related activities. There are, however, notable redundancies amongst the 18 items. For example, item 13 (*Can you write a short sentence with a pencil or an ordinary pen*?) and item 14 (*Can you write a letter* *with a pencil or an ordinary pen?*) may not together provide significantly more information beyond what is captured by either item independently.

Shortening PRO measures would increase the number of outcomes that can be measured in studies; however, no standard methods currently exist (20-23). Traditionally, items have been deleted based on item-total correlations or the goal of maintaining or improving factorial structure by removing items with low factor loadings or high residuals (23,24) or through qualitative analysis of item content (23). Modern techniques, such as item response theory (25), allow for detailed item evaluation to identify problematic items, but still leave the final selection of items to the researcher's prerogative rather than objective and reproducible criteria.

Optimal test assembly (OTA) (26), used frequently for item selection in designing high-stakes educational tests (27), incorporates the results of item response theory models to select the optimal subset of an item pool that best satisfies objective, pre-specified constraints, such as content- or precision-related requirements. To the best of our knowledge, OTA has not been used previously in health research. Nonetheless, these methods have the potential to empirically guide the shortening of previously validated PRO measures by optimizing performance based on objective, replicable procedures.

The objective of the present study was to develop a short form of the CHFS using OTA. To do this, we: (1) verified the unidimensionality of the scale using confirmatory factor analysis; (2) applied OTA methods in order to obtain maximally precise candidate short forms of each possible length; and (3) selected the shortest possible short form that demonstrated statistical equivalency, based on tests of convergent validity, to the full-length scale.

PATIENTS AND METHODS

Sample and Procedure

The sample consisted of patients enrolled in the Scleroderma Patient-centered Intervention Network (SPIN) Cohort (8) from 21 centers in Canada, the United States, and the United Kingdom who completed study questionnaires from March 2014 through June 2015. To be included in the SPIN Cohort, patients must have a confirmed diagnosis of SSc according to 2013 ACR/EULAR classification criteria (28), be ≥ 18 years of age, have the ability to give informed consent, be fluent in English or French, and be able to respond to questionnaires via the Internet. Eligible patients are invited by SPIN center physicians or supervised nurse coordinators to participate, and written informed consent is obtained. To initiate patient registration, the local SPIN physician or nurse coordinator completes an online medical data record and an automated email is then sent to the patient with instructions for activating their account. Participants complete SPIN Cohort measures online upon enrollment and subsequently once every 3 months. Only patients with complete CHFS data at baseline in English were included in the present study.

Measures

SPIN physicians provided medical information, including time since onset of the first non-Raynaud's phenomenon symptom, SSc subtype (limited or diffuse) (29), modified Rodnan skin score, and presence of puffy fingers, sclerodactyly, skin thickening of the fingers, fingertip pitting scars, digital ulcers, and small joint contractures (30). Patients provided demographic data and completed PROs.

The 18-item CHFS (14) measures the ability to perform daily hand-related activities. Items reflecting five content areas (kitchen, dressing oneself, hygiene, the office, and other) are scored on a Likert scale from 0 (*yes, without difficulty*) to 5 (*impossible*). Total CHFS scores range from 0 to 90, and higher scores indicate more hand disability. The CHFS, when clinician-completed, has shown excellent intra- and interrater reliability (intraclass correlation coefficients of 0.97 and 0.96, respectively) (14), good convergent validity with functional disability measures, and sensitivity to changes in hand function (14,31). Validity and reliability of the self-report version have been confirmed in SSc (15,16).

The Health Assessment Questionnaire Disability Index (HAQ-DI) and the Physical Function domain of the 29-item Patient Reported Outcomes Measurement Information System (PROMIS-29; profile version 1.0 or 2.0) were used to establish convergent validity. The HAQ-DI assesses disability within 8 categories measured over the past 7 days: dressing/grooming, arising, eating, walking, hygiene, reach, grip, and common daily activities. Each item is rated on a 4-point scale, ranging from 0 (*without any difficulty*) to 3 (*unable to do*). The highest score from each category, indicating greater disability, determines the score for that category, and the total score is the mean of the 8 category scores, ranging from 0 (*no disability*) to 3 (*severe disability*). The HAQ-DI is widely used in patients with rheumatologic diseases, and is a valid measure of functional disability in SSc (32).

The Physical Function domain of the PROMIS-29 assesses functional ability. This domain consists of four items measuring capacity to complete day-to-day activities, scored on a Likert scale from 1 (*unable to do*) to 5 (*without any difficulty*). The summed score of the four items is standardized based on norms from the general United States population (mean = 50; standard deviation = 10). Higher scores indicate greater physical function. The PROMIS-29 and its subscales have been shown to be valid measures of health status in SSc (33).

Statistical Analysis

To verify the assumption of unidimensionality for the CHFS, a single-factor confirmatory factor analysis model was fit to the CHFS data using a robust weighted least squares estimator (34). Modification indices were calculated to recognize item pairs for which measurement errors correlated highly. If there was theoretical justification for shared effects within these pairs of items, we allowed their errors to covary if this improved model fit (35). Model fit was evaluated via a mean- and variance-adjusted chi-square test statistic (34), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). The CFI, TLI and RMSEA indices were prioritized, as the chi-square test may reject models despite good fit because it is highly dependent on sample size (36). Values of CFI and TLI ≥ 0.95 and RMSEA ≤ 0.10 were considered to indicate good fit (37,38).

Next, a generalized partial credit item response theory model (25,39) was fit to all 18 CHFS items. For each item, the model estimates (1) thresholds for the levels of disability in hand function at which patients are more likely to endorse a given response category instead of the category below, and (2) a discrimination parameter, which measures the strength of the relationship between that item and the underlying construct, functional disability of the hand (θ). Item information functions were then estimated. The test information function (TIF), calculated as the sum of the item information functions, measures the total amount of Fisher's information contained in the CHFS as a function of the latent trait, θ. Fisher's information, which is inversely related to standard error of measurement, summarizes the degree of precision of measurements of the latent construct (25).

OTA was then used to create candidate short forms of each possible length by selecting items to maximize the TIF (26). Thus, for each possible short form length, a single optimal short form, which included a subset of the 18 total CHFS items, was generated. OTA methods use mixed integer programming to optimize an objective function subject to a series of user-defined constraints. In this case, we adopted a maximin procedure (26,40) to select items that maximized the height of each short form's TIF, maintaining the same relative shape of the full form's TIF.

This approach yields short forms that measure the latent trait with the same relative precision across the continuum of the latent trait as the full form, while minimizing absolute loss of information. Based on previously established guidelines for best performance of this OTA procedure, the relative shape of the TIF was anchored at five points across the spectrum of disability in hand function (θ = -1,1,2,3,5) (26). For each candidate short form, the height of the TIF and percentage of the full-form TIF maintained in the short form were calculated, as well as the average total information across the latent trait spectrum, as a percentage of full-form total information.

For each candidate short form and the full-length form, patients were scored in two ways. First, summed scores were computed since summed scores are typically relied upon for clinical use. Second, factor scores of disability in hand function, which are assumed to have a standard normal distribution in the population, were estimated for each patient via an application of Bayes' theorem. Because of the well-known limitations of the summed score under the generalized partial credit model, factor scores were considered to provide a more consistent estimate of the latent trait (41,42).

For each candidate short form and the full-length CHFS, Cronbach's α was used to assess the internal consistency reliability. For concurrent validity, Pearson's product-moment correlation coefficients between summed scores and factor scores on each candidate form with summed and factor scores for the full-length form were calculated. Convergent validity was assessed via correlations between summed and factor scores on each candidate form and summed scores on the HAQ-DI and the PROMIS-29 Physical Function domain. Patients with missing data for either of the convergent validity measures were not included in the corresponding correlation calculations. We hypothesized that higher scores on the CHFS and its short forms would be associated with higher scores on the HAQ-DI (more disability) but lower scores for the PROMIS-29 Physical Function domain (greater physical function).

OTA methods generate optimal candidate short forms of each possible length, but do not readily provide criteria by which to select the best short-form length. By nature of eliminating items, the short forms necessarily will have lower information as compared with the full CHFS. Beforehand, there is no obvious threshold at which one would conclude that a short form has adequate information, so the validity of the short forms must be assessed in order to find a balance between shortening the scale and retaining its measurement ability. Thus, two criteria were used for selecting which candidate short form should be chosen. First, we required that the selected short form maintain high concurrent validity ($r > 0.90$) and high internal consistency ($\alpha \geq 0.90$) and demonstrate statistical equivalence with the full CHFS for measures of convergent validity. Equivalence testing, which has origins in clinical trials, is used to test whether the difference between two effect measures (e.g., treatment effect for two drugs) is within a pre-specified range (43). The equivalence testing paradigm, contrary to traditional hypothesis testing, tests a null hypothesis that there will be a difference between the two effect measures equal or greater than a pre-specified threshold against the alternative of equivalence or no difference. In our study, we specified a null hypothesis that the magnitude of the difference between each convergent validity correlation for the candidate short form and its corresponding correlation for the full CHFS would be ≥ 0.05 (44). To assess statistical significance, we applied a Bonferonni correction factor for each of 66 possible comparisons (summed score and factor score x 2 measures x 16 short forms of length 2-17 items plus single-item score for 2 measures; $p < 7.58\times10^{-4}$) to maintain the family-wise Type I error rate of $\alpha = 0.05$.

The confirmatory factor analysis was done using Mplus 7 (34). All other analyses were done using R version 3.2.1 (45). The generalized partial credit model was fit using the **ltm** package (46). The OTA analysis was conducted using the **lpSolveAPI** package (47).

RESULTS

There were 601 patients who completed the CHFS. Of these, 596 (99%) also completed the HAQ-DI, and 595 (99%) completed the PROMIS-29 Physical Function domain. The mean age was 55.4 years, 87% were women, and 42% had diffuse SSc. The mean score on the CHFS was 14.4 (SD = 16.7). CHFS scores in patients with diffuse SSc were substantially higher than patients with limited SSc (see Table 1 for descriptive statistics).

Confirmatory Factor Analysis

A unidimensional confirmatory factor analysis model of the CHFS items, where covariance of item residuals was restricted to zero, resulted in less than ideal fit (χ^2[df = 135] = 1509.3, p < 0.0001, TLI = 0.966, CFI = 0.970, RMSEA = 0.130). Modification indices suggested that allowing residuals of items 9 and 10 and items 13 and 14 to covary would improve model fit. Items 9 and 10 both assess the ability to perform actions involved in dressing oneself, and the content of items 13 and 14 involves writing with a pencil. The model was refitted, allowing the residuals for these two item pairs to covary, and fit improved (χ^2[df = 133] = 866.0, p < 0.0001, TLI = 0.982, CFI = 0.984, RMSEA = 0.096). Factor loadings for the items were all very high (> 0.82) with the majority over 0.90.

Item Response Theory Model and Optimal Test Assembly

The generalized partial credit model was fit to the 18 items of the CHFS. Item content along with the discrimination parameters of the model are shown in Table 2. The items with greatest discriminative ability and, thus, the greatest influence on the TIF, were items 1 (*Can you hold a bowl?*), 3 (*Can you hold a plate full of food?*), 7 (*Can you prick things well with a fork?*), 16 (*Can you cut a piece of paper with scissors?*) and 18 (*Can you turn a key in a lock?*). Figure 1 shows the individual item information functions generated by the generalized partial credit model and the aggregate TIF.

Considering each possible subset of the 18 items, the OTA procedure selected the short form of each size that maximally maintained the shape of the TIF for the full-length form. The items chosen for each of the short form sizes are shown in Table 3. Several patterns emerged from the OTA selection procedure. First, items 12, 13 and 14 were only selected in the longest short forms and, thus, quickly dropped from smaller short forms. These items all had low discrimination parameters (see Table 2). Second, items 1, 3, 7 and 9 were included in all forms of at least 6 items. All of these items had very high information at certain points on the continuum of disability or had fairly high information consistently across the continuum. Third, some items seemed to alternate in their selection into short forms. For example, items 15 and 18 were often mutually exclusive in the smaller short form sizes. The content of both items relates to opening a lock or door, and there was little added measurement value in including both.

The TIFs for each candidate short form are summarized in Table 4. As expected, the height of the TIF and percentage of information at each of five points across the latent spectrum and average information across the entire spectrum show a consistent decrease in information as the length of the form decreases. This drop in information translates into an increase in the standard error of measurement for the latent trait as the length of the short form decreases. However, despite this loss of information, all internal and external validity correlations remained consistently high, even for short forms containing a small number of items (see Appendix Tables 1 and 2).

Selection of final short form

The equivalency analysis presented in Table 5 assesses which candidate short forms maintained a reasonably equivalent level of concurrent and convergent validity. The 6-item short form and all short forms with at least 9 items demonstrated statistically significant equivalence, following Bonferroni correction, to all correlations between summed and factor scores of the full CHFS with the HAQ-DI and PROMIS-29 Physical Function scores. Although the 5-item short form satisfied statistical equivalence for three of the correlations, the correlation between the factor scores and the PROMIS-29 Physical Function scores failed to demonstrate equivalence to the corresponding full CHFS correlation ($p = 0.003 > 7.58\times10^{-4}$). Thus, the 6-item optimal short form (CHFS-6; see Appendix 1) was the shortest candidate form to fulfill our equivalence requirement.

The CHFS-6 had a Cronbach's α of 0.932 and a correlation with the full 18-item CHFS scores for summed scores of r = 0.980 (95% confidence interval [CI] 0.976 to 0.983) and factor scores of r = 0.970 (95% CI 0.965 to 0.975) (see Appendix Tables 1 and 2). The summed scores on the CHFS-6 maintained strong positive correlations with the HAQ-DI (r = 0.790, 95% CI 0.758 to 0.819) and moderate negative correlations with the PROMIS-29 Physical Function domain (r = -0.544, 95% CI -0.599 to -0.485). The TIF of the CHFS-6 as compared to the full length form is shown in Figure 1. The CHFS-6 retained, on average, 38.6% of the Fisher information of the full form (see Table 4), corresponding to 1.61 times the standard error of measurement on average between the short and full forms.

DISCUSSION

This study demonstrated how OTA methods, which have been used extensively in high-stakes educational testing, may also be used to create valid PRO short forms based on objective, pre-specified constraints in health research. The main finding of the study was that the 18-item

CHFS could be shortened to a 6-item version (see Appendix 1) with minimal loss of information and maintaining high internal validity and similar convergent validity with the HAQ-DI and PROMIS-29 Physical Function domain. The summed scores of the CHFS-6 and the full CHFS correlated at $r = 0.980$. Cronbach's α was 0.932 compared to 0.974 for the full CHFS, and all correlations with convergent validity measures were similar and within pre-specified ranges for statistical equivalence to the correlations between these measures and the full CHFS. The items of the CHFS-6 included three related to eating (holding a bowl; holding a plate of food; holding a fork), one related to food preparation (peeling fruit), one related to dressing (buttoning shirts), and one related to the ability to use a key to unlock doors.

Compared to patients with greater hand disability, the CHFS-6 had relatively higher standard error among patients with minimal disability ($\theta < -1$). This occurred for two reasons. First, the short-forming procedure prioritized maximizing information where the original form measures hand disability well (θ between -1 and 5). Second, the minimal estimated factor score for patients in our sample was $\theta = -1.26$, resulting in little data from patients in the lower end.

The exact specification of the OTA procedure that we used resulted in candidate short forms that were not required to be nested, meaning that the items of one short form were not required to be contained in all larger short forms. For example, item 8 is in the 2-item short form, but does not appear in the 3-item or 4-item forms. This reflects the shape-maintaining property of our OTA specification that allows for a push-and-pull dynamic between items that have more information at different locations of the latent trait continuum. However, if the creation of nested short forms is desired, the methodology used in this study could be easily adapted to satisfy this constraint by selecting each subsequent optimal candidate short form only from the items appearing on the optimal short form one item longer. Similarly, OTA does not automatically consider content validity, but these constraints such as setting a maximum number of items per content area may be added as desired. However, if a short form indeed performs virtually equivalently to its parent form, despite the lack of content in some sub-areas due to elimination of items, this suggests that the content may not be necessary for optimal measurement, regardless of theoretical suggestions otherwise.

Although this study focused on the development and validation of a short form of the CHFS for patients with SSc, the OTA approach that we used could be applied with other patient populations and other measures for developing short forms of PRO measures. The present study, however, represents only a first step in using OTA methods to attempt to standardize processes for developing optimally functioning shortened PRO measures in health research. The SPIN Cohort is a convenience sample, and therefore may not be representative of the SSc population. Patients in the present study, for instance, had somewhat lower hand disability, on average, compared to other SSc cohorts where the CHFS has been used (15,16).

Further research is needed to determine the robustness of the OTA procedure with other measures and patient cohorts and to compare to other short-forming methods. The Bonferroni correction used to account for issues of multiple testing may be overly conservative, albeit easily applied, and alternative approaches may be preferred. Furthermore, additional research is needed on how best to use OTA to shorten scales in the context of multidimensionality. Finally, OTA is an exploratory, data-driven approach, and results of this study should be replicated.

In sum, this study demonstrated how OTA methods can be used to develop and validate short forms of PRO measures based on pre-specified and objective criteria for determining both the number of items to include in the short form and the specific items to be included. Application of OTA methods to the 18-item CHFS in a large sample of SSc patients resulted in a

6-item version with minimal loss of information and minimal change to indices of reliability and convergent validity compared to the 18-item form.

REFERENCES

1. Bren L. The importance of patient-reported outcomes...it's all about the patients. FDA Consum 2006;40:26-32.
2. Fairclough DL. Patient reported outcomes as endpoints in medical research. Stat Methods Med Res 2004;13:115-38.
3. Rumsfeld JS, Ho PM. Depression and cardiovascular disease: A call for recognition. Circulation 2005;111:250-3.
4. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45:S3-S11.
5. Mayes MD. Systemic sclerosis. In: Klippel, JH, Stone JH, Crofford LJ, White PH, editors. Primer on the rheumatic diseases. 13th edn. New York: Springer; 2008. p. 343-62.
6. Firestein GS, Kelley WN, Budd RC, Gabriel SE, McInnes IB, O'Dell JR, editors. Kelley's textbook of rheumatology. 9th edn. Philadelphia: Saunders Elsevier; 2012.
7. Meier FM, Frommer KW, Dinser R, Walker UA, Czirjak L, Denton CP, et al. Update on the profile of the EUSTAR cohort: an analysis of the EULAR Scleroderma Trials and Research group database. Ann Rheum Dis 2012;71:1355-60.
8. Kwakkenbos L, Jewett LR, Baron M, Bartlett SJ, Furst D, Gottesman K, et al. The Scleroderma Patient-centered Intervention Network (SPIN) Cohort: protocol for a cohort multiple randomised controlled trial (cmRCT) design to support trials of psychosocial and rehabilitation interventions in a rare disease context. BMJ Open 2013;3:e003563.

. Galluccio F, Walker UA, Nihtyanova S, Moinzadeh P, Hunzelmann N, Krieg T, et al. Registries in systemic sclerosis: a worldwide experience. Rheumatology (Oxford) 2011;50:60-8.

0. Entin MA, Wilkinson RD. Scleroderma hand: a reappraisal. Orthop Clin North Am 1973;4:1031-8.

1. Poole JL. Musculoskeletal rehabilitation in the person with scleroderma. Curr Opin Rheumatol 2010;22:205-12.

2. Bassel M, Hudson M, Taillefer SS, Schieir O, Baron M, Thombs BD. Frequency and impact of symptoms experienced by patients with systemic sclerosis: results from a Canadian National Survey. Rheumatology 2011;50:762-7.

3. Kallen MA, Mayes MD, Kriseman YL, de Achaval SB, Cox VL, Suarez-Almazor ME. The Symptom Burden Index: Development and initial findings from use with patients with systemic sclerosis. J Rheumatol 2010;37:1692-8.

4. Duruöz MT, Poiraudeau S, Fermanian J, Menkes CJ, Amor B, Dougados M, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 1996;23:1167-72.

5. Brower LM, Poole JL. Reliability and validity of the Duruoz Hand Index in persons with systemic sclerosis (scleroderma). Arthritis Rheum 2004;51:805–9.

6. Rannou F, Poiraudeau S, Berezné A, Baubet T, Le-Guern V, Cabane J, et al. Assessing disability and quality of life in systemic sclerosis: construct validities of the Cochin hand function scale, health assessment questionnaire (HAQ), systemic sclerosis HAQ, and medical outcomes study 36-item short form health survey. Arthritis Rheum 2007;57:94-102.

7. Mouthon L, Carpentier PH, Lok C, Clerson P, Gressin V, Hachulla E, et al. Ischemic digital ulcers affect hand disability and pain in systemic sclerosis. J Rheumatol 2014;41:1317-23.

8. Bongi SM, Del Rosso A, Galluccio F, Tai G, Sigismondi F, Passalacqua M, et al. Efficacy of a tailored rehabilitation program for systemic sclerosis. Clin Exp Rheumatol 2009;27:S44-S50.

9. Bongi SM, Del Rosso A, Galluccio F, Sigismondi F, Miniati I, Conforti ML, et al. Efficacy of connective tissue massage and Mc Mennell joint manipulation in the rehabilitative treatment of the hands in systemic sclerosis. Clin Rheumatol 2009;28:1167-73.

0. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess 2000;12:102-11.

1. Stanton JM, Sinar EF, Balzer WK, Smith PC. Issues and strategies for reducing the length of self-report scales. Pers Psychol 2002;55:167-94.

2. Kruyen PM, Emons WH, Sijtsma K. On the shortcomings of shortened tests: A literature review. International Journal of Testing 2013;13:223-48.

3. Goetz C, Coste J, Lemetayer F, Rat AC, Montel S, Recchia S, et al. Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. J Clin Epidemiol 2013;66:710-8.

4. Jewett LR, Hudson M, Haythornthwaite JA, Heinberg L, Wigley FM, Baron M, et al. Development and validation of the brief-satisfaction with appearance scale for systemic sclerosis. Arthritis Care Res 2010;62:1779-86.

5. van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer; 1997.

6. van der Linden WJ. Linear models for optimal test design. New York: Springer; 2006.

7. Kuhn JT, Kiefer T. Optimal test assembly in practice: The design of the Austrian educational standards assessment in mathematics. Zeitschrift für Psychologie 2013;221:190-200.

8. van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis Rheum 2013;65:2737-47.

9. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger TA Jr, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202-5.

0. Clements P, Lachenbruch P, Siebold J, White B, Weiner S, Martin R, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol 1995;22:1281-5.

1. Poiraudeau S, Chevalier X, Conrozier T, Flippo RM, Lioté F, Noël E, et al. Reliability, validity, and sensitivity to change of the Cochin hand functional disability scale in hand osteoarthritis. Osteoarthritis Cartilage 2001;9:570-7.

2. Poole JL, Steen VD. The use of the Health Assessment Questionnaire (HAQ) to determine physical disability in systemic sclerosis. Arthritis Rheum 1991;4:27-31.

3. Hinchcliff M, Beaumont JL, Thavarajah K, Varga J, Chung A, Podlusky S, et al. Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item health profile and Functional Assessment of Chronic Illness Therapy–Dyspnea Short Form. Arthritis Care Res 2011;63:1620-8.

4. Muthen LK, Muthen BO. Mplus User's Guide: Statistical Analysis with Latent Variables. 7th edn. Los Angeles: Muthen & Muthen; 2012.

5. McDonald RP, Ho MH. Principles and practice in reporting structural equation analyses. Psychol Methods 2002;7:64-82.

6. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114:552-66.

7. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling 1999;6:1-55.

8. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff Points in RMSEA test statistic in structural equation models. Sociol Methods Res 2008;36:462-94.

9. Muraki E. A generalized partial credit model: Application of an EM algorithm. Appl Psych Meas 1992;16:159-76.

0. van der Linden WJ, Boekkooi-Timminga E. A maximin model for IRT-based test design with practical constraints. Psychometrika 1989;54:237-47.

1. van der Ark LA. Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. Psychometrika 2005;70:283-304.

2. Harel D. The effect of model misspecification for polytomous logistic adjacent category item response theory models. Ph. D. thesis, McGill University; 2014.

3. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med 2011;26:192-6.

4. Counsell A, Cribbie RA. Equivalence tests for comparing correlation and regression coefficients. Br J Math Stat Psychol 2015;68:292-309.

5. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2014. URL http://www.R-project.org/.
6. Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. J Stat Softw 2006;17:1-25.
7. Diao Q, van der Linden WJ. Automated test assembly using lp_solve version 5.5 in R. Appl Psych Meas 2011;35:398-409.

Table 1. Patient demographic and disease characteristics (N = 601)

Sociodemographic variables	Values
Age, years; mean ± SD (range)	55.4 ± 11.9 (18.6 – 84.7)
Women; n (%)	524 (87)
Education >12 years; n (%)	483 (80)
Currently employed; n (%)	248 (41)
Married / cohabitating; n (%)	442 (74)
Time since the onset of the first non-Raynaud's symptoms, years[a]; mean ± SD (range)	11.8 ± 8.8 (0.1 – 46.2)
Patients with diffuse SSc[b]; n (%)	250 (42)
MRSS[c]; mean ± SD (range)	8.2 ± 9.2 (0 – 48)
Puffy fingers[d]; n (%)	371 (65)
Sclerodactyly[e]; n (%)	503 (84)
Skin thickening of the fingers[f]; n (%)	338 (56)
Fingertip pitting scars[g]; n (%)	250 (42)
Digital ulcers[h,i]; n (%)	229 (40)
Moderate to severe small joint contractures[j]; n (%)	145 (25)
CHFS score; mean [median] ± SD (range)	14.4 [8] ± 16.7 (0 – 88)
Mean score [median] ± SD (range) in diffuse SSc subset	20.5 [15] ± 19.2 (0 – 88)
Mean score [median] ± SD (range) in limited SSc subset	10.0 [4] ± 12.9 (0 – 62)
HAQ-DI score[k]; mean [median] ± SD (range)	0.81 [0.75] ± 0.69 (0 – 3)
PROMIS-29 Physical Function score[l]; mean [median] ± SD (range)	42.5 [41.8] ± 8.7 (22.9 – 56.9)

Abbreviations: n = number; SD = standard deviation; CHFS = Cochin Hand Function Scale; MRSS = modified Rodnan skin score; SSc = systemic sclerosis; HAQ-DI = Health Assessment Questionnaire disability index; PROMIS-29 = 29-item Patient Reported Outcomes Measurement Information System.

[i]Considered to have digitial ulcer if had digital pulp (volar), distal to distal interphalangeal joints, or elsewhere on the finger, and provided a response to both of these items.

Due to missing values: [a]N = 555; [b]N = 597; [c]N = 503; [d]N = 571; [e]N = 598; [f]N = 600; [g]N =591; [h]N = 579; [j]N = 574; [k]N = 596; [l]N = 595

Table 2. CHFS items and discrimination parameters from the generalized partial credit model

Item Number	Description	Discrimination Parameter
1	Can you hold a bowl?	2.33
2	Can you seize a full bottle and raise it?	1.84
3	Can you hold a plate full of food?	2.25
4	Can you pour liquid from a bottle into a glass?	2.10
5	Can you unscrew the lid from a jar opened before?	1.27
6	Can you cut meat with a knife?	1.98
7	Can you prick things well with a fork?	2.65
8	Can you peel fruit?	1.92
9	Can you button your shirt?	1.76
10	Can you open and close a zipper?	1.99
11	Can you squeeze a new tube of toothpaste?	1.61
12	Can you hold a toothbrush efficiently?	1.66
13	Can you write a short sentence with a pencil or an ordinary pen?	1.52
14	Can you write a letter with a pencil or an ordinary pen?	1.23
15	Can you turn a round door knob?	2.11
16	Can you cut a piece of paper with scissors?	2.29
17	Can you pick up coins from a table top?	1.41
18	Can you turn a key in a lock?	2.32

Table 3. Optimal short forms of each length

Short Form Length	1 Hold Bowl	2 Raise Bottle	3 Hold Plate	4 Pour Liquid	5 Unscrew Lid	6 Cut Meat	7 Prick Fork	8 Peel Fruit	9 Button Shirt	10 Zipper	11 Tooth-paste Tube	12 Hold Tooth-brush	13 Write Short	14 Write Letter	15 Door Knob	16 Cut Paper	17 Pick-up Coins	18 Turn Key
	Item Number (X marks inclusion)																	
1																		X
2							X	X										
3							X		X									X
4							X		X	X					X			
5	X						X	X	X									X
6	X		X				X	X	X									X
7	X		X				X		X						X	X	X	
8	X		X				X		X						X	X	X	X
9	X		X	X			X	X	X							X	X	X
10	X		X	X			X	X	X	X					X	X	X	
11	X		X	X			X	X	X	X					X	X	X	X
12	X	X	X	X		X	X	X	X	X					X		X	X
13	X	X	X	X	X		X	X	X	X					X	X	X	X
14	X	X	X	X	X	X	X	X	X	X	X					X	X	X
15	X	X	X	X	X	X	X	X	X	X	X				X	X	X	X
16	X	X	X	X	X	X	X	X	X	X	X		X		X	X	X	X
17	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X
18	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 4. Test information values for optimal short forms

	Test Information Function (% of full form)					
Short Form Length	**θ = -1**	**θ = 1**	**θ = 2**	**θ = 3**	**θ = 5**	**Average Information (% of full form)**
1	0.14 (4.6)	3.16 (8.1)	3.13 (6.8)	2.34 (5.4)	0.25 (4.5)	6.8
2	0.37 (12.4)	4.51 (11.6)	7.66 (16.7)	5.71 (13.0)	0.77 (13.6)	13.3
3	0.57 (19.1)	7.57 (19.5)	8.87 (19.3)	8.72 (19.9)	1.11 (19.7)	19.7
4	0.72 (24.2)	9.66 (24.9)	11.61 (25.3)	10.63 (24.3)	1.73 (30.8)	24.9
5	1.00 (33.4)	12.10 (31.2)	15.39 (33.5)	13.61 (31.1)	1.88 (33.4)	32.1
6	**1.08 (36.0)**	**14.43 (37.2)**	**17.92 (39.0)**	**16.82 (38.4)**	**2.69 (47.8)**	**38.6**
7	1.30 (43.6)	16.62 (42.8)	19.59 (42.7)	19.69 (44.9)	2.73 (48.5)	43.2
8	1.44 (48.1)	19.79 (51.0)	22.72 (49.5)	22.03 (50.3)	2.98 (53.0)	50.0
9	1.69 (56.6)	21.50 (55.4)	25.60 (55.8)	24.64 (56.2)	3.07 (54.6)	55.6
10	1.85 (61.7)	23.58 (60.7)	28.34 (61.8)	26.55 (60.6)	3.70 (65.8)	60.8
11	1.99 (66.3)	26.75 (68.9)	31.47 (68.6)	28.90 (66.0)	3.95 (70.3)	67.6
12	2.14 (71.7)	28.65 (73.8)	32.83 (71.5)	31.29 (71.4)	4.15 (73.8)	72.0
13	2.40 (80.3)	30.13 (77.6)	35.19 (76.7)	33.77 (77.1)	4.25 (75.7)	76.7
14	2.43 (81.2)	31.39 (80.8)	36.80 (80.2)	35.72 (81.5)	4.77 (84.9)	81.0
15	2.60 (86.9)	34.13 (87.9)	40.29 (87.8)	38.15 (87.1)	4.85 (86.2)	87.1
16	2.72 (90.9)	35.66 (91.8)	42.03 (91.6)	40.30 (92.0)	5.14 (91.5)	91.6
17	2.87 (96.0)	37.30 (96.1)	44.15 (96.2)	41.67 (95.1)	5.32 (94.7)	95.6
18	2.99 (100.0)	38.83 (100.0)	45.89 (100.0)	43.82 (100.0)	5.62 (100.0)	100.0

Bold values represent those of the final selected short form.

Table 5. Equivalency analysis results

	Difference in external validity correlations of summed and factor scores of full Cochin Hand Function Scale and candidate short versions [$r_{full} - r_{short}$], and p-values for equivalency within ± 0.05			
	HAQ-DI Score Correlation		**PROMIS-29 Physical Function Correlation**	
Short Form Length	*Summed Score*	*Factor Score*	*Summed Score*	*Factor Score*
1	0.093 (p > 0.99)	NA	-0.071 (p = 0.89)	NA
2	0.077 (p > 0.99)	0.080 (p > 0.99)	-0.060 (p = 0.78)	-0.069 (p = 0.90)
3	0.040 (p = 0.09)	0.045 (p = 0.28)	-0.038 (p = 0.13)	-0.045 (p = 0.34)
4	0.045 (p = 0.24)	0.051 (p = 0.53)	-0.043 (p = 0.25)	-0.047 (p = 0.40)
5	0.025 (p < 0.0001)	0.026 (p < 0.0001)	-0.025 (p < 0.001)	-0.026 (p < 0.01)
6	**0.016 (p < 0.0001)**	**0.015 (p < 0.0001)**	**-0.017 (p < 0.0001)**	**-0.016 (p < 0.0001)**
7	0.023 (p < 0.0001)	0.025 (p < 0.0001)	-0.035 (p < 0.01)	-0.041 (p = 0.12)
8	0.020 (p < 0.0001)	0.020 (p < 0.0001)	-0.031 (p < 0.001)	-0.036 (p = 0.02)
9	0.015 (p < 0.0001)	0.014 (p < 0.0001)	-0.019 (p < 0.0001)	-0.019 (p < 0.0001)
10	0.017 (p < 0.0001)	0.017 (p < 0.0001)	-0.021 (p < 0.0001)	-0.022 (p < 0.0001)
11	0.014 (p < 0.0001)	0.015 (p < 0.0001)	-0.019 (p < 0.0001)	-0.020 (p < 0.0001)
12	0.005 (p < 0.0001)	0.005 (p < 0.0001)	-0.008 (p < 0.0001)	-0.007 (p < 0.0001)
13	0.005 (p < 0.0001)	0.005 (p < 0.0001)	-0.008 (p < 0.0001)	-0.009 (p < 0.0001)
14	0.001 (p < 0.0001)	0.000 (p < 0.0001)	-0.005 (p < 0.0001)	-0.005 (p < 0.0001)
15	0.002 (p < 0.0001)	0.001 (p < 0.0001)	-0.008 (p < 0.0001)	-0.009 (p < 0.0001)
16	0.002 (p < 0.0001)	0.000 (p < 0.0001)	-0.006 (p < 0.0001)	-0.005 (p < 0.0001)
17	-0.001 (p < 0.0001)	-0.001 (p < 0.0001)	-0.001 (p < 0.0001)	-0.002 (p < 0.0001)

Bold values represent those of the final selected short form.

FIGURES

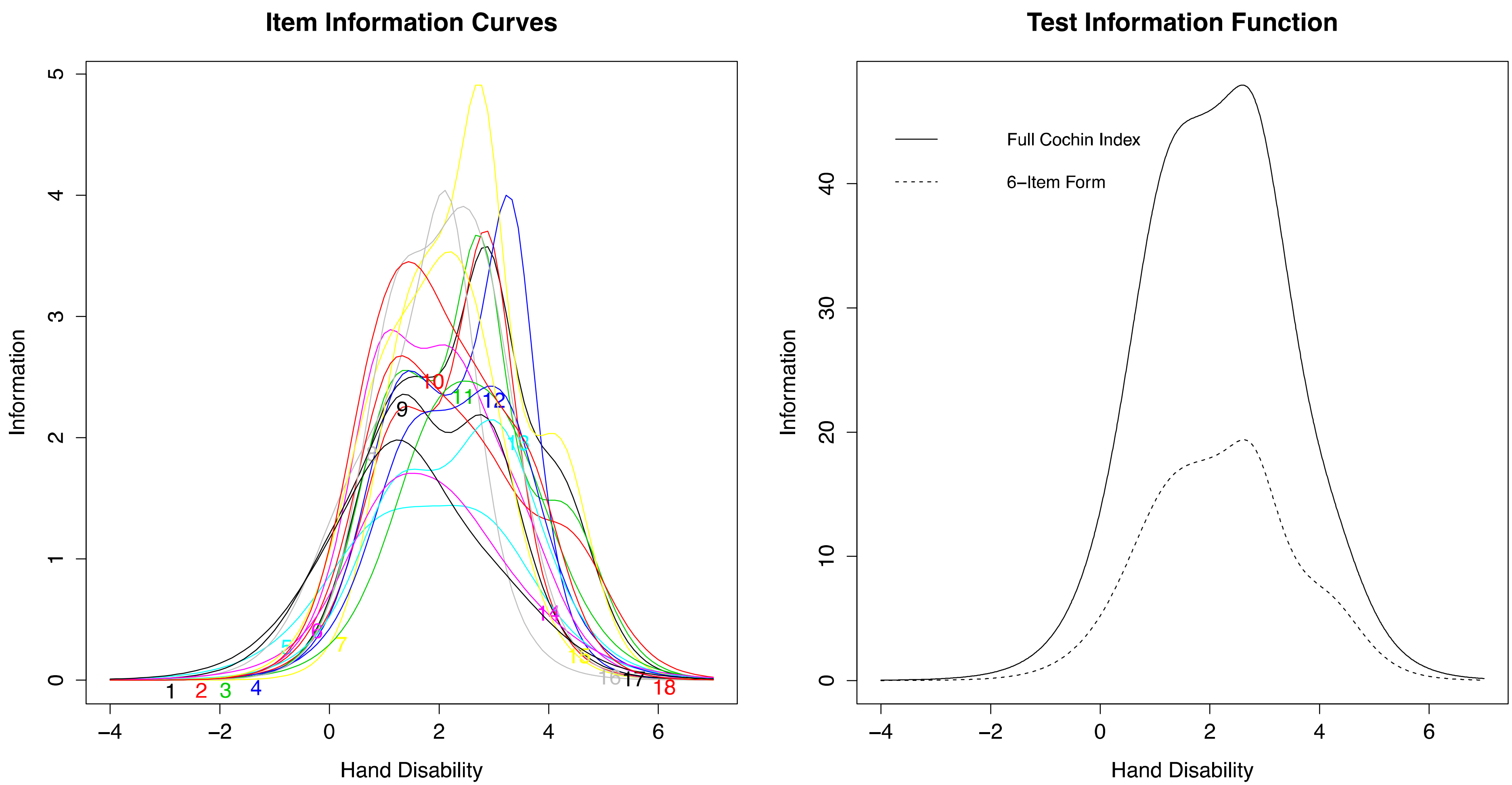

Figure 1. Item and test information curves of the CHFS. The left hand plot shows the 18 individual item information curves, labelled by color. The right hand plot compares the test information functions of the full CHFS (solid line) and CHFS-6 (dashed line)

APPENDIX 1: The Cochin Hand Function Scale 6-Item Short Form (CHFS-6)

Answer the following questions regarding your ability WITHOUT the help of any assistive devices, during the past month.

	Yes, without difficulty	Yes, with a little difficulty	Yes, with some difficulty	Yes, with much difficulty	Nearly impossible to do	Impossible
1. Can you hold a bowl?	☐	☐	☐	☐	☐	☐
2. Can you hold a plate full of food?	☐	☐	☐	☐	☐	☐
3. Can you prick things well with a fork?	☐	☐	☐	☐	☐	☐
4. Can you peel fruit?	☐	☐	☐	☐	☐	☐
5. Can you button your shirt?	☐	☐	☐	☐	☐	☐
6. Can you turn a key in a lock?	☐	☐	☐	☐	☐	☐

APPENDIX 2

Appendix Table 1. Classical test theoretic properties of optimal short forms

Short Form Length	Cronbach's α	Correlation of Summed Score with Full Form Score (95% CI)	Correlation of Summed Score with HAQ-DI Score (95% CI)	Correlation of Summed Score with PROMIS-29 Physical Function Score (95% CI)
1	NA	0.877 (0.857, 0.894)	0.714 (0.672, 0.751)	-0.490 (-0.549, -0.427)
2	0.796	0.925 (0.913, 0.936)	0.730 (0.690, 0.765)	-0.501 (-0.559, -0.439)
3	0.865	0.954 (0.946, 0.961)	0.766 (0.731, 0.798)	-0.523 (-0.579, -0.462)
4	0.912	0.956 (0.949, 0.963)	0.762 (0.726, 0.794)	-0.518 (-0.574, -0.457)
5	0.919	0.974 (0.970, 0.978)	0.782 (0.749, 0.811)	-0.537 (-0.592, -0.477)
6	**0.932**	**0.980 (0.976, 0.983)**	**0.790 (0.758, 0.819)**	**-0.544 (-0.599, -0.485)**
7	0.939	0.983 (0.980, 0.985)	0.783 (0.750, 0.812)	-0.526 (-0.582, -0.466)
8	0.949	0.985 (0.982, 0.987)	0.787 (0.755, 0.816)	-0.531 (-0.586, -0.470)
9	0.953	0.988 (0.985, 0.989)	0.791 (0.759, 0.820)	-0.542 (-0.596, -0.483)
10	0.959	0.989 (0.987, 0.991)	0.790 (0.758, 0.818)	-0.540 (-0.595, -0.481)
11	0.964	0.990 (0.988, 0.992)	0.792 (0.760, 0.820)	-0.542 (-0.596, -0.482)
12	0.966	0.993 (0.991, 0.994)	0.802 (0.771, 0.828)	-0.554 (-0.607, -0.495)
13	0.967	0.993 (0.992, 0.994)	0.802 (0.771, 0.829)	-0.554 (-0.607, -0.495)
14	0.968	0.995 (0.994, 0.996)	0.805 (0.775, 0.832)	-0.556 (-0.609, -0.498)
15	0.971	0.996 (0.995, 0.997)	0.805 (0.774, 0.831)	-0.553 (-0.607, -0.495)
16	0.972	0.998 (0.998, 0.999)	0.805 (0.774, 0.831)	-0.555 (-0.609, -0.497)
17	0.973	0.999 (0.999, 0.999)	0.808 (0.778, 0.834)	-0.560 (-0.613, -0.503)
18	0.974	1.000 (1.000, 1.000)	0.807 (0.777, 0.833)	-0.561 (-0.614, -0.504)

Bold values represent those of the final selected short form.

Appendix Table 2. Correlations for factor scores estimated from the generalized partial credit model

Short Form Length	Correlation of Factor Score with Full Form Factor Score (95% CI)	Correlation of Factor Score with HAQ-DI Score (95% CI)	Correlation of Factor Score with PROMIS-29 Physical Function Score (95% CI)
1	NA	NA	NA
2	0.892 (0.875, 0.908)	0.742 (0.704, 0.776)	-0.527 (-0.583, -0.467)
3	0.937 (0.927, 0.946)	0.777 (0.743, 0.807)	-0.551 (-0.605, -0.493)
4	0.941 (0.931, 0.950)	0.772 (0.737, 0.802)	-0.549 (-0.603, -0.491)
5	0.965 (0.959, 0.970)	0.796 (0.765, 0.824)	-0.570 (-0.622, -0.514)
6	**0.970 (0.965, 0.975)**	**0.807 (0.777, 0.833)**	**-0.581 (-0.632, -0.525)**
7	0.974 (0.969, 0.978)	0.797 (0.766, 0.824)	-0.556 (-0.609, -0.497)
8	0.978 (0.975, 0.981)	0.802 (0.771, 0.829)	-0.561 (-0.614, -0.503)
9	0.984 (0.982, 0.987)	0.808 (0.778, 0.834)	-0.578 (-0.629, -0.521)
10	0.985 (0.983, 0.987)	0.805 (0.775, 0.832)	-0.575 (-0.626, -0.519)
11	0.987 (0.985, 0.989)	0.808 (0.778, 0.834)	-0.577 (-0.628, -0.521)
12	0.990 (0.988, 0.991)	0.817 (0.788, 0.842)	-0.589 (-0.639, -0.534)
13	0.992 (0.991, 0.993)	0.817 (0.788, 0.842)	-0.588 (-0.638, -0.533)
14	0.994 (0.993, 0.995)	0.822 (0.794, 0.847)	-0.591 (-0.641, -0.536)
15	0.995 (0.994, 0.996)	0.821 (0.793, 0.846)	-0.588 (-0.638, -0.533)
16	0.998 (0.998, 0.998)	0.822 (0.794, 0.846)	-0.592 (-0.642, -0.537)
17	0.999 (0.999, 0.999)	0.823 (0.795, 0.847)	-0.594 (-0.644, -0.540)
18	1.000 (1.000, 1.000)	0.822 (0.794, 0.847)	-0.597 (-0.646, -0.542)

Bold values represent those of the final selected short form.