Differential Item Functioning on the Cochin Hand Function Scale among People with Systemic Sclerosis by Language, Sex, and Disease Subtype: a Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Running head: Differential Item Functioning for the 18-Item Cochin Hand Function Scale

Mingyao Xu, BS[1], Daphna Harel, PhD[1], Marie-Eve Carrier, MSc[2], Linda Kwakkenbos, PhD[3-5], Susan J. Bartlett, PhD[6,7], Karen Gottesman[8], Geneviève Guillot[9], Laura Hummers, MD[10], Vanessa L. Malcarne, PhD[11,12], Michelle Richard[13], Brett D. Thombs, PhD[2,6, 14-17] on behalf of the SPIN Investigators[18]

[1] Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York, USA; [2] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada; [3] Department of Clinical Psychology, Radboud University, Nijmegen, the Netherlands; [4] Department of IQ Healthcare, Radboud University Medical Center, Nijmegen, the Netherlands; [5] Department of Psychiatry, Radboud University Medical Center, Nijmegen, the Netherlands; [6] Department of Medicine, McGill University, Montreal, Quebec, Canada; [7] Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada; [8] National Scleroderma Foundation, Los Angeles, California, USA; [9] Sclérodermie Québec, Longueuil, Quebec, Canada; [10] Johns Hopkins University School of Medicine, Division of Rheumatology, Baltimore, Maryland, USA; [11] Department of Psychology, San Diego State University, San Diego, California, USA; [12] San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California, USA; [13]

Scleroderma Atlantic, Halifax, Nova Scotia, Canada; [14] Department of Psychiatry, McGill University, Montreal, Quebec, Canada; [15] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada; [16] Department of Psychology, McGill University, Montreal, Quebec, Canada; [17] Biomedical Ethics Unit, McGill University, Montreal, Quebec, Canada;

[18] SPIN Investigators: Catherine Fortuné, Ottawa Scleroderma Support Group, Ottawa, Ontario, Canada; Amy Gietzen, National Scleroderma Foundation, Tri-State Chapter, Buffalo, New York, USA; Marie Hudson, McGill University, Montreal, Quebec, Canada; Amanda Lawrie-Jones, Scleroderma Australia, and Scleroderma Victoria, Melbourne, Australia; Maureen D. Mayes, Department of Internal Medicine, University of Texas McGovern School of Medicine, Houston, Texas, USA; Warren R. Nielson, Department of Psychology, Western University, and Lawson Research Institute, London, Ontario, Canada; Maureen Sauvé, Scleroderma Society of Ontario, and Scleroderma Canada, Hamilton, Ontario, Canada; Robyn K. Wojeck, University of Rhode Island, Kingston, Rhode Island, USA; Claire Elizabeth Adams, Jewish General Hospital, Montreal, Quebec, Canada; Richard S. Henry, Jewish General Hospital, Montreal, Quebec, Canada; Shervin Assassi, University of Texas McGovern School of Medicine, Houston, Texas, USA; Andrea Benedetti, McGill University, Montreal, Quebec, Canada; Ghassan El-Baalbaki, Université du Québec à Montréal, Montreal, Quebec, Canada; Kim Fligelstone, Scleroderma & Raynaud's UK, London, UK; Tracy Frech, University of Utah, Salt Lake City, Utah, USA; Monique Hinchcliff, Yale School of Medicine, New Haven, Connecticut, USA; Sindhu R. Johnson, Toronto Scleroderma Program, Mount Sinai Hospital, Toronto Western Hospital, and University of Toronto, Toronto, Ontario, Canada; Maggie Larche, McMaster University, Hamilton, Ontario, Canada; Catarina Leite, University of Minho, Braga, Portugal; Christelle

Nguyen, Université Paris Descartes, Université de Paris, Paris, France, and Assistance Publique - Hôpitaux de Paris, Paris, France; Karen Nielsen, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Janet Pope, University of Western Ontario, London, Ontario, Canada; François Rannou, Université Paris Descartes, Université de Paris, Paris, France, and Assistance Publique - Hôpitaux de Paris, Paris, France; Tatiana Sofia Rodriguez-Reyna, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; Anne A. Schouffoer, Leiden University Medical Center, Leiden, The Netherlands; Maria E. Suarez-Almazor, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; Christian Agard, Centre Hospitalier Universitaire - Hôtel-Dieu de Nantes, Nantes, France; Marc André, Centre Hospitalier Universitaire Gabriel-Montpied, Clermont-Ferrand, France; Elana J. Bernstein, Columbia University, New York, New York, USA; Sabine Berthier, Centre Hospitalier Universitaire Dijon Bourgogne, Dijon, France; Lyne Bissonnette, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Alessandra Bruns, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Carlotta Cacciatore, Assistance Publique – Hôpitaux de Paris, Hôpital St-Louis, Paris, France; Patricia Carreira, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Marion Casadevall, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Benjamin Chaigne, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Lorinda Chung, Stanford University, Stanford, California, USA; Benjamin Crichi, Assistance Publique - Hôpitaux de Paris, Hôpital St-Louis, Paris, France; Robyn Domsic, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; James V. Dunne, St. Paul's Hospital and University of British Columbia, Vancouver, British Columbia, Canada; Bertrand Dunogue, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Regina Fare, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Dominique Farge-Bancel, Assistance Publique -

Hôpitaux de Paris, Hôpital St-Louis, Paris, France; Paul R. Fortin, CHU de Québec - Université Laval, Quebec, Quebec, Canada; Jessica Gordon, Hospital for Special Surgery, New York City, New York, USA; Brigitte Granel-Rey, Université, and Assistance Publique - Hôpitaux de Marseille, Hôpital Nord, Marseille, France; Aurélien Guffroy, Les Hôpitaux Universitaires de Strasbourg, Nouvel Hôpital Civil, Strasbourg, France; Genevieve Gyger, Jewish General Hospital and McGill University, Montreal, Quebec, Canada; Eric Hachulla, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, Sabrina Hoa, Centre hospitalier de l'Université de Montréal – CHUM, Montreal, Quebec, Canada; Alena Ikic, CHU de Québec - Université Laval, Quebec, Quebec; Niall Jones, University of Alberta, Edmonton, Alberta, Canada; Nader Khalidi, McMaster University, Hamilton, Ontario, Canada; Kimberly Lakin, Hospital for Special Surgery, New York City, New York, USA; Marc Lambert, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; David Launay, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Yvonne C. Lee, Northwestern University, Chicago, Illinois, USA; Hélène Maillard, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Nancy Maltez, University of Ottawa, Ottawa, Ontario, Canada; Joanne Manning, Salford Royal NHS Foundation Trust, Salford, UK; Isabelle Marie, CHU Rouen, Hôpital de Bois-Guillaume, Rouen, France; Maria Martin Lopez, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Thierry Martin, Les Hôpitaux Universitaires de Strasbourg, Nouvel Hôpital Civil, Strasbourg, France; Ariel Masetto, Université de Sherbrooke, Sherbrooke, Quebec, Canada; François Maurier, Uneos - Groupe hospitalier associatif, Metz, France; Arsene Mekinian, Assistance Publique - Hôpitaux de Paris, Hôpital St-Antoine, Paris, France; Sheila Melchor Díaz, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Mandana Nikpour, St Vincent's

Hospital and University of Melbourne, Melbourne, Victoria, Australia; Louis Olagne, Centre Hospitalier Universitaire Gabriel-Montpied, Clermont-Ferrand, France; Vincent Poindron, Les Hôpitaux Universitaires de Strasbourg, Nouvel Hôpital Civil, Strasbourg, France; Susanna Proudman, Royal Adelaide Hospital and University of Adelaide, Adelaide, South Australia, Australia; Alexis Régent, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Sébastien Rivière, Assistance Publique - Hôpitaux de Paris, Hôpital St-Antoine, Paris, France; David Robinson, University of Manitoba, Winnipeg, Manitoba, Canada; Esther Rodríguez Almazar, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Sophie Roux, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Perrine Smets, Centre Hospitalier Universitaire Gabriel-Montpied, Clermont-Ferrand, France; Vincent Sobanski, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Robert Spiera, Hospital for Special Surgery, New York City, 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; Pearce Wilcox, St. Paul's Hospital and University of British Columbia, Vancouver, British Columbia, Canada; Mara Cañedo Ayala, Jewish General Hospital, Montreal, Quebec, Canada; Vanessa Cook, Jewish General Hospital, Montreal, Quebec, Canada; Sophie Hu, Jewish General Hospital, Montreal, Quebec, Canada; Bianca Matthews, Jewish General Hospital, Montreal, Quebec, Canada; Elsa-Lynn Nassar, Jewish General Hospital, Montreal, Quebec, Canada; Marieke A. Neyer, Jewish General Hospital, Montreal, Quebec, Canada; Julia Nordlund, Jewish General Hospital, Montreal, Quebec, Canada; Sabrina Provencher, Jewish General Hospital, Montreal, Quebec, Canada.

Financial Support: The Scleroderma Patient-centered Intervention Network (SPIN) Cohort has received funding from the Canadian Institutes of Health Research (CIHR); the Lady Davis Institute for Medical Research of the Jewish General Hospital, Montreal, Quebec, Canada; the Jewish General Hospital Foundation, Montreal, Quebec, Canada; McGill University, Montreal, Quebec, Canada; the Scleroderma Society of Ontario; Scleroderma Canada; Sclérodermie Québec; Scleroderma Manitoba; Scleroderma Atlantic; the Scleroderma Association of BC; Scleroderma SASK; Scleroderma Australia; Scleroderma New South Wales; Scleroderma Victoria; and Scleroderma Queensland. Dr. Thombs was supported by a Tier 1 Canada Research Chair, outside of the present work. No sponsor had any role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Conflict of Interests: The authors have no conflicts of interest to declare.

Address for Correspondence: Daphna Harel, PhD; New York University; 246 Greene Street, New York, NY, 10003; E-mail: daphna.harel@nyu.edu; ORCID: 0000-0001-7015-5989

Word Count: 2861

ABSTRACT

Objective: The Cochin Hand Function Scale (CHFS) is commonly used among people with systemic sclerosis (SSc), including study participants who speak different languages, are of different sexes, or have different disease subtypes. It is not known, however, whether CHFS displays differential item functioning (DIF) and if scores from participants in different groups can be treated equivalently. We evaluated the degree that the CHFS generates scores that are comparable across language, sex, and disease subtype.

Methods: We included participants enrolled in the Scleroderma Patient-centered Intervention Network (SPIN) Cohort who completed the CHFS at their baseline assessment between April 2014 and September 2020. Confirmatory factor analysis (CFA) was used to test unidimensionality, and Multiple Indicator Multiple Cause (MIMIC) models were used for DIF analysis based on language, sex, and disease subtype. Both intraclass correlation coefficient (ICC) and Pearson's correlation were calculated using factor scores obtained from unadjusted and DIF-adjusted MIMIC models to evaluate agreement and correlation between scores.

Results: 2155 participants were included. CFA with covarying error terms supported a good fit of the model ($\chi 2[127] = 1754.671$, $P < 0.001$, TLI = 0.985, CFI = 0.987, RMSEA = 0.077). Nine items displayed statistically significant DIF for language of administration, 10 items for sex, and 10 items for disease subtype. However, the overall impact of DIF was negligible when comparing factor scores that did and did not account for DIF (ICC= 0.999, r = 0.999).

Conclusion: The CHFS has score comparability in SSc regardless of participants' language, sex, and disease subtype.

SIGNIFICANCE AND INNOVATION

- Hand function is an important contributor to disability in systemic sclerosis (SSc), and the Cochin Hand Function Scale (CHFS) is commonly used in SSc clinical trials and multi-national observational studies.
- This is the first study to evaluate if CHFS items display differential item functioning by language (English and French), sex, and disease subtype.
- Some CHFS items display differential item functioning for participants taking the CHFS in different languages, are of different sexes, and have different disease subtypes, but the impact on total scale scores is negligible.
- The CHFS can be used and compared among participants with SSc across different languages, sexes, and disease subtypes.

Systemic sclerosis (SSc; scleroderma) is a rare chronic autoimmune disease characterized by fibrosis of the skin and internal organs (1). Digital ulcers, contractures, and deformities of the hand can lead to decreased flexion and limited extension (2). These symptoms impact hand function and can result in substantial impairment (3). The Cochin Hand Function Scale (CHFS) was developed to measure the functional ability of the hand among people with rheumatic diseases (4) and has been validated (5,6) and used extensively in SSc (2,6–9). The self-report CHFS consists of 18 items used to assess a person's ability to perform daily hand-related activities (4).

The cross-language validity of the CHFS is important in SSc because SSc is a rare disease (10), and people who complete a scale in different languages are commonly included in the same study (11), especially when the study is carried out in countries or regions with more than one commonly spoken language, such as Canada (e.g., French and English). Additionally, for rare diseases such as SSc, international collaboration and recruitment of participants from different countries who use different languages is often necessary to include sufficient numbers of participants in a given study (7,9,12,13).

In addition, because approximately 85% of people with SSc are female (14,15), it is important to ensure the measurements obtained from the CHFS are comparable regardless of sex. Previous validations have been done with very small numbers of male participants, and thus it is hard to evaluate the equivalence of measurement. For example, out of 40 participants in the first study that validated the CHFS, then called the Duruöz Hand Index, in SSc, only 6 participants were male (5).

SSc has two main subtypes - limited and diffuse (16), and disease severity, which is reflected in subtypes, is an important indicator of hand function (12). Therefore, it is important to assess the degree to which scores from the CHFS may systematically differ by disease subtype.

Differential item functioning (DIF) occurs when members of one group (e.g., English-language responders) have a different expected score on an item compared to members of another group (e.g., French-language responders), after controlling for any differences in the construct being measured (e.g., hand function) (17,18). Therefore, the responses to an item are influenced, not only by the level of the hand function the person has, but also by the grouping factor (e.g., whether they completed the scale in French or English).

The purpose of the study was to evaluate whether: (1) the CHFS displays DIF with respect to language (English or French), sex (male or female), and disease subtype (limited or diffuse); and (2) if any identified statistically significant DIF influences CHFS scores to a non-negligible extent.

METHODS

This was a cross-sectional study evaluating baseline data from the Scleroderma Patient-centered Intervention Network (SPIN) Cohort (7). A protocol was published online prior to study initiation (https://osf.io/qb8m3/). Because of overlap with previous studies, we adopted part of the methods from previous work (12), including the description of the SPIN Cohort in the Participants and Procedure section, and procedures and study variables in the Measures section. This is in line with guidance from the Text Recycling Research Project (19).

Participants and Procedure

The SPIN Cohort is a convenience sample. Eligible patients at SPIN recruiting sites are invited by the attending physician or a nurse coordinator to participate. Eligible participants must

be classified as having SSc according to 2013 ACR/EULAR classification criteria; ≥ 18 years of age; and fluent in English, French, or Spanish (20). After written informed consent is obtained, the recruiting site physician or nurse coordinator completes and submits an online medical data form. An automated email is then sent to participants with instructions on activating their SPIN online account and completing measures. SPIN Cohort participants complete outcome measures via an online portal upon enrolment and subsequently every three months. The SPIN Cohort consists of data from 51 centers in Canada, the United States, the United Kingdom, France, Spain, Mexico, and Australia. The SPIN Cohort study was approved by the Research Ethics Committee of the Centre intégré universitaire de santé et de services sociaux du Centre-Ouest-de-l'Île-de-Montréal (#MP-05-2013-150) and by the ethics committees of all recruiting sites.

The present study used baseline assessment data from participants enrolled between April 2014 and September 2020 who completed the CHFS in English or French only, and with complete item-level data for the CHFS, and complete data on language of instrument completion, sex, and disease subtype.

Measures

Sociodemographic and medical data

Participants provided marital status, years of education, number of cigarettes smoked per week, and number of alcoholic drinks per week. SPIN physicians completed a medical data form that included all items of the 2013 ACR/EULAR SSc classification criteria (20) and provided age, sex, time since the first non-Raynaud's phenomenon symptoms and diagnosis, SSc subtype (limited or diffuse cutaneous SSc) (16), presence of overlap syndromes (systemic lupus erythematosus, rheumatoid arthritis, Sjögren's syndrome, idiopathic inflammatory myositis), and presence of joint contractures (no/mild (0–25%) versus moderate/severe (>25%) limit in range of

motion). Standard numeric rating scales were completed by patients for Raynaud's severity in the past week and severity of finger ulcers, ranging from 0 (not severe at all) to 10 (unbearable).

In the SPIN cohort, participants self-report race or ethnicity data using the standard categories that are used in each country. Because categories differ across countries, and categories used in one country may not be recognized by participants from other countries, we characterized study participants by aggregating them as White, Black, and Other. The categories used in each country are presented in Supplemental Material A.

Hand Function (CHFS)

The 18-item CHFS (4) was developed to measure the ability to perform daily hand-related activities. Items reflecting 5 content areas (i.e., kitchen, dressing oneself, hygiene, writing/typing, other) are scored on a 0-5 Likert scale (0=without difficulty; 5=impossible). The total score is obtained by adding the scores of all items (range 0-90), and higher scores indicate more difficulty in hand function. Validity and reliability of the CHFS have been confirmed in SSc (5,6).

Statistical analysis

Descriptive statistics were calculated for all variables and all participants in the sample. We fit a unidimensional confirmatory factor analysis (CFA) model to the CHFS data using a robust weighted least squares variance estimator (21) to test the unidimensionality of the underlying latent trait (hand function). We chose to assess a unidimensional model in order to evaluated whether the standard practice of scoring the CHFS with a simple summed score is justified. To evaluate the unidimensional model, we determined fit via a mean- and variance-adjusted chi-square test statistic, 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 is highly dependent on the sample size of the study and may reject the model despite its good fit (22). Values of CFI and TLI ≥0.95 and RMSEA ≤0.08 were considered to indicate a good fit for the model (23–25). We also calculated modification indices to recognize item pairs for which measurement errors correlate highly (26). If there was also theoretical justification for shared effects within these pairs of items, we then allowed their errors to co-vary if this improved model fit.

We then used Multiple Indicator Multiple Cause (MIMIC) models to determine if items of the CHFS exhibited DIF when different groups were compared by language, sex, and disease subtype. We first fit a baseline MIMIC model that only included the path between the grouping variables and the latent variable, hand function (27,28). This model is a unidimensional CFA model with the additional paths between all groups and the latent variable to capture any mean differences in scores for participants in different groups. Next, we used the constrained baseline approach. Specifically, for each grouping variable, we fit 18 augmented models, each with paths from the grouping variable to the individual CHFS item. For each of the grouping variables, we iterated this process 18 times for each item separately. Meanwhile, we noted the statistical significance of the coefficient of the path between the grouping variables and each item. Once we identified all the items that displayed DIF and which grouping variable(s) were the sources of DIF, we constructed the final MIMIC model by adding paths between all the DIF items and the corresponding grouping variables to the baseline MIMIC model, even if these paths were no longer statistically significant in the final DIF-adjusted MIMIC model in order to be conservative in our model choice. We did not employ a Type 1 error correction for the p values from the original sets of models in order to capture any possible DIF across items.

Lastly, we assessed the effect of DIF on latent factor scores. This is important because we included 2155 participants and, due to the large sample size, we expected to detect statistically significant DIF for potentially many items. Use of an effect size measure indicates whether any statistically significant DIF has an actual, meaningful impact on the reason the CHFS is administered – to obtain scores for hand function for participants. Ideally, clinical decisions are based on highly precise estimates and effect sizes and not on analyses of statistical significance. Therefore we calculated the agreement between the scores obtained from the MIMIC baseline model and the final DIF-adjusted MIMIC model through the intraclass correlation coefficient (ICC) and its 95% confidence interval (29). As a secondary measure, we also calculated Pearson's correlation coefficient and its 95% confidence interval (30). Upon analyzing all the MIMIC models, we identified if any CHFS item exhibited DIF and which grouping variable(s) contributed to DIF, as well as whether any observed DIF impacted the factor scores that were estimated from the participants' responses. A high ICC or correlation would indicate that any statistically significant DIF had meaningful impact, while a low ICC or correlation would indicate that although there was statistically significant DIF, it may not have clinical impact.

All analyses were conducted in R (31), with the CFA and MIMIC models fit using the MplusAutomation package (32).

Sample size calculation

Recommendations for CFA sample size vary. In the present study, we performed a single-factor CFA and multiple MIMIC models with 18 indicators, using a sample of 2155 participants. This number significantly surpassed the minimum sample size recommended by many established recommendations and standards (33,34) to ensure a substantial agreement between true sample characteristics and model estimates.

RESULTS

Sample characteristics

Within the SPIN Cohort, 2240 participants had complete data for all CHFS items with 2178 of those in English or French. However, only 2155 participants had complete data for all variables in the CFA and MIMIC model analyses (i.e., sex, disease subtype) and were included in this study. There were 1882 females (87.3%) and 273 males (12.7%; see Table 1); 1459 people responded to the CHFS in English (67.7%) and 696 in French (32.3%); 842 respondents presented with diffuse SSc (39.1%); and 1313 respondents presented with limited or sine SSc (60.9%). 1788 (83.0%) self identified as White. Most participants were married or living as married (61.8%). The mean time since first non-Raynaud's symptoms was 11.1 (SD=8.8) years, and the mean time since diagnosis was 9.4 (SD=8.1) years. The mean CHFS score was 13.5 (SD=16.1). There were 63 (2.9%) participants with sine disease subtype who were grouped together with the 1250 (58.0%) participants with limited disease subtype for all following analyses.

Confirmatory factor analysis

A unidimensional CFA model of the CHFS items, where covariance of item residuals was restricted to zero, resulted in less than ideal fit ($\chi 2[135] = 5232.629$, $P < 0.001$, TLI = 0.955, CFI = 0.960, RMSEA = 0.132).

The modification indices suggested allowing error measurements of the following items to covary: items 1 and 2, items 2 and 3, items 2 and 4, items 3 and 4, items 9 and 10, items 9 and 12, items 9 and 17, and items 13 and 14. For example, item 13 measures how well participants can write a short sentence with a pencil or an ordinary pen, and item 14 measures how well participants can write a letter with a pencil or an ordinary pen, which are extremely similar. Due

to high degree of similarity across the content or wording of these CHFS items, we allowed all pairs of items with large modification indices to have correlated covariance terms until the CFA model had adequate fit. Therefore, the CFA model was refitted with allowing the error terms of these items to covary, and the refitted model indicated a good fit ($\chi 2[127] = 1754.671$, $P < 0.001$, TLI = 0.985, CFI = 0.987, RMSEA = 0.077).

DIF analysis

The baseline MIMIC model with paths between each grouping variable and the latent variable demonstrated good fit ($\chi 2[178] = 2173.740$, $P < 0.001$, TLI = 0.982, CFI = 0.984, RMSEA = 0.072). The baseline MIMIC model's parameters can be found in Table 2.

Using iterations to identify DIF for each grouping variable, we found that 9 items displayed DIF for the grouping variable of language of CHFS administration, 10 items displayed DIF for the grouping variable of the respondent's sex, and 10 items displayed DIF for the grouping variable of the respondent's disease subtype. See Table 3 for the p-values of each of statistically significant paths in the MIMIC models.

Table 4 shows the final MIMIC model parameters after correcting for DIF. Estimated group differences on the latent factor did not differ meaningfully depending on whether we controlled for DIF. The difference between the two language groups (French - English) on the latent factor was not statistically significant for either the model with DIF adjustment (standardized mean differences [SMD] = -0.048, 95% CI -0.150 to 0.053, p = 0.352) or without adjustment (SMD = -0.049, 95% CI -0.149 to 0.052, p = 0.343). The difference between the two sex groups (Male - Female) on the latent factor was statistically significant for both the final MIMIC model with DIF adjustment (SMD = -0.282, 95% CI -0.432 to -0.131, $p < 0.001$) and the baseline MIMIC model (SMD = -0.292, 95% CI -0.439 to -0.146, $p < 0.001$). The difference

between the two disease subtype groups (Diffuse - Limited) on the latent factor was statistically significant for both the model with DIF adjustment (SMD = 0.624, 95% CI 0.526 to 0.722, $p < 0.001$) and without adjustment (SMD = 0.638, 95% CI 0.541 to 0.735, $p < 0.001$).

The ICC between the factor scores obtained from the baseline MIMIC model and the ones from the final MIMIC model was 0.999 (95% CI 0.999, 0.999). Pearson's correlation coefficient between the factor scores obtained from the baseline MIMIC model and the ones from the final MIMIC model was 0.998 (95% CI 0.998, 0.998).

DISCUSSION

We tested the unidimensional structure of the CHFS and examined whether there were meaningful differences in measurement properties on the latent variable with three grouping variables - language, sex, and disease subtype - in a sample of participants with SSc. We confirmed the unidimensionality of the latent trait and found that while there was statistically significant DIF in items of the CHFS, the overall impact of DIF on scores was negligible.

Although there was statistically significant DIF for 9 items between English- and French-language participants, 10 items between male and female participants, and 10 items between participants with limited and diffuse disease subtype, the cumulative effect of DIF was minimal and did not meaningfully influence estimates of hand function differences of participants, regardless of their language, sex, or disease subtype. The high Pearson's correlation (0.998) and ICC (0.999) between factor scores from models that did and did not account for DIF allowed us to conclude that CHFS scores of French- and English-language, male and female, diffuse and limited subtype participants can be aggregated and compared without concerns of bias due to the grouping factors we studied. The lack of impact of DIF on the CHFS may be due, in part, to the wording of the items. Specifically, all items assess concrete abilities to perform a certain task

that requires the use of hands, and not abstract concepts. This, in turn, may reduce the likelihood of DIF based on participant characteristics.

The present study is the first to assess DIF of CHFS using MIMIC models and the first to compare measurement properties based on participants' language, sex, and disease subtype. Our findings have important implications for research. This study's result demonstrated the comparability of CHFS scores across English and French-languages in SSc. Furthermore, regardless of participants' sex and disease subtype, their CHFS scores can be compared without scaling or DIF correction. Considering SSc is a rare disease, with its overall pooled prevalence of SSc approximately 17.6 per 100,000 people (35), local or regional samples can be limited. Our study supports the use of the CHFS in larger-scale collaborations and promotes broader utilization in international participants cohorts, such as the SPIN Cohort. Additionally, interventions and treatments aimed at improving hand functionality have been shown to reduce symptom burden among individuals with SSc to some degree (8). The CHFS is a valid outcome measure that can be used to measure hand function in patients with SSc across language, sex, and disease severity. Future work may investigate sensitivity to change for the CHFS, therefore allowing it to be used to test interventions and treatments.

There are several noteworthy strengths of our study, including its international cohort recruited from 51 clinical sites, its large sample, and the assessment of measurement properties among people with SSc in multiple languages. Although this study focused on determining the impact of DIF for the CHFS for people with SSc based on their language, sex, and disease subtype, the MIMIC models we used could be applied to other participant populations and other measures for DIF identification and correction.

The present study, however, represents only a first step in using the DIF approach to attempt to standardize processes for validating CHFS among people with SSc with different backgrounds and medical histories. There are also limitations to our study. First, the SPIN Cohort is a convenience sample, and thus may not represent the SSc population. For example, the cohort participants completed all the required measures online. Second, the examination of DIF was limited to English- and French-speaking participants, and therefore the generalizability of the finding based on our sample population is unknown. Third, we only examined uniform DIF in this study with the assumption of a constant relationship between measures and grouping variables; we did not examine non-uniform DIF (36). We only examined the differences in mean across groups and did not examine the patterns. However, because in practice the CHFS is scored with a summed score that does not allow for varying factor loadings, any non-uniform DIF would not influence how the CHFS is scored. Lastly, future research may investigate whether our results are replicable under other well-known methods for DIF detection, such as those that use item response theory methods.

Overall, the results of this study indicated that while the CHFS displayed statistically significant DIF across language of administration, participant sex, and disease subtype, the impact of this DIF was negligible on scores obtained from the scale. This means that participants' CHFS scores can be compared without DIF adjustment, which supports the use of the CHFS in studies that administer the scale in different languages or recruit participants with SSc of different sexes or with different levels of disease severity.

REFERENCES

1. Allanore Y, Simms R, Distler O, et al. Systemic sclerosis. Nat Rev Dis Primer. 2015;1(1):15002.

2. Mouthon L, Carpentier PH, Lok C, et al. Ischemic digital ulcers affect hand disability and pain in systemic sclerosis. J Rheumatol. 2014;41(7):1317–23.

3. Young A, Namas R, Dodge C, et al. Hand impairment in systemic sclerosis: various manifestations and currently available treatment. Curr Treat Options Rheumatol. 2016;2(3):252–69.

4. Duruöz MT, Poiraudeau S, Fermanian J, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol. 1996;23(7):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(5):805–9.

6. Rannou F, Poiraudeau S, Berezné A, 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(1):94–102.

7. Kwakkenbos L, Jewett LR, Baron M, 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(8):e003563.

8. Rannou F, Boutron I, Mouthon L, et al. Personalized physical therapy versus usual care for patients with systemic sclerosis: a randomized controlled trial. Arthritis Care Res. 2017;69(7):1050–9.

9. Kwakkenbos L, Carrier ME, Welling J, et al. Randomized controlled trial of an internet-based self-guided hand exercise program to improve hand function in people with systemic sclerosis: the Scleroderma Patient-centered Intervention Network Hand Exercise Program (SPIN-HAND) Trial. Trials. 2022;23(1):994.

10. Bergamasco A, Hartmann N, Wallace L, et al. Epidemiology of systemic sclerosis and systemic sclerosis-associated interstitial lung disease. Clin Epidemiol. 2019;11:257–73.

11. Bairwa D, Kavadichanda CG, Adarsh MB, et al. Cultural adaptation, translation and validation of Cochin Hand Function Scale and evaluation of hand dysfunction in systemic sclerosis. Clin Rheumatol. 2021;40(5):1913–22.

12. Kwakkenbos L, Sanchez TA, Turner KA, et al. The association of sociodemographic and disease variables with hand function: a Scleroderma Patient-centered Intervention Network Cohort study. Clin Exp Rheumatol. 2018;36 Suppl 113(4):88–94.

13. Levis AW, Harel D, Kwakkenbos L, et al. 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 Cohort study: using optimal test assembly to shorten the CHFS. Arthritis Care Res. 2016;68(11):1704–13.

14. Peoples C, Medsger TA, Lucas M, et al. Gender differences in systemic sclerosis: relationship to clinical features, serologic status and outcomes. J Scleroderma Relat Disord. 2016;1(2):177–240.

15. Hughes M, Pauling JD, Armstrong-James L, et al. Gender-related differences in systemic sclerosis. Autoimmun Rev. 2020;19(4):102494.

16. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol. 1988;15(2):202–5.

17. Holland PW, Wainer H. (eds.) Differential Item Functioning. Routledge; 2012.

18. Chen WH, Revicki D. Differential Item Functioning (DIF). In: Michalos AC. (ed.) Encyclopedia of quality of life and well-being research. Springer; 2014:1611–4.

19. Hall S, Moskovitz C, Pemberton M. Understanding text recycling: a guide for researchers. Text Recycling Research Project. 2021. https://textrecycling.org. Accessed April 6, 2023.

20. van den Hoogen F, Khanna D, Fransen J, et al. Classification criteria for systemic sclerosis: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis. 2013;72(11):1747–55.

21. Muthén LK, Muthén BO. Mplus User's Guide. 8th Edition. Los Angeles, CA: Muthén & Muthén; 2017.

22. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull. 1993;114(3):552–66.

23. Chen F, Curran PJ, Bollen KA, et al. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociol Methods Res. 2008;36(4):462–94.

24. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999;6(1):1–55.

25. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Methods Res. 1992;21(2):230–58.

26. McDonald RP, Ho MHR. Principles and practice in reporting structural equation analyses. Psychol Methods. 2002;7(1):64–82.

27. Chun S, Stark S, Kim ES, et al. MIMIC methods for detecting DIF among multiple groups: exploring a new sequential-free baseline procedure. Appl Psychol Meas. 2016;40(7):486–99.

28. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. J Psychopathol Behav Assess. 2009;31(4):320–30.

29. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res. 2017;8(4):187–91.

30. Liu J, Tang W, Chen G, et al. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch Psychiatry. 2016;28(2):115–20.

31. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. https://www.R-project.org. Accessed April 6, 2023.

32. Hallquist MN, Wiley JF. MplusAutomation : an R package for facilitating large-scale latent variable analyses in Mplus. Struct Equ Model Multidiscip J. 2018;25(4):621–38.

33. Mundfrom DJ, Shaw DG, Ke TL. Minimum sample size recommendations for conducting factor analyses. Int J Test. 2005;5(2):159–68.

34. Wolf EJ, Harrington KM, Clark SL, et al. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. Educ Psychol Meas. 2013;73(6):913–34.

35. Bairkdar M, Rossides M, Westerlind H, et al. Incidence and prevalence of systemic sclerosis globally: a comprehensive systematic review and meta-analysis. Rheumatology. 2021;60(7):3121–33.

36. Woods CM, Grimm KJ. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. Appl Psychol Meas. 2011;35(5):339–61.

TABLES

Table 1. Participant demographic and. disease characteristics (n=2155)

Variable	N (%) or Mean (SD)
Demographic	
English language, n (%)	1459 (67.7)
Female sex, n (%)	1882 (87.3)
Race or ethnicity,[a] n (%) #	
White	1788 (83.0)
Black	149 (6.9)
Other	216 (10.0)
Age in years, mean (SD)	55.0 (12.6)
Marital status, n (%)	
Married	1332 (61.8)
Living as married	196 (9.1)
Separated or divorced	257 (11.9)
Widowed	97 (4.5)
Single	273 (12.7)
Education in years, mean (SD)	14.9 (3.7)
Alcohol consumption(drinks/week), n (%)	
0	1224 (56.8)
1-7	773 (35.9)
8+	158 (7.3)
Cigarette consumption (cigarettes/day), n (%)	
0	2001 (92.9)
1-9	75 (3.5)
10-19	57 (2.6)
20+	22 (1.0)
Disease characteristics	
Time since onset first non-Raynaud's symptom in years, mean (SD) †	11.1 (8.8)
Time since diagnosis, n (%) ‡	9.4 (8.1)
Disease subtype, n (%)	
Limited	1250 (58.0)
Diffuse	842 (39.1)

Sine	63 (3.0)
Patient-reported severity of Raynaud's, mean (SD) §	3.8 (2.8)
Patient-reported severity of finger ulcers, mean (SD) ¶	1.7 (2.7)
Small joint contractures, n (%)	
None or mild	1505 (69.8)
Moderate	383 (17.8)
Severe	151 (7.0)
Not available	116 (5.4)
Large joint contractures, n (%)	
None or mild	1743 (80.9)
Moderate	185 (8.6)
Severe	70 (3.2)
Not available	157 (7.3)
Presence of systemic lupus erythematosus, n (%)	63 (2.9)
Presence of Sjögren's syndrome, n (%)	164 (7.6)
Presence of rheumatoid arthritis, n (%)	119 (5.5)
Presence of idiopathic inflammatory myositis, n (%)	107 (5.0)
Cochin Hand Function Scale total score, mean (SD)	13.5 (16.1)

SSc = systemic sclerosis; CHFS = Cochin Hand Function Scale

[a] Because ethnicity/race information is collected differently across countries, it is aggregated here into the categories "White", "Black", and "Other". See Supplementary Material A for further details about race or ethnicity grouping.
N = 2153; † N = 1975; ‡ N = 2072; § N = 2132; ¶ N = 2131

Table 2. Factors Loadings of the Baseline MIMIC Model

Variable	Estimate	95% Confidence Interval
Item 1 Hold bowl	0.874	(0.858, 0.890)
Item 2 Raise bottle	0.817	(0.797, 0.836)
Item 3 Hold plate	0.857	(0.841, 0.874)
Item 4 Pour liquid	0.865	(0.849, 0.881)
Item 5 Unscrew lid	0.810	(0.793, 0.828)
Item 6 Cut meat	0.887	(0.874, 0.899)
Item 7 Prick fork	0.876	(0.857, 0.896)
Item 8 Peel fruit	0.895	(0.883, 0.907)
Item 9 Button shirt	0.870	(0.858, 0.883)
Item 10 Zipper	0.879	(0.865, 0.893)
Item 11 Toothpaste tube	0.875	(0.857, 0.893)
Item 12 Hold toothbrush	0.865	(0.848, 0.882)
Item 13 Write short	0.842	(0.824, 0.860)
Item 14 Write letter	0.786	(0.766, 0.807)
Item 15 Doorknob	0.885	(0.873, 0.897)
Item 16 Cut paper	0.897	(0.884, 0.911)
Item 17 Pick up coins	0.854	(0.841, 0.867)
Item 18 Turn key	0.906	(0.895, 0.917)
Item 2 with Item 1	0.110	(0.095, 0.125)
Item 2 with Item 3	0.164	(0.145, 0.182)
Item 2 with Item 4	0.144	(0.126, 0.161)
Item 3 with Item 4	0.124	(0.107, 0.141)
Item 9 with Item 10	0.103	(0.086, 0.120)
Item 9 with Item 17	0.069	(0.056, 0.082)
Item 11 with Item 12	0.087	(0.071, 0.103)
Item 13 with Item 14	0.243	(0.221, 0.266)
Hand function on language	-0.049	(-0.149, 0.052)
Hand function on sex	-0.292	(-0.439, -0.146)
Hand function on disease subtype	0.638	(0.541, 0.735)

Table 3. P-values for Items Displaying DIF

	Initial Models			MIMIC Models		
Variable	Item on Language	Item on Sex	Item on Disease Subtype	Item on Language	Item on Sex	Item on Disease Subtype
Item 2	<0.001	0.001	–	<0.001	<0.001	–
Item 3	<0.001	–	<0.001	<0.001	–	0.029
Item 4	0.028	0.006	0.035	0.065	0.006	0.186
Item 5	<0.001	<0.001	–	<0.001	<0.001	–
Item 6	0.026	0.001	<0.001	0.095	<0.001	<0.001
Item 8	–	–	0.003	–	–	0.002
Item 9	–	<0.001	–	–	<0.001	–
Item 10	–	–	0.003	–	–	0.020
Item 11	–	–	0.012	–	–	0.035
Item 12	0.006	0.004	0.031	0.013	0.006	0.058
Item 13	<0.001	0.001	0.002	<0.001	<0.001	0.001
Item 14	<0.001	–	–	<0.001	–	–
Item 15	–	0.006	–	–	0.002	–
Item 16	–	–	0.020	–	–	0.011
Item 17	–	<0.001	<0.001	–	<0.001	–
Item 18	0.006	–	–	0.008	–	–

Table 4. Factors Loadings of the final MIMIC Model

Variable	Estimate (95% CI)	Item on Language Estimate (95% CI)	Item on Sex Estimate (95% CI)	Item on Disease Subtype Estimate (95% CI)
Item 1	0.874 (0.858, 0.890)	–	–	–
Item 2	0.816 (0.797, 0.836)	0.195 (0.121, 0.270)	-0.198 (-0.304, -0.092)	–
Item 3	0.858 (0.841, 0.874)	0.426 (0.356, 0.496)	–	-0.070 (-0.133, -0.007)
Item 4	0.865 (0.849, 0.881)	0.069 (-0.004, 0.141)	-0.131 (-0.224, -0.037)	-0.043 (-0.106, 0.020)
Item 5	0.81 (0.793, 0.828)	-0.18 (-0.256, -0.104)	-0.342 (-0.448, -0.236)	–
Item 6	0.886 (0.874, 0.899)	-0.057 (-0.124, 0.010)	-0.187 (-0.278, -0.096)	0.132 (0.072, 0.192)
Item 7	0.876 (0.857, 0.896)	–	–	–
Item 8	0.895 (0.883, 0.906)	–	–	0.093 (0.034, 0.151)
Item 9	0.871 (0.858, 0.883)	–	0.185 (0.097, 0.273)	–
Item 10	0.879 (0.865, 0.893)	–	–	-0.072 (-0.132, -0.011)
Item 11	0.875 (0.857, 0.893)	–	–	-0.078 (-0.15, -0.006)
Item 12	0.865 (0.848, 0.882)	-0.100 (-0.179, -0.021)	0.140 (0.039, 0.241)	0.069 (-0.002, 0.14)
Item 13	0.842 (0.824, 0.86)	-0.170 (-0.248, -0.093)	0.171 (0.079, 0.264)	-0.115 (-0.182, -0.047)
Item 14	0.787 (0.766, 0.807)	-0.152 (-0.23, -0.074)	–	–
Item 15	0.885 (0.873, 0.897)	–	-0.142 (-0.233, -0.05)	–
Item 16	0.897 (0.884, 0.911)	–	–	0.079 (0.018, 0.14)
Item 17	0.854 (0.840, 0.867)	–	0.211 (0.109, 0.314)	0.174 (0.112, 0.235)
Item 18	0.906 (0.895, 0.917)	-0.087 (-0.150, -0.023)	–	–
Item 1 with Item 2	0.110 (0.095, 0.125)			

Item 2 with Item 3	0.164 (0.145, 0.182)
Item 2 with Item 4	0.144 (0.126, 0.161)
Item 3 with Item 4	0.124 (0.106, 0.141)
Item 9 with Item 10	0.103 (0.086, 0.120)
Item 9 with Item 12	0.087 (0.071, 0.103)
Item 9 with Item 17	0.069 (0.056, 0.082)
Item 13 with Item 14	0.243 (0.220, 0.266)
Hand function on language	-0.048 (-0.150, 0.053)
Hand function on sex	-0.282 (-0.432, -0.131)
Hand function on disease subtype	0.624 (0.526, 0.722)