BMJ Open

Assessing Differential Item Functioning for the Social Appearance Anxiety Scale: A Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037639
Article Type:	Original research
Date Submitted by the Author:	12-Feb-2020
Complete List of Authors:	Sommer, Sophia; New York University, Applied Statistics, Social Science, and Humanities; New York University, PRIISM Applied Statistics Center Harel, Daphna; New York University, Applied Statistics, Social Science, Humanities; New York University, PRIISM Applied Statistics Center Kwakkenbos, Linda; Radboud University, Behavioral Science Institute Carrier, Marie-Eve; Jewish General Hospital, Lady Davis Institute for Medical Research Gholizadeh, Shadi; San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology; McGill University, Psychiatry Gottesman, Karen; Scleroderma Foundation Leite, Catarina; University of Minho, School of Psychology Malcarne, Vanessa; San Diego State University, Psychology; San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology Thombs, Brett; McGill University, Psychiatry
Keywords:	STATISTICS & RESEARCH METHODS, RHEUMATOLOGY, Anxiety disorders < PSYCHIATRY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Assessing Differential Item Functioning for the Social Appearance Anxiety Scale: A Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Sophia J. Sommer, BA[1,2]; Daphna Harel, PhD[1,2]; Linda Kwakkenbos, PhD[3]; Marie-Eve Carrier, MSc[4]; Shadi Gholizadeh, PhD[5,6]; Karen Gottesman, BA[7]; Catarina Leite, BA[8;] Vanessa L. Malcarne, PhD[5,9], Brett D. Thombs, PhD[4,6,10-13]; and the SPIN Investigators[14]

[1]Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York, USA

[2]Center for Practice and Research at the Intersection of Information, Society, and Methodology, New York University, New York, New York, USA

[3]Behavioural Science Institute, Clinical Psychology, Radboud University, Nijmegen, the Netherlands

[4]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

[5]San Diego State University/ University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California, USA

[6]Department of Psychiatry, McGill University, Montreal, Quebec, Canada

[7]Scleroderma Foundation, Los Angeles, California, USA

[8]School of Psychology, University of Minho, Braga, Portugal

[9]Department of Psychology, San Diego State University, San Diego, California, USA

[10]Department of Medicine, McGill University, Montreal, Quebec, Canada

[11]Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

[12]Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada

[13]Department of Psychology, McGill University, Montreal, Quebec, Canada

[14]SPIN Investigators: Murray Baron, McGill University, Montreal, Quebec, Canada; Susan J. Bartlett, McGill University, Montreal, Quebec, Canada, Daniel E. Furst, Division of Rheumatology, Geffen School of Medicine, University of California, Los Angeles, California, USA; Maureen D. Mayes, University of Texas McGovern School of Medicine, Houston, Texas, USA; Luc Mouthon, Université Paris Descartes, Paris, France; Warren R. Nielson, St. Joseph's Health Care, London, Ontario, Canada; Robert Riggs, Scleroderma Foundation, Danvers, Massachusetts, USA; Maureen Sauve, Scleroderma Society of Ontario, Hamilton, Ontario; Fredrick Wigley, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Shervin Assassi, University of Texas McGovern School of Medicine, Houston, Texas, USA; Andrea Benedetti, McGill University, Montreal, Quebec, Canada; Isabelle Boutron, Université Paris Descartes, and Assistance Publique - Hôpitaux de Paris, Paris, France; Angela Costa Maia, University of Minho, Braga, Portugal; Lindsay Cronin, Scleroderma Foundation, Western Pennsylvania Chapter, Pittsburgh, Pennsylvania, USA; Ghassan El-Baalbaki, Université du Québec à Montréal, Montreal, Quebec, Canada; Carolyn Ells, McGill University, Montreal, Quebec, Canada; Stephen Elrod, Scleroderma Foundation, Southern California Chapter, Los Angeles, California, USA; Cornelia van den Ende, Sint Maartenskliniek, Nijmegen, The Netherlands; Kim Fligelstone, Scleroderma & Raynaud's UK, London, UK; Catherine Fortune, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Tracy Frech, University of Utah, Salt Lake City, Utah, USA; Amy Gietzen, Scleroderma Foundation, Tri-State Chapter, Binghamton, New York, USA; Dominique Godard, Association des Sclérodermiques de France,

Sorel-Moussel, France; Geneviève Guillot, Sclérodermie Québec, Montreal, Quebec, Canada; Shirley Haslam, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Monique Hinchcliff, Yale School of Medicine, New Haven, Connecticut, USA; Marie Hudson, McGill University, Montreal, Quebec, Canada; Ann Impens, Midwestern University, Downers Grove, Illinois, USA; Yeona Jang, McGill University, Montreal, Quebec, 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; Annett Körner, McGill University, Montreal, Quebec, Canada; Maggie Larche, McMaster University, Hamilton, Ontario, Canada; Carlo Marra, Memorial University, St. John's, Newfoundland, Canada; Christelle Nguyen, Université Paris Descartes, 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; Alexandra Portales, Asociación Española de Esclerodermia, Madrid, Spain; François Rannou, Université Paris Descartes, and Assistance Publique - Hôpitaux de Paris, Paris, France; Michelle Richard, Scleroderma Society of Nova Scotia, Halifax, Nova Scotia, Canada; Tatiana Sofia Rodriguez Reyna, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; Ken Rozee, Scleroderma Society of Nova Scotia, Halifax, Nova Scotia, Canada; Anne A. Schouffoer, Leiden University Medical Center, Leiden, The Netherlands; Russell J. Steele, Jewish General Hospital and McGill University, Montreal, Quebec, Canada; Nancy Stephens, Scleroderma Foundation, Michigan Chapter, Southflied, Michigan, USA; Maria E. Suarez-Almazor, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; Joep Welling, NVLE Dutch patient organization for systemic autoimmune diseases, Utrecht, The Netherlands, and Federation of European

Scleroderma Associations, Brussel, Belgium; Durhane Wong-Rieger, Canadian Organization for Rare Disorders, Toronto, Ontario, Canada; Christian Agard, Centre Hospitalier Universitaire - Hôtel-Dieu de Nantes, Nantes, France; Alexandra Albert, Université Laval, Quebec, Quebec, Canada; Marc André, Centre Hospitalier Universitaire Gabriel-Montpied, Clermont-Ferrand, France; Guylaine Arsenault, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Ilham Benzidia, Assistance Publique - Hôpitaux de Paris, Hôpital St-Louis, Paris, 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; Gilles Boire, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Alessandra Bruns, Université de Sherbrooke, Sherbrooke, Quebec, Canada; 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; Pascal Cohen, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Chase Correia, Northwestern University, Chicago, Illinois, USA; Pierre Dagenais, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Christopher Denton, Royal Free London Hospital, London, UK; Robyn Domsic, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; Sandrine Dubois, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; 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; Alexia Esquinca, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; 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; Anna Gill, Royal Free London Hospital, London, UK; Jessica Gordon, Hospital for Special Surgery, New York City, New York, USA; Brigitte Granel-Rey, Aix Marseille Université, and Assistance Publique - Hôpitaux de Marseille, Hôpital Nord, Marseille, France; Claire Grange, Centre Hospitalier Lyon Sud, Lyon, France; Genevieve Gyger, Jewish General Hospital and McGill University, Montreal, Quebec, Canada; Eric Hachulla, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Pierre-Yves Hatron, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Sabrina Hoa, CHUM - Centre hospitalier de l'Université de Montréal, Montreal, Quebec, Canada; Ariane L Herrick, University of Manchester, Salford Royal NHS Foundation Trust, Manchester, UK; Adrian Hij, Assistance Publique - Hôpitaux de Paris, Hôpital St-Louis, Paris, France; Alena Ikic, Université Laval, Quebec, Quebec, Canada; Niall Jones, University of Alberta, Edmonton, Alberta, Canada; Artur Jose de B. Fernandes, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Suzanne Kafaja, University of California, Los Angeles, California, USA; Nader Khalidi, McMaster University, Hamilton, Ontario, Canada; 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; Patrick Liang, Université de Sherbrooke, Sherbrooke, Quebec, Canada; 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, 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, Hôpitaux Privés de Metz, Hôpital Belle-Isle, Metz, France; Arsene Mekinian, Assistance Publique - Hôpitaux de Paris, Hôpital St-Antoine, Paris, France; Sheila Melchor, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Mandana Nikpour, St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; 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 Rodriguez, 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; Doug Smith, University of Ottawa, Ottawa, Ontario, Canada; 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; Wendy Stevens, St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; Evelyn Sutton, Dalhousie University, Halifax, Nova Scotia, Canada; Benjamin Terrier, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; 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; Angelica Bourgeault, Jewish General Hospital, Montreal, Quebec, Canada;

Andrea Carboni Jiménez, Jewish General Hospital, Montreal, Quebec, Canada; Sami Harb, Jewish General Hospital, Montreal, Quebec, Canada; Mengxin Tang, Jewish General Hospital, Montreal, Quebec, Canada; Lydia Tao, Jewish General Hospital, Montreal, Quebec, Canada.

Address for Correspondence: Brett D. Thombs, PhD; Jewish General Hospital; 4333 Cote Ste Catherine Road; Montréal, Québec H3T 1E4; Telephone (514) 340-8222 ext. 25112; E-mail: brett.thombs@mcgill.ca.

Word Count: 3,390

ABSTRACT

Objectives: The Social Appearance Anxiety Scale (SAAS) is a 16-item questionnaire developed to evaluate fear of appearance-based evaluation by others. The primary objective of this research was to investigate the existence of differential item functioning (DIF) for the 16 SAAS items, comparing patients who completed the SAAS in English and French, either to confirm that scores are comparable or provide guidance on calculating comparable scores. A secondary research objective was to investigate the existence of DIF based on sex and disease status. A tertiary research objective was to assess DIF related to language, sex, and disease status on the recently developed SAAS-5.

Design: This was a cross-sectional analysis using baseline data from patients enrolled in the Scleroderma Patient-centered Intervention Network (SPIN).

Setting: SPIN patients included in the present study were enrolled at 43 centers in Canada, USA, UK, France, and Australia, with questionnaires completed in April 2014 through July 2019.

Participants: 1640 SPIN patients completed the SAAS in French (N=600) or English (N=1040).

Primary and secondary measures: The SAAS was collected along with demographic and disease characteristics.

Results: Six items were identified with statistically significant language-based DIF, four with sex-based DIF, and one with disease type-based DIF. However, GPCM factor scores before and after accounting for DIF were similar (Pearson correlation > .99), and individual score differences were small. This was true for both the full and shortened versions of the SAAS.

Conclusion: SAAS and SAAS-5 scores are comparable across language, sex, and disease-type, despite small differences in how patients respond to some items.

Keywords: Patient Reported Outcome Measure; Differential Item Functioning; Generalized

Partial Credit Model; Systemic Sclerosis

STRENGHTS AND LIMITATIONS OF THIS STUDY

- This study showed that SAAS and SAAS-5 scores are comparable across language (English and French), sex, and disease-type groups
- This indicates that responses can be combined in future analyses.
- These findings are only generalizable to adults with scleroderma and should be confirmed for other populations.

Introduction

A desire to improve the patient-centered focus of healthcare research has led to the development and increased use of patient-reported outcome (PRO) measures aimed at a wide range of human experiences, including patient-perceived health, well-being and psychological status [1]. This is particularly important in chronic diseases that lead to symptoms that are not directly measurable [2]. Many PRO measures have been translated into multiple languages, which is relevant in treatment centers where more than one language is common, as well as in rare disease research, which often involves collaboration and communication across sites in multiple countries [3]. In these situations, outcomes measured in more than one language are commonly combined in analyses.

In order to compare PROs across language and cultural groups, it is important to ensure that all patients interpret and respond to the questionnaire items in equivalent ways, and not based on idiosyncratic differences due to differing cultural norms, systematic differences in interpretation, or indirect translations of some items [4]. If this is not the case, then items or questions are said to have differential item functioning (DIF). When DIF is present, patients with equal underlying levels of the construct, or latent trait, measured by that scale will respond differently to the same item [5].

Systemic sclerosis (SSc) is a rare, multi-system autoimmune disorder with heterogeneous symptomatology characterized by microvascular damage and fibrosis in multiple organs [6 7]. Changes in appearance are common and can include telangiectasias, hypo- and hyper-pigmentation, loss of skin folds, loss of flexibility of the lips, digital ulcers, and hand contractures [6 8]. These changes in appearance are often in socially relevant areas of the body, such as the face and hands, and can lead to problems with social interactions and increased social appearance anxiety [9].

The Social Appearance Anxiety Scale (SAAS) is a 16-item scale, which aims to measure patients' fear of appearance-based evaluation [10]. Among people with SSc, the SAAS may be used for both individual-level treatment plans and larger scale research, evaluating the impact of potential interventions. The Scleroderma Patient-centered Intervention Network (SPIN) Cohort is a web-based, international cohort designed to collect PROs at regular intervals and as a framework to conduct trials of psychosocial and rehabilitation interventions for patients with SSc [11]. Depending on their native language, participants enrolled in SPIN may complete the SAAS in French, English, or Spanish; however, no research has yet confirmed that SAAS scores are comparable across these language groups.

A recent study developed a shortened version of the SAAS consisting of five items (SAAS-5) for use in patients with SSc [12]. The use of shortened versions, such as the SAAS 5, has the potential to decrease patient burden and increase data quality [13]. However, it is of interest to determine whether the shortened version exhibits DIF.

Therefore, the primary purpose of this analysis is to investigate the comparability of responses to versions of the SAAS administered in different languages. As a secondary research objective, comparability of SAAS scores with respect to disease type and sex were also assessed. A tertiary research objective was to assess the comparability of SAAS scores on the 5-item shortened version.

Materials and Methods

Patients and procedures

The sample consisted of patients enrolled in the SPIN Cohort with complete data study questionnaires from initial enrollment sessions between April 2014 through July 2019. Participants in the SPIN Cohort were enrolled at 43 centers in Canada, USA, UK, France, and Australia. To be eligible for the SPIN Cohort, participants must be classified as having SSc according to the 2013 ACR/EULAR classification criteria [14], confirmed by a SPIN physician, be at least 18 years of age, have the ability to give informed consent, and be fluent in English, French or Spanish. However, the present study only included patients who completed study questionnaires in English or French, as the sample size of Spanish patients was too small to be included at the time of the analyses. Exclusion criteria for participation in the SPIN Cohort include not having access to the internet or otherwise not being able to respond to questionnaires via the internet. The SPIN sample is a convenience sample. Eligible participants are invited by the attending physician or a supervised nurse coordinator to participate in the SPIN Cohort, and written informed consent is obtained. The local SPIN physician or supervised nurse coordinator then completes a medical data form that is submitted online to initiate participants registration in the SPIN Cohort. After completion of online registration, an automated welcoming email is sent to participants with instructions to on how to activate their SPIN online account and how to complete the SPIN Cohort patient measures online. SPIN Cohort participants complete outcome measures via the internet upon enrollment and subsequently every 3 months. The SPIN Cohort study was approved by the Research Ethics Committee of the Jewish General Hospital, Montreal, Canada (MP-05-2013-150, 12-123) and by the Institutional Reviews Boards of each participating center.

Measures

Demographics and disease characteristics. Demographic and disease variables included age, sex, race/ethnicity, marital status, education level, time since diagnosis, and SSc subtype. Disease subtypes were classified as limited or diffuse. Limited disease was defined as skin involvement distal to the knees and elbows only, whereas diffuse disease included more extensive skin involvement [15]. The country of patient recruitment and language of assessment were also recorded.

Social appearance anxiety scale (SAAS). The SAAS consists of 16 items assessing patients' self-reported anxiety about appearance-based evaluation. The SAAS was initially validated among three samples of undergraduate students (n=512, 853, and 541, respectively) [10]. In this population, the SAAS was shown to have unifactorial structure, high internal consistency, high test-retest reliability, and was positively correlated with other social anxiety measures [10]. A recent study of 938 participants enrolled in the SPIN Cohort demonstrated that the SAAS is a unidimensional, reliable, and valid measurement of social appearance anxiety among people with SSc [16]. The SAAS was initially written in English. The French version used in this study was translated by SPIN investigators using the forward-backward method [17]. For both versions, item responses are recorded on a five point scale (1=Not at all, 5=Extremely). Item 1 ("I feel comfortable with the way I appear to others") is reverse coded before summing across items to produce a total score ranging from 16 to 80. Higher scores indicate higher levels of appearance anxiety.

The SAAS-5, consisting of items 6, 7, 12, 13, and 14 from the original SAAS, was recently developed and validated for use in patients with SSc [12]. Scores on the SAAS-5 range from 5 to 25, with higher scores indicating higher levels of appearance anxiety.

Statistical analysis

The English- and French-speaking samples were compared based on demographic and disease characteristics to identify possible differences between the two language groups.

A generalized partial credit model (GPCM) was then used to model the latent factor (social anxiety with appearance) underlying the SAAS. For each item, a GPCM was used to estimate two types of item-level parameters: 1) thresholds (betas) for the level of the latent factor (theta) at which respondents are more likely to endorse a given response category instead of the category below, and 2) a discrimination parameter (alpha) that measures the strength of the relationship between that item and the underlying latent factor [18].

Item-characteristic curves (ICCs) are often used to visualize these parameters, and Figure 1 shows three examples of ICCs for a hypothetical 5-category item. Each curve in an ICC plot corresponds to a possible categorical response. Along the latent spectrum, the height of each curve indicates the estimated probability that a respondent with a particular level of the latent factor will respond with the corresponding category. Item level thresholds are visualized as the intersections between consecutive curves; discrimination parameters are visualized as the peaked-ness of the curves. When item-level thresholds vary across observed groups, items are said to display uniform DIF. Uniform DIF could be visualized as a horizontal shift of ICC curves for one demographic group compared to the other. Meanwhile, when the discrimination parameter varies across observed groups, items are said to display non-uniform DIF. Non-uniform DIF could be visualized as a change in the peaked-ness of the curves for one demographic group compared to the other.

The *lordif* package in R [19 20] was used to identify items with language-based DIF through an iterative procedure. The algorithm implemented by *lordif* iteratively fits three ordinal logistic models for each item and uses these models to flag items with potential DIF. The first model predicts the probability of each response category using estimated latent factor scores alone, while the second and third models test for uniform and non-uniform DIF, respectively. Once a set of items is flagged, the algorithm then re-estimates latent factor scores using another GPCM that accounts for DIF on those items. DIF is accounted for by allowing item level parameters to vary across groups. The process stops once the same items are repeatedly flagged for DIF [20].

During the iterative search for items with DIF, items were flagged using a chi-square test comparing the first and third models (alpha = 0.01 significance level). Flagged items were then re-examined to distinguish between uniform and non-uniform DIF. This was done by separately comparing the first and second models (to ascertain uniform DIF) and second and third models (to ascertain non-uniform DIF), again using a chi-square test (alpha = 0.01 significance level).

Items with DIF were further investigated by comparing item-level parameters from a GPCM for patients who completed the SAAS in English and French. To visualize and understand differences among the two groups on each item, item true score functions for English- and French-speakers were compared, which show expected responses for items with DIF as a function of estimated latent social appearance anxiety scores accounting for DIF.

The questionnaire-level impact of DIF on estimated latent factor scores was assessed by plotting test characteristic curves, which show expected summed scores on the SAAS as a function of patients' GPCM scores accounting for DIF. As per previous guidelines, impact was numerically assessed by comparing initial scores (not accounting for DIF) to final scores (accounting for DIF), using the Pearson correlation of the two scores and by comparing individual score differences to the standard errors of initial scores [21 22]. To assess whether the correlation significantly differed from 1, a randomization null distribution and p-values were obtained by randomly permuting group labels 1000 times and re-estimating scores and statistics holding the measurement model fixed across permutations, but re-estimating the item-level parameters based on the permuted dataset.

Lastly, the median and range of score differences (of the difference between scores accounting for and not accounting for DIF) were also calculated, and score differences were plotted against initial scores to find areas of the latent spectrum with highest DIF impact. Before comparison, scores were placed on the same scale using a transformation by Stocking and Lord [23]. This was also done using the *lordif* package, which equates final scores accounting for DIF to initial GPCM estimates using the non-DIF items as anchors.

The same process was repeated to identify and investigate DIF related to sex and disease status, respectively, and additionally for the SAAS-5.

Patient Involvement

SPIN was conceived by a collaboration of investigators and patients. SPIN's Patient Advisory Board advises the SPIN Steering Committee on priorities for investigation. Patients were included in the SPIN Publication Committee, which reviewed the proposal for the present study and its methods. Two patients were co-authors of the present report.

Results

Table 1 shows demographic information and disease characteristics for all patients and by assessment language.

Table 1. Demographic and disease characteristics by assessment language

Variable	All Patients (N = 1640)	English-Speaking Patients (N = 1040)	French-Speaking Patients (N = 600)
Mean age, years (SD)[a]	55.1 (12.5)	55.7 (11.7)	54.0 (13.8)
Female (%)	87.2	87.6	86.5
Mean SAAS summed score (SD)	29.1 (13.7)	28.3 (13.2)	30.5 (14.5)
Diffuse disease type (%)	39.0	42.4	33.2
Mean time since diagnosis, years (SD)	9.2 (7.9)	9.7 (8.0)	8.5 (7.6)
Married or common law (%)	71.2	73.3	67.5
At least 12 years of education (%)	85.7	94.2	70.8
Race[b]			
White (%)	83.6	83.9	83.0
Black (%)	7.1	6.1	8.8
Other (%)	9.3	10.0	8.2
Country of patient recruitment			
Canada (%)	24.9	28.7	18.5
USA (%)	35.5	55.9	0.2
UK (%)	9.7	15.3	0.0
France (%)	29.8	0.1	81.3
Australia (%)	0.1	0.1	0.0

Due to missing values: a) N=1036 for the English cohort, b) N=1038 for the English cohort

The English and French samples included 1040 and 600 patients, respectively. The mean age was 55.7 years (SD = 11.8) for the English sample and 54.0 years (SD = 13.8) for the French sample. In both groups, the majority of patients identified as female (87.6% and 86.5% in the English and French samples, respectively), and mean time since diagnosis was about 9 years (9.7 years [SD = 8.0] and 8.5 years [SD = 7.6] for the English and French samples, respectively). Mean summed scores on the SAAS were also similar across groups: 28.3 (SD = 13.2) for the English sample and 30.5 (SD = 14.5) for the French sample. Fewer than half of patients had diffuse disease status (42.4% and 33.2% in the English and French samples, respectively). The majority of patients were married (73.3% and 67.5% in the English and French samples, respectively), had at least 12 years of education (94.2% and 70.8% in the English and French samples, respectively), and identified as white (83.9% and 83.0% in the English and French samples, respectively). About half (56%) of the English sample was recruited in the United States and 29% were recruited in Canada. Meanwhile, most of the French sample was recruited in France (81.3%).

DIF Analysis

Six of the 16 SAAS items (Table 2) were identified as having statistically significant ($p < .01$) language-based DIF: items 2, 5, 8, 11, 12, and 13. Only item 11 was identified as having non-uniform DIF. Item true score functions for these six items are shown in Figure 2. For most items with uniform DIF, French speakers' expected item level responses were slightly higher than their English-speaking counterparts with equal levels of appearance anxiety. This pattern is reversed for item 2.

Table 2. SAAS Items

Item #	Item Text
1	I feel comfortable with the way I appear to others
2	I feel nervous when having my picture taken
3	I get tense when it is obvious people are looking at me
4	I am concerned people won't like me because of the way I look
5	I worry that others talk about flaws in my appearance when I am not around
6	I am concerned people will find me unappealing because of my appearance
7	I am afraid people find me unattractive
8	I worry that my appearance will make life more difficult for me
9	I am concerned that I have missed out on opportunities because of my appearance
10	I get nervous when talking to people because of the way I look
11	I feel anxious when other people say something about my appearance
12	I am frequently afraid that I won't meet others' standards of how I should look
13	I worry people will judge the way I look negatively
14	I am uncomfortable when I think others are noticing flaws in my appearance
15	I worry that a romantic partner will/would leave me because of my appearance
16	I am concerned that people think I am not good looking

Test characteristic curves for the English and French cohorts are plotted in Figure 3, while Figure 4 shows score differences based on GPCMs that do and do not account for DIF. At the questionnaire level, French speakers are expected to have slightly larger summed scores on the SAAS as compared to English speakers with the same level of appearance anxiety. The correlation between the two sets of GPCM scores was 0.99977 (95% confidence interval: [0.99975, 0.99979], $p < 0.001$). At the individual level, the median GPCM score difference (scores accounting for DIF minus scores that do not account for DIF) was 0.0049, and differences in factor scores ranged from -0.078 to 0.065. No individual score differences exceeded the standard errors of initial estimates. Patients with the largest score differences had initial GPCM scores around -0.5 and 1.0, whereas individuals whose initial estimated anxiety level was extreme (low or high) or average had smaller DIF impact.

Four items were identified as having sex-based DIF (all uniform): items 2, 4, 9, and 14. Only item 2 exhibited both language and sex based DIF. Item true score functions suggest that females tend to give slightly higher categorical responses than equally anxious males on items 2 and 14 and lower responses on items 4 and 9. Meanwhile, the test characteristic curves for males and females were practically indistinguishable, suggesting that equally anxious males and females have almost identical expected summed scores. The correlation between the two sets of GPCM scores was 0.99985 (95% confidence interval: [0.99983, 0.99986], $p=0.003$). At the individual level, the median score difference based on a GPCM was 0.0020, and differences in factor scores ranged from -0.047 to 0.135. No individual score differences exceeded the standard errors of initial estimates. The largest score differences were observed for individuals whose initial GPCM score was low (around -1.0 in this dataset); individuals with average or high estimated anxiety levels had comparatively low DIF impact.

Only one item (item 9) was identified as having DIF related to disease-type (non-uniform). This item was also identified as having sex-based DIF, but not language-based DIF. Among patients with low appearance anxiety, those with limited disease are expected to give smaller categorical responses to item 9 than patients with diffuse disease and equal levels of appearance anxiety; this pattern is reversed at the higher end of the latent spectrum. At the questionnaire level, expected summed scores were nearly identical across disease-type groups. The correlation between the two sets of GPCM scores was 0.99996 (95% confidence interval: [0.99996, 0.99997], $p < 0.001$). At the individual level, the median GPCM score difference was 0.001 and these differences in factor scores ranged from -0.101 to 0.080. No individual score differences exceeded the standard errors of initial estimates. The largest score differences were observed for individuals whose initial GPCM estimate was around 0, or slightly below.

For the SAAS-5, only item 12 was flagged for language based DIF, while item 14 was still flagged for gender-based DIF. In both cases, the correlation between factor scores was still high: 0.99995 for language-based DIF (95% confidence interval: [0.99995, 0.99996], $p = 0.107$) and 0.99971 for gender-based DIF (95% confidence interval: [0.99969, 0.99974], $p = 0.018$).

Discussion

This study investigated whether the SAAS displays DIF across language, sex, and disease subtype groups among people with SSc. Nine items were flagged for language-based DIF (8 uniform, 1 non-uniform), four were flagged with sex-based DIF (all uniform), and only one was flagged with disease-type based DIF (non-uniform). In reviewing translations of the items flagged with language-based DIF, we did not observe any clear differences. Similarly negligible levels of DIF were found for the SAAS-5.

For all three analyses on the full-length SAAS, the high (>0.99) Pearson correlations between the two GPCM estimates imply that accounting for DIF does not provide much additional information about respondents' comparative levels of social appearance anxiety. The near-zero (<0.01) associated p-values nonetheless suggest that observed correlations are lower than what would be expected by random chance in a no-DIF null condition under identical measurement models. While previous analyses have used Pearson correlations [21 22] to compare GPCM scores that do and do not account for DIF, other research has cautioned against this [24]. Our findings imply that very large correlations between initial and final GPCM estimates may still be smaller than simulated values under a no-DIF condition. Thus, we caution that correlations alone may not be particularly interpretable as a measure of DIF impact.

The relatively small ranges of GPCM score differences in all three analyses nonetheless support the conclusion that accounting for DIF has limited impact on individual estimated scores. In all cases, no individual differences exceeded initial standard errors. Thus, estimated scores accounting for DIF were all within the range of inherent uncertainty in naïve GPCM estimates. The median score difference was largest for language-based DIF and smallest for disease-type-based DIF; however, the range of score differences was smallest in the language-based analysis, due to the existence of a few outliers in the other two cases.

Scatter plots of GPCM score differences as a function of naïve GPCM estimates (see Figure 4 for language-based DIF) show that language-, sex-, and disease-type-based DIF impact is not constant across the latent spectrum. Naïve GPCM estimates near values where GPCM score differences are larger (i.e., near -0.5 and 1 for language-based DIF, -1 for sex-based DIF, and 0 for disease-type-based DIF) may therefore be slightly less certain.

While DIF impact was found to be small for both simple summed scores and naïve GPCM estimates, it is important to note that the choice between these two scoring methods is also relevant [25 26]. This paper explored three different methods for estimating social appearance anxiety levels based on responses to the SAAS: simple summed scores, naïve GPCM factor scores, and GPCM factor scores accounting for DIF. Our analysis aimed to assess comparability of scores across demographic groups, and therefore mainly focused comparison between the two sets of GPCM factor scores; however, much more confidence in individual scores is gained in the jump from simple summed scores to a GPCM factor score, than in the jump from a naïve GPCM factor score to a GPCM factor score accounting for DIF. For example, in this dataset, individuals with the same summed score had naïve GPCM estimates of social appearance anxiety differing by up to 0.92 standardized units. Thus, regardless of whether DIF is accounted for in score calculations, a GPCM-based score or weighted summed score would be preferable over a simple summed score.

Conclusion

In conclusion, this study used an iterative algorithm implemented via the *lordif* package in R to flag items on the SAAS for DIF related to language of test administration, sex, and disease type. After flagging items with DIF, impact was assessed primarily by looking at GPCM score correlations and differences before and after accounting for DIF. While at least one item was flagged for DIF in each analysis, DIF impact was assessed to be small, supporting the conclusion that GPCM scores are comparable across groups produced by these three demographic variables.

FIGURE CAPTIONS

Figure 1. Three possible ICC curves for a 5 category item. The left and middle panels show ICCs for items with the same approximate discrimination parameters (alphas) but different item-level thresholds (betas). The left and right panels show ICCs for items with the same approximate item-level thresholds (betas) but different discrimination parameters (alphas).

Figure 2. Item True Score Functions for the six items identified as having language-based DIF. For items 5, 8, 12, and 13, these plots demonstrate that French speakers are expected to give larger categorical responses than English speakers with equal levels of appearance anxiety. This trend is reversed for item 2, while item 11 demonstrates non-uniform DIF (i.e., the true score functions for English and French speakers cross each other).

Figure 3. Test characteristic curve showing expected summed scores on the SAAS as a function of estimated social appearance anxiety accounting for DIF. Thus, among French and English speakers with the same estimated level of social appearance anxiety, French speakers are expected to have slightly larger summed scores.

Figure 4. The top plot shows GPCM score differences at the questionnaire level (accounting for DIF – not accounting for DIF) compared to factor scores accounting for DIF. The largest score differences occur at estimated appearance anxiety levels .5 standard deviations (sd) below average and 1 sd above average. The figure on the bottom left shows a box plot of these score differences among all respondents. The figure on the bottom right shows these differences by language. Overall differences are small and are mostly negative for English speakers and positive

for French speakers, suggesting that pooled scores from a GPCM will tend to overestimate appearance anxiety for French speakers and underestimate it for English speakers.

AUTHOR CONTRIBUTIONS

DH and BDT were responsible for the study conception. LK, M-EC, VLM, BDT and the SPIN Investigators contributed to data collection. SJS, DH, LK, M-EC, SG, KG, CL, VLM, BDT contributed to data analysis and interpretation. SJS and DH drafted the manuscript. All authors provided a critical revision of the manuscript and approved the final version of the manuscript. DH is the guarantor.

FUNDING STATEMENT

SPIN has been funded by grants from the Canadian Institutes of Health Research (TR3-119192, PJT-148504, PJT-149073, SCT-162963), the Canadian Initiative for Outcomes in Rheumatology Care, and the Arthritis Society. In addition, SPIN has received institutional contributions from the Lady Davis Institute for Medical Research of the Jewish General Hospital, Montreal, Canada and from McGill University, Montreal, Canada. SPIN has also received support from Scleroderma Canada, the Scleroderma Society of Ontario, Sclérodermie Québec, the Scleroderma Society of Nova Scotia, the Scleroderma Association of British Columbia, Scleroderma Manitoba, and the Scleroderma Society of Saskatchewan.

COMPETING INTERESTS STATEMENT

The authors have read and understood the BMJ policy on declaration of interests and declare that they have no competing interests.

DATA SHARING STATEMENT

Data from the SPIN Cohort can be requested from the corresponding author.

References

1. Weldring T, Smith SMS. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health Serv Insights* 2013;6:61-68. doi: 10.4137/HSI.S11093

2. Thombs BD, van Lankveld W, Bassel M, et al. Psychological health and well-being in systemic sclerosis: State of the science and consensus research agenda. *Arthritis Care & Research* 2010;62(8):1181-89. doi: 10.1002/acr.20187

3. Kwakkenbos L, Willems LM, Baron M, et al. The Comparability of English, French and Dutch Scores on the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F): An Assessment of Differential Item Functioning in Patients with Systemic Sclerosis. *PLOS ONE* 2014;9(3):e91979. doi: 10.1371/journal.pone.0091979

4. Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Med Care* 2006;44(11 Suppl 3):S39-49. doi: 10.1097/01.mlr.0000245452.48613.45

5. Holland PW, Wainer H. Differential Item Functioning: Taylor & Francis 2012.

6. Allanore Y, Simms R, Distler O, et al. Systemic sclerosis. (2056-676X (Electronic))

7. Clements PJ, Furst DE. Systemic Sclerosis. 2 ed. Baltimore: Lippincott Williams & Wilkins 2003.

8. Gholizadeh S, Fox RS, Mills SD, et al. Coping with the disfigurement of scleroderma: Facial, skin, and hand changes. Scleroderma. 2 ed: Springer International Publishing 2017:713-21.

9. Kwakkenbos L, Delisle VC, Fox RS, et al. Psychosocial Aspects of Scleroderma. *Rheum Dis Clin North Am* 2015;41(3):519-28. doi: 10.1016/j.rdc.2015.04.010

10. Hart TA, Flora DB, Palyo SA, et al. Development and Examination of the Social Appearance Anxiety Scale. *Assessment* 2008;15(1):48-59. doi: 10.1177/1073191107306673

11. Kwakkenbos L, Jewett Lr Fau - Baron M, Baron M Fau - Bartlett SJ, 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. LID - 10.1136/bmjopen-2013-003563 [doi] LID - e003563 [pii]. (2044-6055 (Electronic))

12. Harel D, Mills SD, Kwakkenbos L, et al. Shortening patient-reported outcome measures through optimal test assembly: application to the Social Appearance Anxiety Scale in the Scleroderma Patient-centered Intervention Network Cohort. *BMJ Open* 2019;9(2):e024010. doi: 10.1136/bmjopen-2018-024010

13. Harel D, Baron M. Methods for shortening patient-reported outcome measures. *Statistical Methods in Medical Research* 2018;28(10-11):2992-3011. doi: 10.1177/0962280218795187

14. van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Ann Rheum Dis* 2013;72(11):1747-55. doi: 10.1136/annrheumdis-2013-204424

15. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. *J Rheumatol* 1988;15(2):202-05.

16. Mills SD, Kwakkenbos L, Carrier M-E, et al. Validation of the Social Appearance Anxiety Scale in Patients With Systemic Sclerosis: A Scleroderma Patient-Centered Intervention

Network Cohort Study. *Arthritis Care & Research* 2018;70(10):1557-62. doi: 10.1002/acr.23514

17. Wild D, Grove A Fau - Martin M, Martin M Fau - Eremenco S, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. (1098-3015 (Print))

18. Muraki E. A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement* 1992;16(2):159-76. doi: 10.1177/014662169201600206

19. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2019.

20. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw* 2011;39(8):1-30.

21. Chung H, Kim J, Cook KF, et al. Testing measurement invariance of the patient-reported outcomes measurement information system pain behaviors score between the US general population sample and a sample of individuals with chronic pain. *Quality of Life Research* 2014;23(1):239-44. doi: 10.1007/s11136-013-0463-0

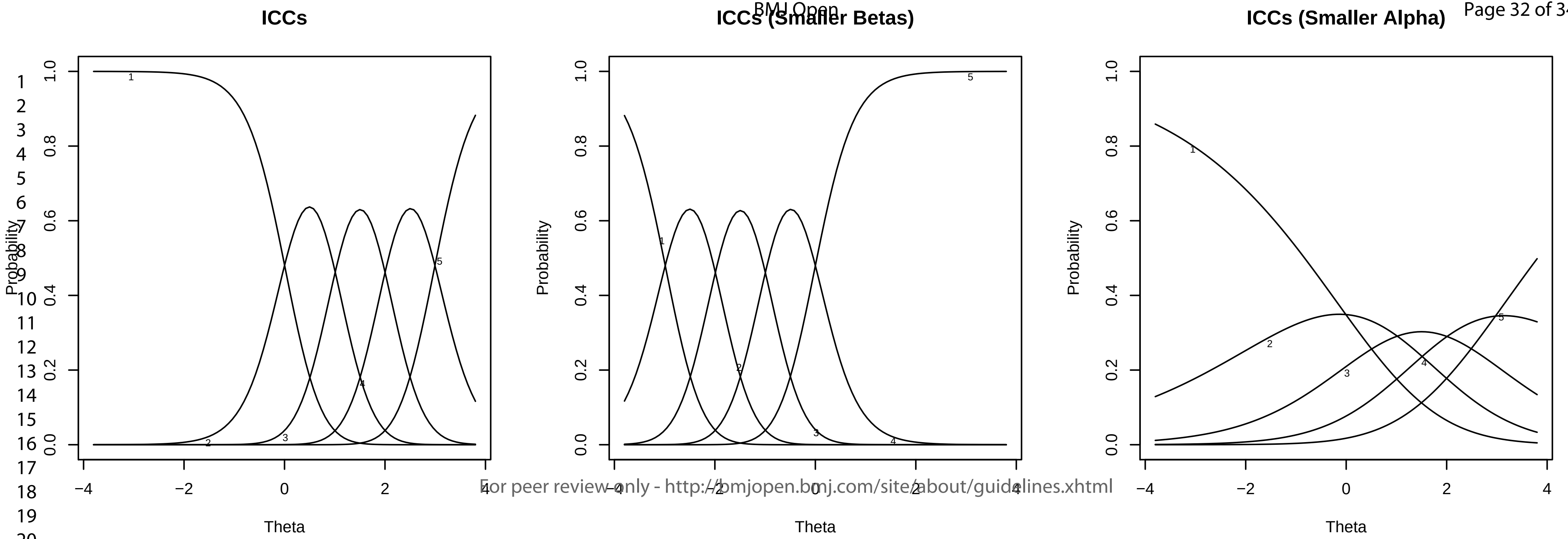
22. Cook KF, Bombardier CH, Bamer AM, et al. Do somatic and cognitive symptoms of traumatic brain injury confound depression screening? *Arch Phys Med Rehabil* 2011;92(5):818-23. doi: 10.1016/j.apmr.2010.12.008

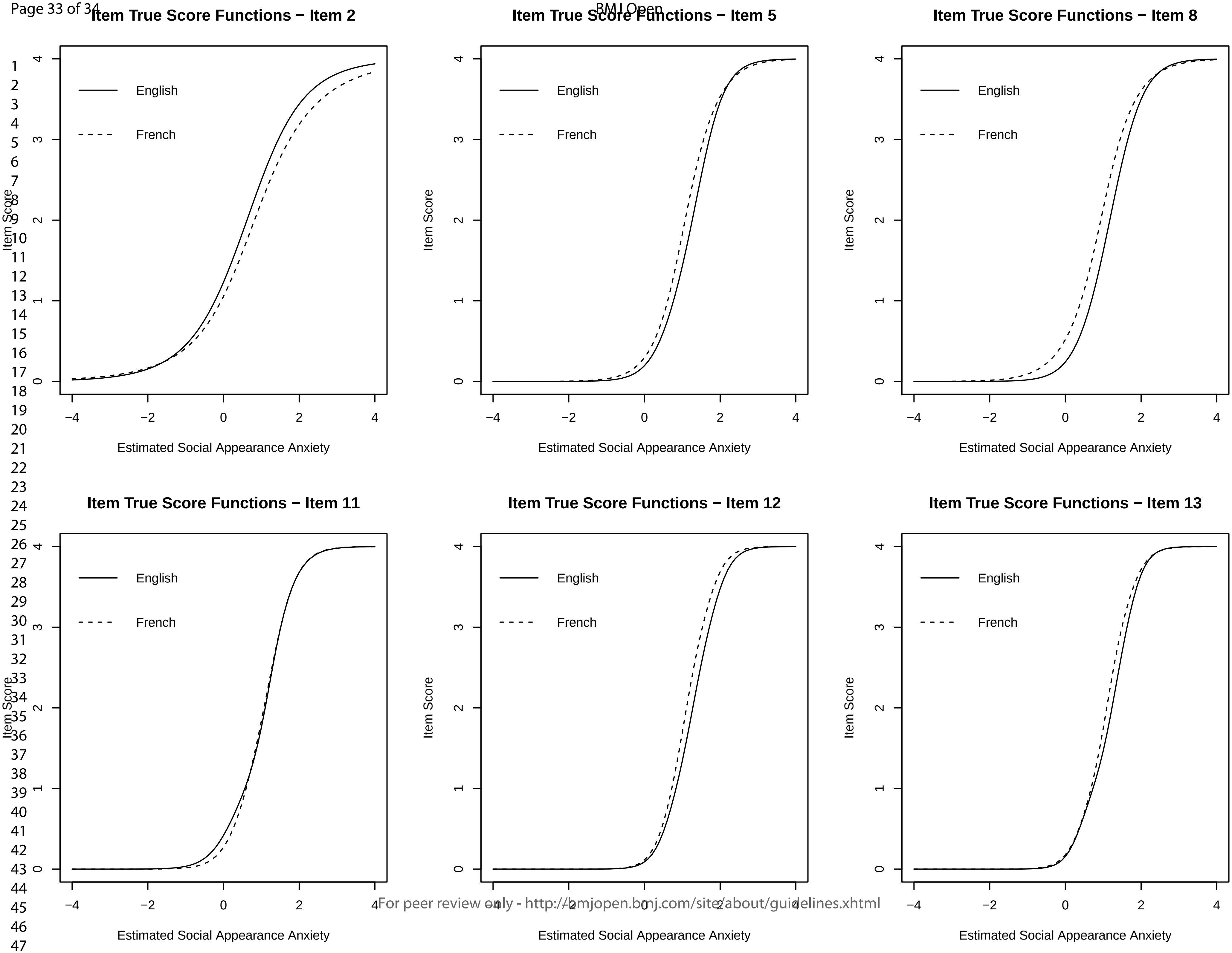
23. Stocking ML, Lord FM. Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement* 1983;7(2):201-10. doi: 10.1177/014662168300700208

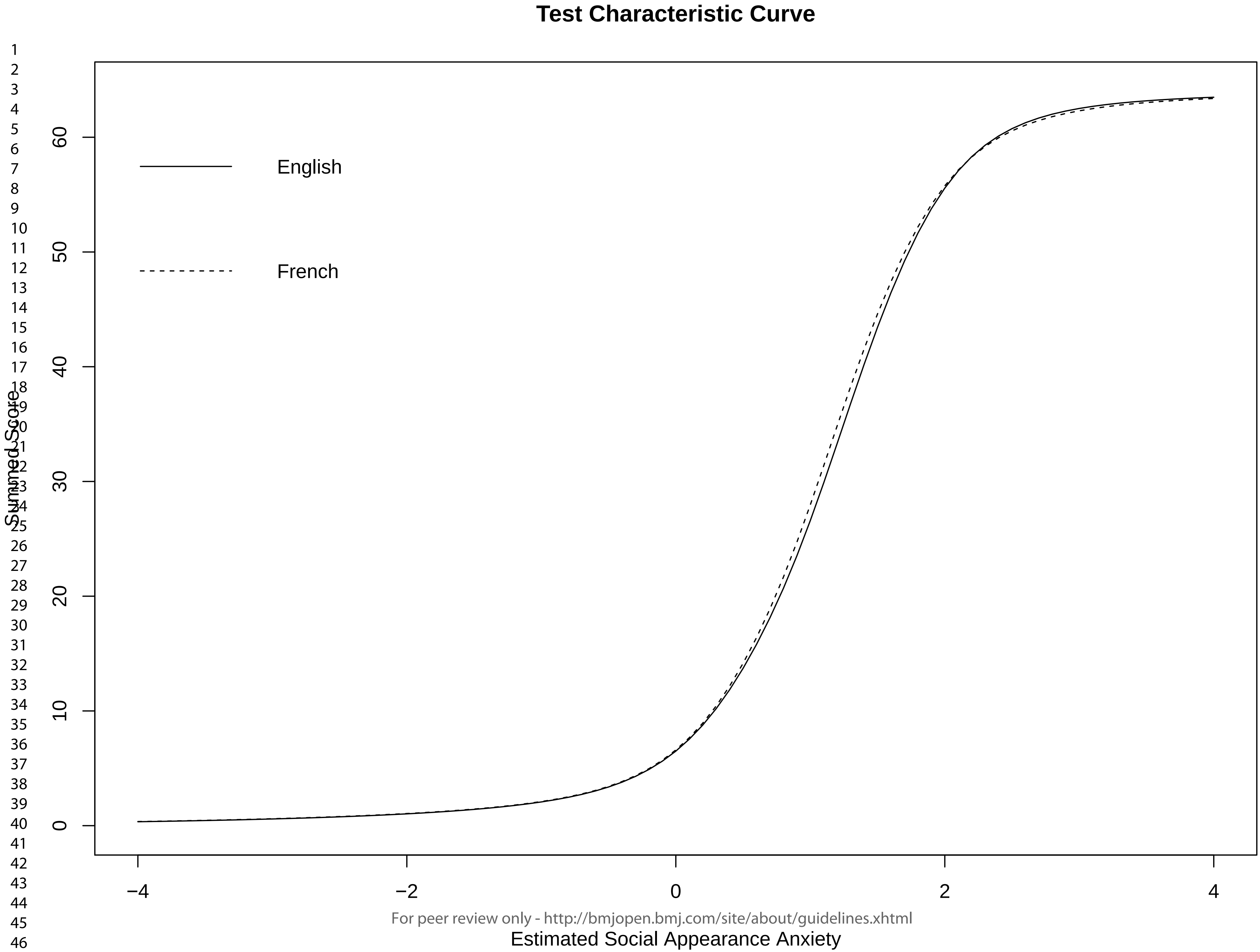
24. Rupp AA, Zumbo BD. A Note on How to Quantify and Report Whether Irt Parameter Invariance Holds: When Pearson Correlations are Not Enough. *Educational and Psychological Measurement* 2004;64(4):588-99. doi: 10.1177/0013164403261051

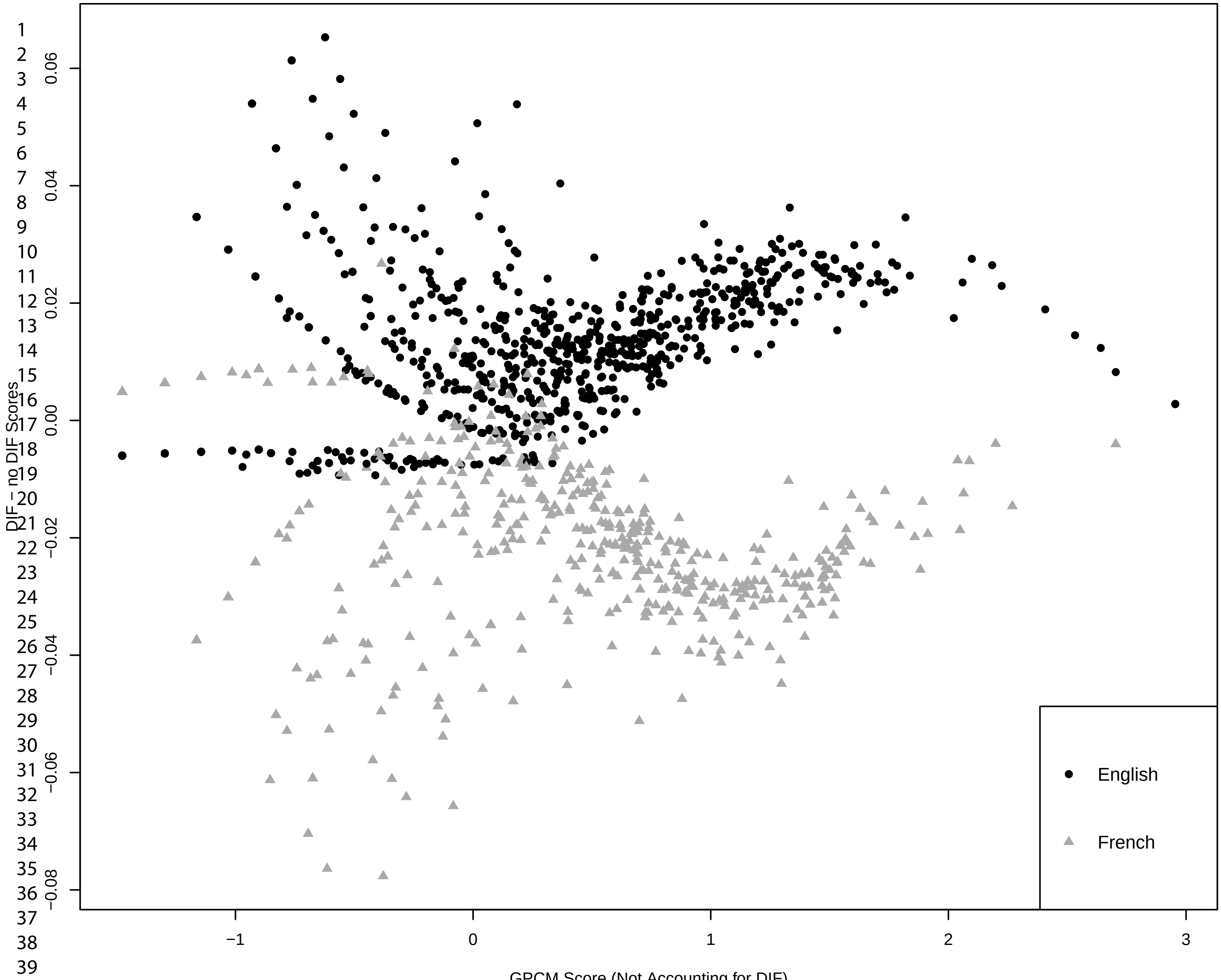
25. Harel D. The effect of model misspecification for polytomous logistic adjacent category item response theory models. McGill University Libraries, 2014.

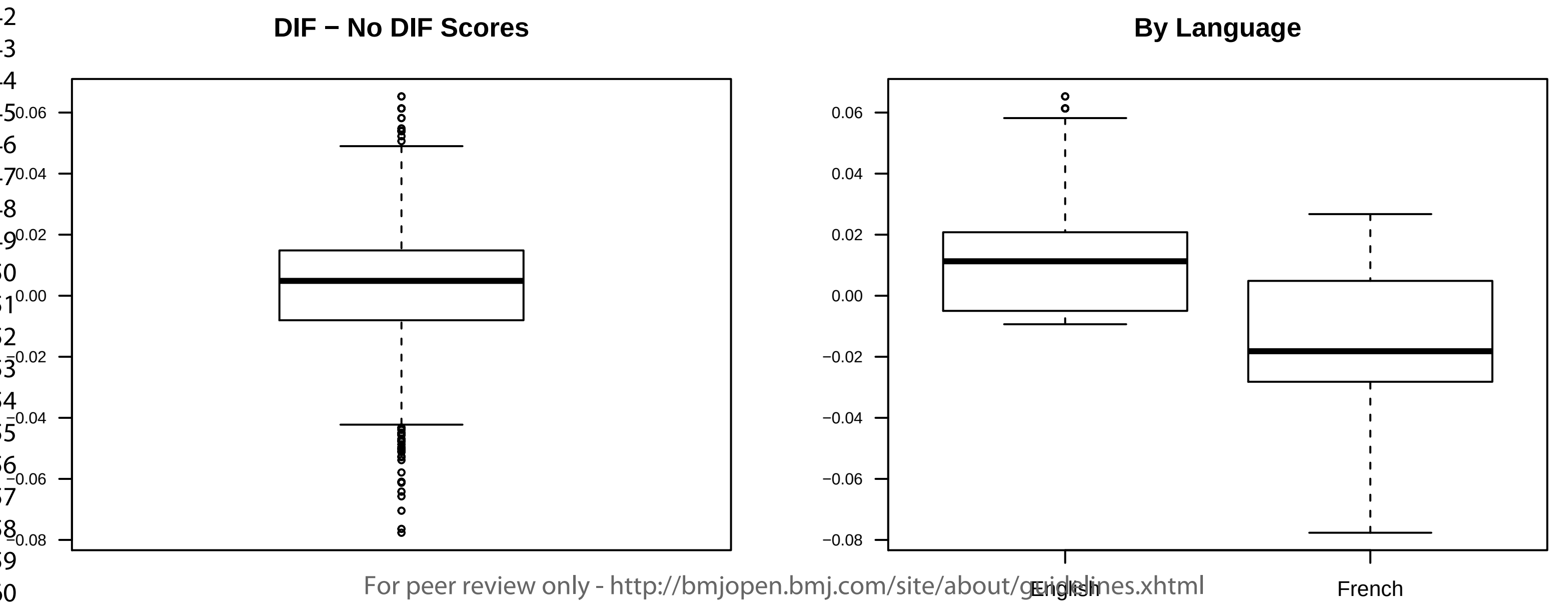
26. van der Ark LA. Stochastic Ordering Of the Latent Trait by the Sum Score Under Various Polytomous IRT Models. *Psychometrika* 2005;70(2):283-304. doi: 10.1007/s11336-000-0862-3


ICCs
ICCs (Smaller Betas)
ICCs (Smaller Alpha)
Probability
Theta
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4


Item True Score Functions − Item 2
Item True Score Functions − Item 5
Item True Score Functions − Item 8
Item True Score Functions − Item 11
Item True Score Functions − Item 12
Item True Score Functions − Item 13
English
French
Item Score
Estimated Social Appearance Anxiety
0
1
2
3
4
−4
−2
0
2
4


Test Characteristic Curve
English
French
Summed Score
0
10
20
30
40
50
60
−4
−2
0
2
4
Estimated Social Appearance Anxiety


DIF − no DIF Scores
GPCM Score (Not Accounting for DIF)
English
French
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−1
0
1
2
3


DIF − No DIF Scores
By Language
English
French
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08

Assessing Differential Item Functioning for the Social Appearance Anxiety Scale: A Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037639.R1
Article Type:	Original research
Date Submitted by the Author:	15-Jul-2020
Complete List of Authors:	Sommer, Sophia; New York University, Applied Statistics, Social Science, and Humanities; New York University, PRIISM Applied Statistics Center Harel, Daphna; New York University, Applied Statistics, Social Science, Humanities; New York University, PRIISM Applied Statistics Center Kwakkenbos, Linda; Radboud University, Behavioral Science Institute Carrier, Marie-Eve; Jewish General Hospital, Lady Davis Institute for Medical Research Gholizadeh, Shadi; San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology; McGill University, Psychiatry Gottesman, Karen; Scleroderma Foundation Leite, Catarina; University of Minho, School of Psychology Malcarne, Vanessa; San Diego State University, Psychology; San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology Thombs, Brett; McGill University, Psychiatry
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Rheumatology
Keywords:	STATISTICS & RESEARCH METHODS, RHEUMATOLOGY, Anxiety disorders < PSYCHIATRY

Assessing Differential Item Functioning for the Social Appearance Anxiety Scale: A Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study

Sophia J. Sommer, MSc[1,2]; Daphna Harel, PhD[1,2]; Linda Kwakkenbos, PhD[3]; Marie-Eve Carrier, MSc[4]; Shadi Gholizadeh, PhD[5,6]; Karen Gottesman, BA[7]; Catarina Leite, BA[8;] Vanessa L. Malcarne, PhD[5,9], Brett D. Thombs, PhD[4,6,10-13]; and the SPIN Investigators[14]

[1]Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York, USA

[2]Center for Practice and Research at the Intersection of Information, Society, and Methodology, New York University, New York, New York, USA

[3]Behavioural Science Institute, Clinical Psychology, Radboud University, Nijmegen, the Netherlands

[4]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

[5]San Diego State University/ University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California, USA

[6]Department of Psychiatry, McGill University, Montreal, Quebec, Canada

[7]Scleroderma Foundation, Los Angeles, California, USA

[8]School of Psychology, University of Minho, Braga, Portugal

[9]Department of Psychology, San Diego State University, San Diego, California, USA

[10]Department of Medicine, McGill University, Montreal, Quebec, Canada

[11]Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

[12]Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada

[13]Department of Psychology, McGill University, Montreal, Quebec, Canada

[14]SPIN Investigators: Murray Baron, McGill University, Montreal, Quebec, Canada; Susan J. Bartlett, McGill University, Montreal, Quebec, Canada, Daniel E. Furst, Division of Rheumatology, Geffen School of Medicine, University of California, Los Angeles, California, USA; Maureen D. Mayes, University of Texas McGovern School of Medicine, Houston, Texas, USA; Luc Mouthon, Université Paris Descartes, Paris, France; Warren R. Nielson, St. Joseph's Health Care, London, Ontario, Canada; Robert Riggs, Scleroderma Foundation, Danvers, Massachusetts, USA; Maureen Sauve, Scleroderma Society of Ontario, Hamilton, Ontario; Fredrick Wigley, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Shervin Assassi, University of Texas McGovern School of Medicine, Houston, Texas, USA; Andrea Benedetti, McGill University, Montreal, Quebec, Canada; Isabelle Boutron, Université Paris Descartes, and Assistance Publique - Hôpitaux de Paris, Paris, France; Angela Costa Maia, University of Minho, Braga, Portugal; Lindsay Cronin, Scleroderma Foundation, Western Pennsylvania Chapter, Pittsburgh, Pennsylvania, USA; Ghassan El-Baalbaki, Université du Québec à Montréal, Montreal, Quebec, Canada; Carolyn Ells, McGill University, Montreal, Quebec, Canada; Stephen Elrod, Scleroderma Foundation, Southern California Chapter, Los Angeles, California, USA; Cornelia van den Ende, Sint Maartenskliniek, Nijmegen, The Netherlands; Kim Fligelstone, Scleroderma & Raynaud's UK, London, UK; Catherine Fortune, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Tracy Frech, University of Utah, Salt Lake City, Utah, USA; Amy Gietzen, Scleroderma Foundation, Tri-State Chapter, Binghamton, New York, USA; Dominique Godard, Association des Sclérodermiques de France,

Sorel-Moussel, France; Geneviève Guillot, Sclérodermie Québec, Montreal, Quebec, Canada; Shirley Haslam, Scleroderma Society of Ontario, Hamilton, Ontario, Canada; Monique Hinchcliff, Yale School of Medicine, New Haven, Connecticut, USA; Marie Hudson, McGill University, Montreal, Quebec, Canada; Ann Impens, Midwestern University, Downers Grove, Illinois, USA; Yeona Jang, McGill University, Montreal, Quebec, 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; Annett Körner, McGill University, Montreal, Quebec, Canada; Maggie Larche, McMaster University, Hamilton, Ontario, Canada; Carlo Marra, Memorial University, St. John's, Newfoundland, Canada; Christelle Nguyen, Université Paris Descartes, 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; Alexandra Portales, Asociación Española de Esclerodermia, Madrid, Spain; François Rannou, Université Paris Descartes, and Assistance Publique - Hôpitaux de Paris, Paris, France; Michelle Richard, Scleroderma Society of Nova Scotia, Halifax, Nova Scotia, Canada; Tatiana Sofia Rodriguez Reyna, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; Ken Rozee, Scleroderma Society of Nova Scotia, Halifax, Nova Scotia, Canada; Anne A. Schouffoer, Leiden University Medical Center, Leiden, The Netherlands; Russell J. Steele, Jewish General Hospital and McGill University, Montreal, Quebec, Canada; Nancy Stephens, Scleroderma Foundation, Michigan Chapter, Southflied, Michigan, USA; Maria E. Suarez-Almazor, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; Joep Welling, NVLE Dutch patient organization for systemic autoimmune diseases, Utrecht, The Netherlands, and Federation of European

Scleroderma Associations, Brussel, Belgium; Durhane Wong-Rieger, Canadian Organization for Rare Disorders, Toronto, Ontario, Canada; Christian Agard, Centre Hospitalier Universitaire - Hôtel-Dieu de Nantes, Nantes, France; Alexandra Albert, Université Laval, Quebec, Quebec, Canada; Marc André, Centre Hospitalier Universitaire Gabriel-Montpied, Clermont-Ferrand, France; Guylaine Arsenault, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Ilham Benzidia, Assistance Publique - Hôpitaux de Paris, Hôpital St-Louis, Paris, 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; Gilles Boire, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Alessandra Bruns, Université de Sherbrooke, Sherbrooke, Quebec, Canada; 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; Pascal Cohen, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; Chase Correia, Northwestern University, Chicago, Illinois, USA; Pierre Dagenais, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Christopher Denton, Royal Free London Hospital, London, UK; Robyn Domsic, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; Sandrine Dubois, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; 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; Alexia Esquinca, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico; 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; Anna Gill, Royal Free London Hospital, London, UK; Jessica Gordon, Hospital for Special Surgery, New York City, New York, USA; Brigitte Granel-Rey, Aix Marseille Université, and Assistance Publique - Hôpitaux de Marseille, Hôpital Nord, Marseille, France; Claire Grange, Centre Hospitalier Lyon Sud, Lyon, France; Genevieve Gyger, Jewish General Hospital and McGill University, Montreal, Quebec, Canada; Eric Hachulla, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Pierre-Yves Hatron, Centre Hospitalier Régional Universitaire de Lille, Hôpital Claude Huriez, Lille, France; Sabrina Hoa, CHUM - Centre hospitalier de l'Université de Montréal, Montreal, Quebec, Canada; Ariane L Herrick, University of Manchester, Salford Royal NHS Foundation Trust, Manchester, UK; Adrian Hij, Assistance Publique - Hôpitaux de Paris, Hôpital St-Louis, Paris, France; Alena Ikic, Université Laval, Quebec, Quebec, Canada; Niall Jones, University of Alberta, Edmonton, Alberta, Canada; Artur Jose de B. Fernandes, Université de Sherbrooke, Sherbrooke, Quebec, Canada; Suzanne Kafaja, University of California, Los Angeles, California, USA; Nader Khalidi, McMaster University, Hamilton, Ontario, Canada; 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; Patrick Liang, Université de Sherbrooke, Sherbrooke, Quebec, Canada; 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, 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, Hôpitaux Privés de Metz, Hôpital Belle-Isle, Metz, France; Arsene Mekinian, Assistance Publique - Hôpitaux de Paris, Hôpital St-Antoine, Paris, France; Sheila Melchor, Servicio de Reumatologia del Hospital 12 de Octubre, Madrid, Spain; Mandana Nikpour, St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; 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 Rodriguez, 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; Doug Smith, University of Ottawa, Ottawa, Ontario, Canada; 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; Wendy Stevens, St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; Evelyn Sutton, Dalhousie University, Halifax, Nova Scotia, Canada; Benjamin Terrier, Assistance Publique - Hôpitaux de Paris, Hôpital Cochin, Paris, France; 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; Angelica Bourgeault, Jewish General Hospital, Montreal, Quebec, Canada;

Andrea Carboni Jiménez, Jewish General Hospital, Montreal, Quebec, Canada; Sami Harb, Jewish General Hospital, Montreal, Quebec, Canada; Mengxin Tang, Jewish General Hospital, Montreal, Quebec, Canada; Lydia Tao, Jewish General Hospital, Montreal, Quebec, Canada.

Address for Correspondence: Brett D. Thombs, PhD; Jewish General Hospital; 4333 Cote Ste Catherine Road; Montréal, Québec H3T 1E4; Telephone (514) 340-8222 ext. 25112; E-mail: brett.thombs@mcgill.ca.

Word Count: 3,352

ABSTRACT

Objectives: The Social Appearance Anxiety Scale (SAAS) is a 16-item questionnaire developed to evaluate fear of appearance-based evaluation by others. The primary objective of this research was to investigate the existence of differential item functioning (DIF) for the 16 SAAS items, comparing patients who completed the SAAS in English and French, either to confirm that scores are comparable or provide guidance on calculating comparable scores. A secondary research objective was to investigate the existence of DIF based on sex and disease status. A tertiary research objective was to assess DIF related to language, sex, and disease status on the recently developed SAAS-5.

Design: This was a cross-sectional analysis using baseline data from patients enrolled in the Scleroderma Patient-centered Intervention Network (SPIN).

Setting: SPIN patients included in the present study were enrolled at 43 centers in Canada, USA, UK, France, and Australia, with questionnaires completed in April 2014 through July 2019.

Participants: 1640 SPIN patients completed the SAAS in French (N=600) or English (N=1040).

Primary and secondary measures: The SAAS was collected along with demographic and disease characteristics.

Results: Six items were identified with statistically significant language-based DIF, four with sex-based DIF, and one with disease type-based DIF. However, factor scores before and after accounting for DIF were similar (Pearson correlation > .99), and individual score differences were small. This was true for both the full and shortened versions of the SAAS.

Conclusion: SAAS and SAAS-5 scores are comparable across language, sex, and disease-type, despite small differences in how patients respond to some items.

Keywords: Patient Reported Outcome Measure; Differential Item Functioning; Generalized

Partial Credit Model; Systemic Sclerosis

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study uses a large cohort of patients which provides robust results, allowing for insights into the larger population of adults with scleroderma.
- Patients in the sample were required to have internet access in order to complete study questionnaires, which may limit generalizability of these findings due to selection bias.
- These findings are only generalizable to adults with scleroderma and should be confirmed for other populations.

Introduction

A desire to improve the patient-centered focus of healthcare research has led to the development and increased use of patient-reported outcome (PRO) measures aimed at a wide range of human experiences, including patient-perceived health, well-being and psychological status [1]. This is particularly important in chronic diseases that lead to symptoms that are not directly measurable [2]. Many PRO measures have been translated into multiple languages, which is relevant in treatment centers where more than one language is common, as well as in rare disease research, which often involves collaboration and communication across sites in multiple countries [3]. In these situations, outcomes measured in more than one language are commonly combined in analyses.

In order to compare PROs across language and cultural groups, it is important to ensure that all patients interpret and respond to the questionnaire items in equivalent ways, and not based on idiosyncratic differences due to differing cultural norms, systematic differences in interpretation, or indirect translations of some items [4]. If this is not the case, then items or questions are said to have differential item functioning (DIF). When DIF is present, patients with equal underlying levels of the construct, or latent trait, measured by that scale will respond differently to the same item [5].

Systemic sclerosis (SSc) is a rare, multi-system autoimmune disorder with heterogeneous symptomatology characterized by microvascular damage and fibrosis in multiple organs [6 7]. Changes in appearance are common and can include telangiectasias, hypo- and hyper-pigmentation, loss of skin folds, loss of flexibility of the lips, digital ulcers, and hand contractures [6 8]. These changes in appearance are often in socially relevant areas of the body, such as the face and hands, and can lead to problems with social interactions and increased social appearance anxiety [9].

The Social Appearance Anxiety Scale (SAAS) is a 16-item scale, which aims to measure patients' fear of appearance-based evaluation [10]. Among people with SSc, the SAAS may be used for both individual-level treatment plans and larger scale research, evaluating the impact of potential interventions. The Scleroderma Patient-centered Intervention Network (SPIN) Cohort is a web-based, international cohort designed to collect PROs at regular intervals and as a framework to conduct trials of psychosocial and rehabilitation interventions for patients with SSc [11]. Depending on their native language, participants enrolled in SPIN may complete the SAAS in French, English, or Spanish; however, no research has yet confirmed that SAAS scores are comparable across these language groups.

A recent study developed a shortened version of the SAAS consisting of five items (SAAS-5) for use in patients with SSc [12]. The use of shortened versions, such as the SAAS 5, has the potential to decrease patient burden and increase data quality [13]. However, it is of interest to determine whether the shortened version exhibits DIF.

Therefore, the primary purpose of this analysis is to investigate the comparability of responses to versions of the SAAS administered in different languages. As a secondary research objective, comparability of SAAS scores with respect to disease type and sex were also assessed. A tertiary research objective was to assess the comparability of SAAS scores on the 5-item shortened version.

Materials and Methods

Patients and procedures

The sample consisted of patients enrolled in the SPIN Cohort with complete data study questionnaires from initial enrollment sessions between April 2014 through July 2019. Participants in the SPIN Cohort were enrolled at 43 centers in Canada, USA, UK, France, and Australia. To be eligible for the SPIN Cohort, participants must be classified as having SSc according to the 2013 ACR/EULAR classification criteria [14], confirmed by a SPIN physician, be at least 18 years of age, have the ability to give informed consent, and be fluent in English, French or Spanish. However, the present study only included patients who completed study questionnaires in English or French, as the sample size of Spanish patients was too small to be included at the time of the analyses. Exclusion criteria for participation in the SPIN Cohort include not having access to the internet or otherwise not being able to respond to questionnaires via the internet. The SPIN sample is a convenience sample. Eligible participants are invited by the attending physician or a supervised nurse coordinator to participate in the SPIN Cohort, and written informed consent is obtained. The local SPIN physician or supervised nurse coordinator then completes a medical data form that is submitted online to initiate participants registration in the SPIN Cohort. After completion of online registration, an automated welcoming email is sent to participants with instructions to on how to activate their SPIN online account and how to complete the SPIN Cohort patient measures online. SPIN Cohort participants complete outcome measures via the internet upon enrollment and subsequently every 3 months. The SPIN Cohort study was approved by the Research Ethics Committee of the Jewish General Hospital, Montreal, Canada (MP-05-2013-150, 12-123) and by the Institutional Reviews Boards of each participating center.

Measures

Demographics and disease characteristics. Demographic and disease variables included age, sex, race/ethnicity, marital status, education level, time since diagnosis, and SSc subtype. Disease subtypes were classified as limited or diffuse. Limited disease was defined as skin involvement distal to the knees and elbows only, whereas diffuse disease included more extensive skin involvement [15]. The country of patient recruitment and language of assessment were also recorded.

Social appearance anxiety scale (SAAS). The SAAS consists of 16 items assessing patients' self-reported anxiety about appearance-based evaluation. The SAAS was initially validated among three samples of undergraduate students (n=512, 853, and 541, respectively) [10]. In this population, the SAAS was shown to have unifactorial structure, high internal consistency, high test-retest reliability, and was positively correlated with other social anxiety measures [10]. A recent study of 938 participants enrolled in the SPIN Cohort demonstrated that the SAAS is a unidimensional, reliable, and valid measurement of social appearance anxiety among people with SSc [16]. The SAAS was initially written in English. The French version used in this study was translated by SPIN investigators using the forward-backward method [17]. For both versions, item responses are recorded on a five point scale (1=Not at all, 5=Extremely). Item 1 ("I feel comfortable with the way I appear to others") is reverse coded before summing across items to produce a total score ranging from 16 to 80. Higher scores indicate higher levels of appearance anxiety.

The SAAS-5, consisting of items 6, 7, 12, 13, and 14 from the original SAAS, was recently developed and validated for use in patients with SSc [12]. Scores on the SAAS-5 range from 5 to 25, with higher scores indicating higher levels of appearance anxiety.

Statistical analysis

The English- and French-speaking samples were compared based on demographic and disease characteristics to identify possible differences between the two language groups.

A generalized partial credit model (GPCM) was then used to model the latent factor (social anxiety with appearance) underlying the SAAS. For each item, a GPCM was used to estimate two types of item-level parameters: 1) thresholds (betas) for the level of the latent factor (theta) at which respondents are more likely to endorse a given response category instead of the category below, and 2) a discrimination parameter (alpha) that measures the strength of the relationship between that item and the underlying latent factor [18].

Item-characteristic curves (ICCs) are often used to visualize these parameters, and Figure 1 shows three examples of ICCs for a hypothetical 5-category item. Each curve in an ICC plot corresponds to a possible categorical response. Along the latent spectrum, the height of each curve indicates the estimated probability that a respondent with a particular level of the latent factor will respond with the corresponding category. Item level thresholds are visualized as the intersections between consecutive curves; discrimination parameters are visualized as the peaked-ness of the curves. When item-level thresholds vary across observed groups, items are said to display uniform DIF. Uniform DIF could be visualized as a horizontal shift of ICC curves for one demographic group compared to the other. Meanwhile, when the discrimination parameter varies across observed groups, items are said to display non-uniform DIF. Non-uniform DIF could be visualized as a change in the peaked-ness of the curves for one demographic group compared to the other.

The *lordif* package in R [19 20] was used to identify items with language-based DIF through an iterative procedure. The algorithm implemented by *lordif* iteratively fits three ordinal logistic models for each item and uses these models to flag items with potential DIF. The first model predicts the probability of each response category using estimated latent factor scores alone, while the second and third models test for uniform and non-uniform DIF, respectively. Once a set of items is flagged, the algorithm then re-estimates latent factor scores using another GPCM that accounts for DIF on those items. DIF is accounted for by allowing item level parameters to vary across groups. The process stops once the same items are repeatedly flagged for DIF [20].

During the iterative search for items with DIF, items were flagged using a chi-square test comparing the first and third models (alpha = 0.01 significance level). Flagged items were then re-examined to distinguish between uniform and non-uniform DIF. This was done by separately comparing the first and second models (to ascertain uniform DIF) and second and third models (to ascertain non-uniform DIF), again using a chi-square test (alpha = 0.01 significance level).

Items with DIF were further investigated by comparing item-level parameters from a GPCM for patients who completed the SAAS in English and French. To visualize and understand differences among the two groups on each item, item true score functions for English- and French-speakers were compared, which show expected responses for items with DIF as a function of estimated latent social appearance anxiety scores accounting for DIF.

The questionnaire-level impact of DIF on estimated latent factor scores was assessed by plotting test characteristic curves, which show expected summed scores on the SAAS as a function of patients' GPCM scores accounting for DIF. As per previous guidelines, impact was numerically assessed by comparing initial scores (not accounting for DIF) to final scores (accounting for DIF), using the Pearson correlation of the two scores and by comparing individual score differences to the standard errors of initial scores [21 22]. To assess whether the correlation significantly differed from 1, a randomization null distribution and p-values were obtained by randomly permuting group labels 1000 times and re-estimating scores and statistics holding the measurement model fixed across permutations, but re-estimating the item-level parameters based on the permuted dataset.

Lastly, the median and range of score differences (of the difference between scores accounting for and not accounting for DIF) were also calculated, and score differences were plotted against initial scores to find areas of the latent spectrum with highest DIF impact. Before comparison, scores were placed on the same scale using a transformation by Stocking and Lord [23]. This was also done using the *lordif* package, which equates final scores accounting for DIF to initial GPCM estimates using the non-DIF items as anchors.

The same process was repeated to identify and investigate DIF related to sex and disease status, respectively, and additionally for the SAAS-5.

Patient Involvement

SPIN was conceived by a collaboration of investigators and patients. SPIN's Patient Advisory Board advises the SPIN Steering Committee on priorities for investigation. Patients were included in the SPIN Publication Committee, which reviewed the proposal for the present study and its methods. Two patients were co-authors of the present report.

Results

Table 1. Demographic and disease characteristics by assessment language

Variable	All Patients (N = 1640)	English-Speaking Patients (N = 1040)	French-Speaking Patients (N = 600)
Mean age, years (SD)[a]	55.1 (12.5)	55.7 (11.7)	54.0 (13.8)
Female (%)	87.2	87.6	86.5
Mean SAAS summed score (SD)	29.1 (13.7)	28.3 (13.2)	30.5 (14.5)
Diffuse disease type (%)	39.0	42.4	33.2
Mean time since diagnosis, years (SD)	9.2 (7.9)	9.7 (8.0)	8.5 (7.6)
Married or common law (%)	71.2	73.3	67.5
At least 12 years of education (%)	85.7	94.2	70.8
Race[b]			
White (%)	83.6	83.9	83.0
Black (%)	7.1	6.1	8.8
Other (%)	9.3	10.0	8.2
Country of patient recruitment			
Canada (%)	24.9	28.7	18.5
USA (%)	35.5	55.9	0.2
UK (%)	9.7	15.3	0.0
France (%)	29.8	0.1	81.3
Australia (%)	0.1	0.1	0.0

Due to missing values: a) N=1036 for the English cohort, b) N=1038 for the English cohort

The English and French samples included 1040 and 600 patients, respectively. Table 1 presents descriptive statistics for the full sample, as well as the English and French samples separately.

DIF Analysis

Six of the 16 SAAS items (Table 2) were identified as having statistically significant ($p < .01$) language-based DIF: items 2, 5, 8, 11, 12, and 13. Only item 11 was identified as having non-uniform DIF. Item true score functions for these six items are shown in Figure 2. For most items with uniform DIF, French speakers' expected item level responses were slightly higher than their English-speaking counterparts with equal levels of appearance anxiety. This pattern is reversed for item 2.

Table 2. SAAS Items

Item #	Item Text
1	I feel comfortable with the way I appear to others
2	I feel nervous when having my picture taken
3	I get tense when it is obvious people are looking at me
4	I am concerned people won't like me because of the way I look
5	I worry that others talk about flaws in my appearance when I am not around
6	I am concerned people will find me unappealing because of my appearance
7	I am afraid people find me unattractive
8	I worry that my appearance will make life more difficult for me
9	I am concerned that I have missed out on opportunities because of my appearance
10	I get nervous when talking to people because of the way I look
11	I feel anxious when other people say something about my appearance
12	I am frequently afraid that I won't meet others' standards of how I should look
13	I worry people will judge the way I look negatively
14	I am uncomfortable when I think others are noticing flaws in my appearance
15	I worry that a romantic partner will/would leave me because of my appearance
16	I am concerned that people think I am not good looking

Test characteristic curves for the English and French cohorts are plotted in Figure 3, while Figure 4 shows score differences based on GPCMs that do and do not account for DIF. At the questionnaire level, French speakers are expected to have slightly larger summed scores on the SAAS as compared to English speakers with the same level of appearance anxiety. The correlation between the two sets of GPCM scores was 0.99977 (95% confidence interval: [0.99975, 0.99979], $p < 0.001$). At the individual level, the median GPCM score difference (scores accounting for DIF minus scores that do not account for DIF) was 0.0049, and differences in factor scores ranged from -0.078 to 0.065. No individual score differences exceeded the standard errors of initial estimates. Patients with the largest score differences had initial GPCM scores around -0.5 and 1.0, whereas individuals whose initial estimated anxiety level was extreme (low or high) or average had smaller DIF impact.

Four items were identified as having sex-based DIF (all uniform): items 2, 4, 9, and 14. Only item 2 exhibited both language and sex based DIF. Item true score functions suggest that females tend to give slightly higher categorical responses than equally anxious males on items 2 and 14 and lower responses on items 4 and 9. Meanwhile, the test characteristic curves for males and females were practically indistinguishable, suggesting that equally anxious males and females have almost identical expected summed scores. The correlation between the two sets of GPCM scores was 0.99985 (95% confidence interval: [0.99983, 0.99986], $p=0.003$). At the individual level, the median score difference based on a GPCM was 0.0020, and differences in factor scores ranged from -0.047 to 0.135. No individual score differences exceeded the standard errors of initial estimates. The largest score differences were observed for individuals whose initial GPCM score was low (around -1.0 in this dataset); individuals with average or high estimated anxiety levels had comparatively low DIF impact.

Only one item (item 9) was identified as having DIF related to disease-type (non-uniform). This item was also identified as having sex-based DIF, but not language-based DIF. Among patients with low appearance anxiety, those with limited disease are expected to give smaller categorical responses to item 9 than patients with diffuse disease and equal levels of appearance anxiety; this pattern is reversed at the higher end of the latent spectrum. At the questionnaire level, expected summed scores were nearly identical across disease-type groups. The correlation between the two sets of GPCM scores was 0.99996 (95% confidence interval: [0.99996, 0.99997], $p < 0.001$). At the individual level, the median GPCM score difference was 0.001 and these differences in factor scores ranged from -0.101 to 0.080. No individual score differences exceeded the standard errors of initial estimates. The largest score differences were observed for individuals whose initial GPCM estimate was around 0, or slightly below.

For the SAAS-5, only item 12 was flagged for language based DIF, while item 14 was still flagged for gender-based DIF. In both cases, the correlation between factor scores was still high: 0.99995 for language-based DIF (95% confidence interval: [0.99995, 0.99996], $p = 0.107$) and 0.99971 for gender-based DIF (95% confidence interval: [0.99969, 0.99974], $p = 0.018$).

Discussion

This study investigated whether the SAAS displays DIF across language, sex, and disease subtype groups among people with SSc. Nine items were flagged for language-based DIF (8 uniform, 1 non-uniform), four were flagged with sex-based DIF (all uniform), and only one was flagged with disease-type based DIF (non-uniform). In reviewing translations of the items flagged with language-based DIF, we did not observe any clear differences. Similarly negligible levels of DIF were found for the SAAS-5.

For all three analyses on the full-length SAAS, the high (>0.99) Pearson correlations between the two GPCM estimates imply that accounting for DIF does not provide much additional information about respondents' comparative levels of social appearance anxiety. The near-zero (<0.01) associated p-values nonetheless suggest that observed correlations are lower than what would be expected by random chance in a no-DIF null condition under identical measurement models. While previous analyses have used Pearson correlations [21 22] to compare GPCM scores that do and do not account for DIF, other research has cautioned against this [24]. Our findings imply that very large correlations between initial and final GPCM estimates may still be smaller than simulated values under a no-DIF condition. Thus, we caution that correlations alone may not be particularly interpretable as a measure of DIF impact.

The relatively small ranges of GPCM score differences in all three analyses nonetheless support the conclusion that accounting for DIF has limited impact on individual estimated scores. In all cases, no individual differences exceeded initial standard errors. Thus, estimated scores accounting for DIF were all within the range of inherent uncertainty in naïve GPCM estimates. The median score difference was largest for language-based DIF and smallest for disease-type-based DIF; however, the range of score differences was smallest in the language-based analysis, due to the existence of a few outliers in the other two cases.

Scatter plots of GPCM score differences as a function of naïve GPCM estimates (see Figure 4 for language-based DIF) show that language-, sex-, and disease-type-based DIF impact is not constant across the latent spectrum. Naïve GPCM estimates near values where GPCM score differences are larger (i.e., near -0.5 and 1 for language-based DIF, -1 for sex-based DIF, and 0 for disease-type-based DIF) may therefore be slightly less certain.

While DIF impact was found to be small for both simple summed scores and naïve GPCM estimates, it is important to note that the choice between these two scoring methods is also relevant [25 26]. This paper explored three different methods for estimating social appearance anxiety levels based on responses to the SAAS: simple summed scores, naïve GPCM factor scores, and GPCM factor scores accounting for DIF. Our analysis aimed to assess comparability of scores across demographic groups, and therefore mainly focused comparison between the two sets of GPCM factor scores; however, much more confidence in individual scores is gained in the jump from simple summed scores to a GPCM factor score, than in the jump from a naïve GPCM factor score to a GPCM factor score accounting for DIF. For example, in this dataset, individuals with the same summed score had naïve GPCM estimates of social appearance anxiety differing by up to 0.92 standardized units. Thus, regardless of whether DIF is accounted for in score calculations, a GPCM-based score or weighted summed score would be preferable over a simple summed score.

This study has several limitations. First, DIF was only investigated in the population of adults with scleroderma and results may not generalize to the general population. Second, in order to complete study questionnaires, patients were required to have access to the internet, which may bias the sample. Specifically, those with most severe disease may not be able to type due to the inability to use their fingers or hands. As well, it is possible that the oldest patients would be unable to participate. However, although the SPIN Cohort constitutes a convenience sample of SSc patients receiving treatment at a SPIN recruiting center, and patients at these centers may differ from those in other settings, a comparison between SPIN Cohort participants and the European Scleroderma Trials and Research (EUSTAR) and Canadian Scleroderma Research Group (CSRG) cohorts showed broad comparability [27]. This increases confidence that insights gained from the SPIN Cohort should be generalizable.

Conclusion

In conclusion, this study used an iterative algorithm implemented via the *lordif* package in R to flag items on the SAAS for DIF related to language of test administration, sex, and disease type. After flagging items with DIF, impact was assessed primarily by looking at GPCM score correlations and differences before and after accounting for DIF. While at least one item was flagged for DIF in each analysis, DIF impact was assessed to be small, supporting the conclusion that GPCM scores are comparable across groups produced by these three demographic variables.

FIGURE CAPTIONS

Figure 1. Three possible ICC curves for a 5 category item. The left and middle panels show ICCs for items with the same approximate discrimination parameters (alphas) but different item-level thresholds (betas). The left and right panels show ICCs for items with the same approximate item-level thresholds (betas) but different discrimination parameters (alphas).

Figure 2. Item True Score Functions for the six items identified as having language-based DIF. For items 5, 8, 12, and 13, these plots demonstrate that French speakers are expected to give larger categorical responses than English speakers with equal levels of appearance anxiety. This

trend is reversed for item 2, while item 11 demonstrates non-uniform DIF (i.e., the true score functions for English and French speakers cross each other).

Figure 3. Test characteristic curve showing expected summed scores on the SAAS as a function of estimated social appearance anxiety accounting for DIF. Thus, among French and English speakers with the same estimated level of social appearance anxiety, French speakers are expected to have slightly larger summed scores.

Figure 4. The top plot shows GPCM score differences at the questionnaire level (accounting for DIF – not accounting for DIF) compared to factor scores accounting for DIF. The largest score differences occur at estimated appearance anxiety levels .5 standard deviations (sd) below average and 1 sd above average. The figure on the bottom left shows a box plot of these score differences among all respondents. The figure on the bottom right shows these differences by language. Overall differences are small and are mostly negative for English speakers and positive for French speakers, suggesting that pooled scores from a GPCM will tend to overestimate appearance anxiety for French speakers and underestimate it for English speakers.

AUTHOR CONTRIBUTIONS

DH and BDT were responsible for the study conception. LK, M-EC, VLM, BDT and the SPIN Investigators contributed to data collection. SJS, DH, LK, M-EC, SG, KG, CL, VLM, BDT contributed to data analysis and interpretation. SJS and DH drafted the manuscript. All authors provided a critical revision of the manuscript and approved the final version of the manuscript. DH is the guarantor.

FUNDING STATEMENT

SPIN has been funded by grants from the Canadian Institutes of Health Research (TR3-119192, PJT-148504, PJT-149073, SCT-162963), the Canadian Initiative for Outcomes in Rheumatology Care, and the Arthritis Society. In addition, SPIN has received institutional contributions from the Lady Davis Institute for Medical Research of the Jewish General Hospital, Montreal, Canada and from McGill University, Montreal, Canada. SPIN has also received support from Scleroderma Canada, the Scleroderma Society of Ontario, Sclérodermie Québec, the Scleroderma Society of Nova Scotia, the Scleroderma Association of British Columbia, Scleroderma Manitoba, and the Scleroderma Society of Saskatchewan.

COMPETING INTERESTS STATEMENT

The authors have read and understood the BMJ policy on declaration of interests and declare that they have no competing interests.

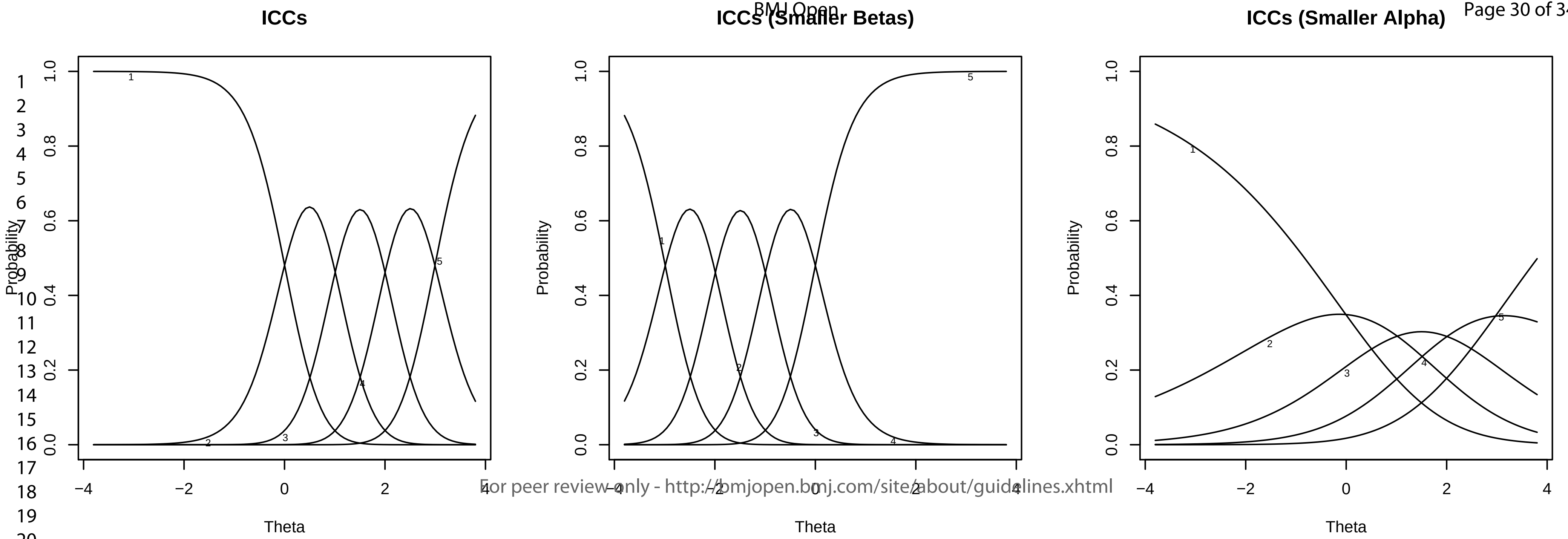
DATA SHARING STATEMENT

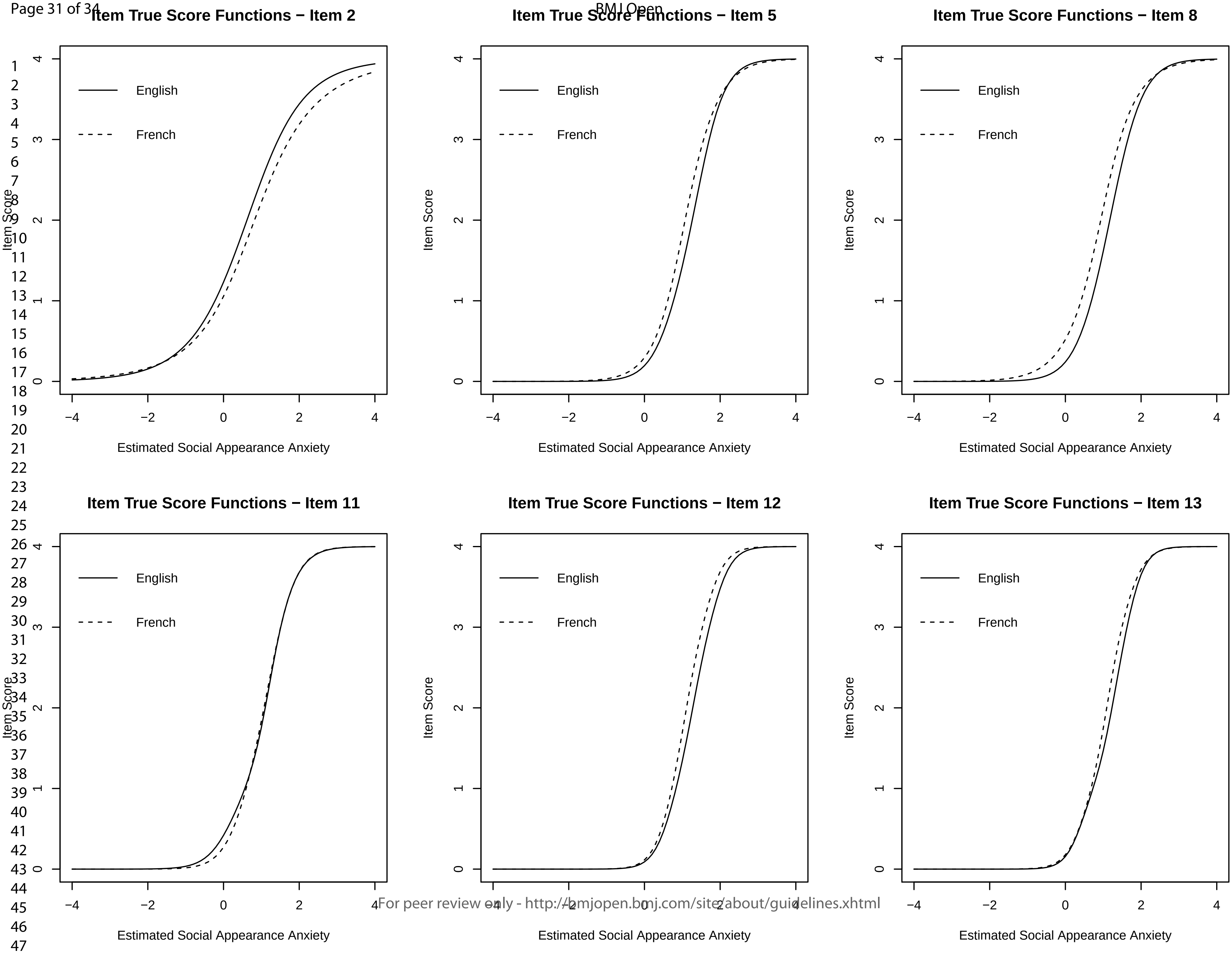
Data from the SPIN Cohort can be requested from the corresponding author.

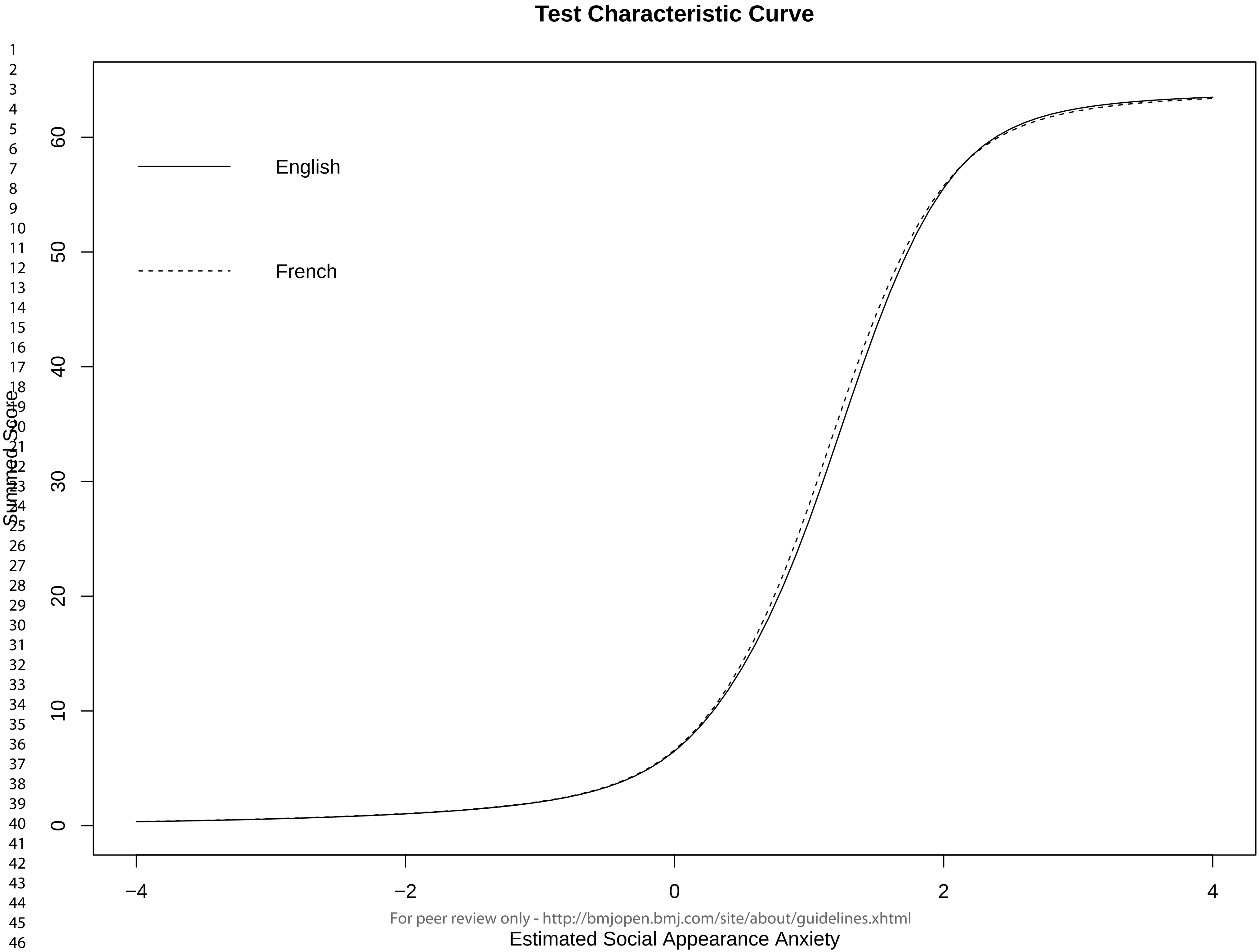
References

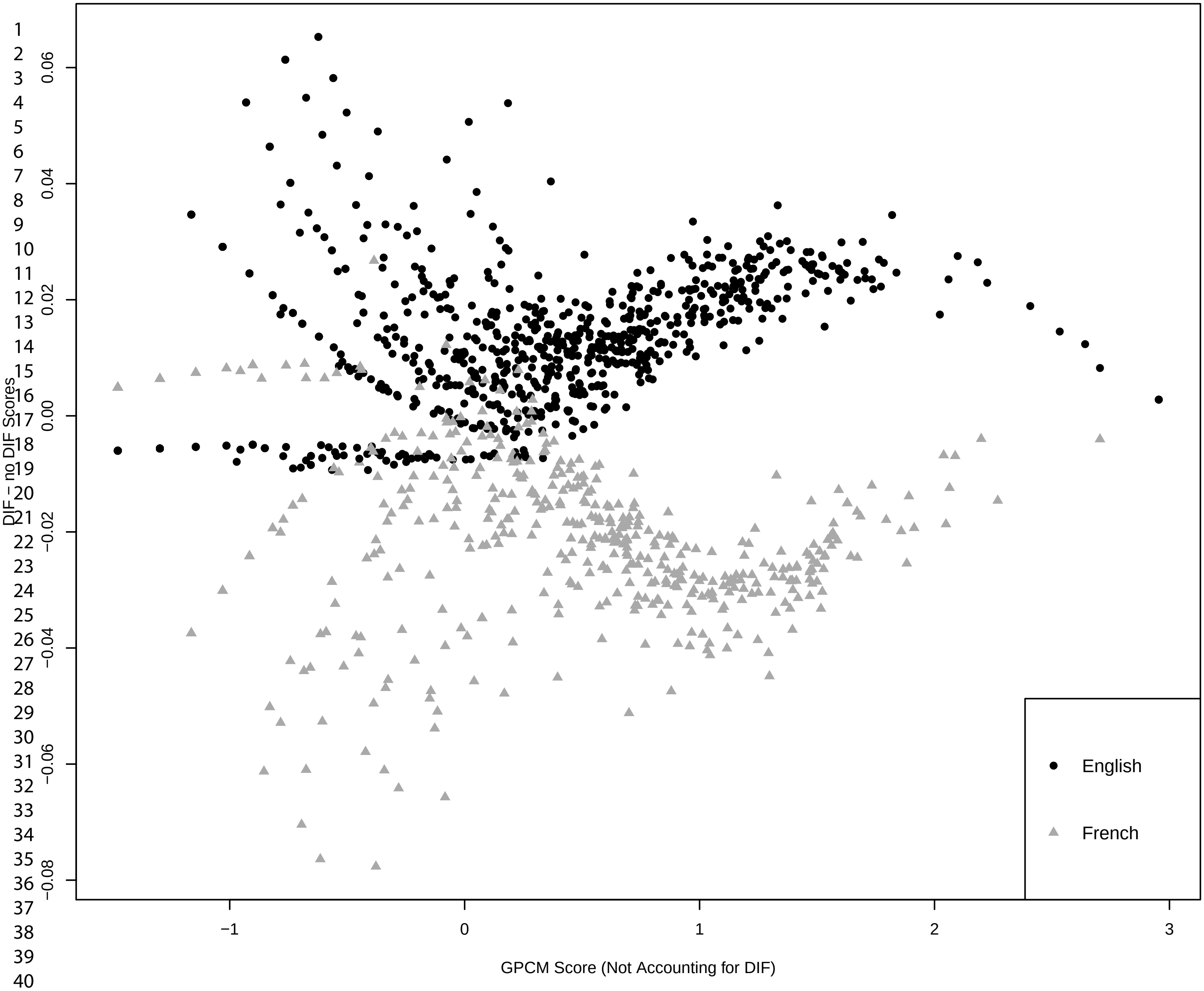
1. Weldring T, Smith SMS. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health Serv Insights* 2013;6:61-68. doi: 10.4137/HSI.S11093
2. Thombs BD, van Lankveld W, Bassel M, et al. Psychological health and well-being in systemic sclerosis: State of the science and consensus research agenda. *Arthritis Care & Research* 2010;62(8):1181-89. doi: 10.1002/acr.20187
3. Kwakkenbos L, Willems LM, Baron M, et al. The Comparability of English, French and Dutch Scores on the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F): An Assessment of Differential Item Functioning in Patients with Systemic Sclerosis. *PLOS ONE* 2014;9(3):e91979. doi: 10.1371/journal.pone.0091979
4. Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Med Care* 2006;44(11 Suppl 3):S39-49. doi: 10.1097/01.mlr.0000245452.48613.45
5. Holland PW, Wainer H. Differential Item Functioning: Taylor & Francis 2012.
6. Allanore Y, Simms R, Distler O, et al. Systemic sclerosis. (2056-676X (Electronic))
7. Clements PJ, Furst DE. Systemic Sclerosis. 2 ed. Baltimore: Lippincott Williams & Wilkins 2003.
8. Gholizadeh S, Fox RS, Mills SD, et al. Coping with the disfigurement of scleroderma: Facial, skin, and hand changes. Scleroderma. 2 ed: Springer International Publishing 2017:713-21.
9. Kwakkenbos L, Delisle VC, Fox RS, et al. Psychosocial Aspects of Scleroderma. *Rheum Dis Clin North Am* 2015;41(3):519-28. doi: 10.1016/j.rdc.2015.04.010
10. Hart TA, Flora DB, Palyo SA, et al. Development and Examination of the Social Appearance Anxiety Scale. *Assessment* 2008;15(1):48-59. doi: 10.1177/1073191107306673
11. Kwakkenbos L, Jewett Lr Fau - Baron M, Baron M Fau - Bartlett SJ, 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. LID - 10.1136/bmjopen-2013-003563 [doi] LID - e003563 [pii]. (2044-6055 (Electronic))
12. Harel D, Mills SD, Kwakkenbos L, et al. Shortening patient-reported outcome measures through optimal test assembly: application to the Social Appearance Anxiety Scale in the Scleroderma Patient-centered Intervention Network Cohort. *BMJ Open* 2019;9(2):e024010. doi: 10.1136/bmjopen-2018-024010
13. Harel D, Baron M. Methods for shortening patient-reported outcome measures. *Statistical Methods in Medical Research* 2018;28(10-11):2992-3011. doi: 10.1177/0962280218795187
14. van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Ann Rheum Dis* 2013;72(11):1747-55. doi: 10.1136/annrheumdis-2013-204424
15. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. *J Rheumatol* 1988;15(2):202-05.

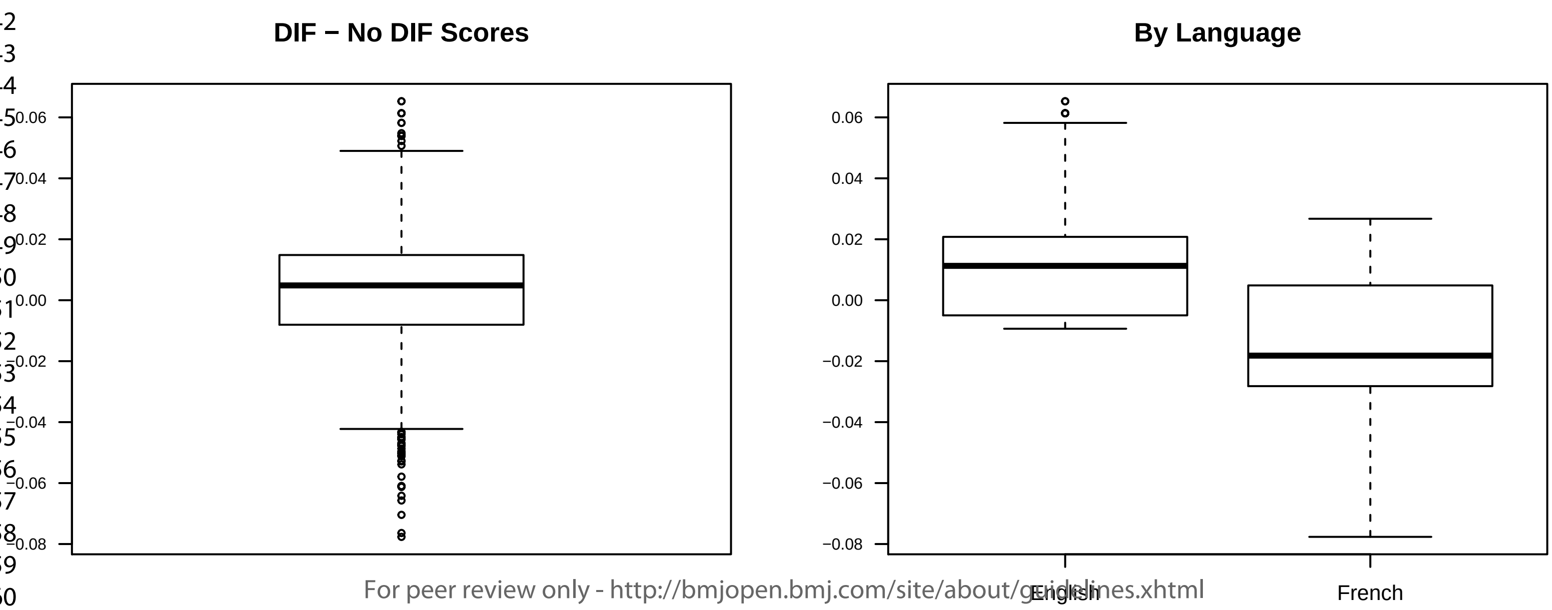
16. Mills SD, Kwakkenbos L, Carrier M-E, et al. Validation of the Social Appearance Anxiety Scale in Patients With Systemic Sclerosis: A Scleroderma Patient-Centered Intervention Network Cohort Study. *Arthritis Care & Research* 2018;70(10):1557-62. doi: 10.1002/acr.23514
17. Wild D, Grove A Fau - Martin M, Martin M Fau - Eremenco S, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. (1098-3015 (Print))
18. Muraki E. A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement* 1992;16(2):159-76. doi: 10.1177/014662169201600206
19. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2019.
20. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw* 2011;39(8):1-30.
21. Chung H, Kim J, Cook KF, et al. Testing measurement invariance of the patient-reported outcomes measurement information system pain behaviors score between the US general population sample and a sample of individuals with chronic pain. *Quality of Life Research* 2014;23(1):239-44. doi: 10.1007/s11136-013-0463-0
22. Cook KF, Bombardier CH, Bamer AM, et al. Do somatic and cognitive symptoms of traumatic brain injury confound depression screening? *Arch Phys Med Rehabil* 2011;92(5):818-23. doi: 10.1016/j.apmr.2010.12.008
23. Stocking ML, Lord FM. Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement* 1983;7(2):201-10. doi: 10.1177/014662168300700208
24. Rupp AA, Zumbo BD. A Note on How to Quantify and Report Whether Irt Parameter Invariance Holds: When Pearson Correlations are Not Enough. *Educational and Psychological Measurement* 2004;64(4):588-99. doi: 10.1177/0013164403261051
25. Harel D. The effect of model misspecification for polytomous logistic adjacent category item response theory models. McGill University Libraries, 2014.
26. van der Ark LA. Stochastic Ordering Of the Latent Trait by the Sum Score Under Various Polytomous IRT Models. *Psychometrika* 2005;70(2):283-304. doi: 10.1007/s11336-000-0862-3
27. Dougherty DH, Kwakkenbos L, Carrier ME, et al. The Scleroderma Patient-Centered Intervention Network Cohort: baseline clinical features and comparison with other large scleroderma cohorts. (1462-0332 (Electronic))


ICCs
ICCs (Smaller Betas)
ICCs (Smaller Alpha)
Probability
Theta
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
1
2
3
4
5


Item True Score Functions − Item 2
Item True Score Functions − Item 5
Item True Score Functions − Item 8
Item True Score Functions − Item 11
Item True Score Functions − Item 12
Item True Score Functions − Item 13
English
French
Item Score
Estimated Social Appearance Anxiety
0
1
2
3
4
−4
−2
0
2
4


Test Characteristic Curve
English
French
Summed Score
0
10
20
30
40
50
60
−4
−2
0
2
4
Estimated Social Appearance Anxiety


0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
DIF − no DIF Scores
−1
0
1
2
3
GPCM Score (Not Accounting for DIF)
English
French


DIF − No DIF Scores
By Language
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
English
French

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	8
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	8
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	10
Objectives	3	State specific objectives, including any prespecified hypotheses	11
Methods			
Study design	4	Present key elements of study design early in the paper	12
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	12
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	12
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	13
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	13
Bias	9	Describe any efforts to address potential sources of bias	n/a
Study size	10	Explain how the study size was arrived at	12
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	13
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	14 - 16
		(*b*) Describe any methods used to examine subgroups and interactions	n/a
		(*c*) Explain how missing data were addressed	n/a
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	n/a
		(*e*) Describe any sensitivity analyses	n/a
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	18
		(b) Give reasons for non-participation at each stage	n/a
		(c) Consider use of a flow diagram	n/a
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	17
		(b) Indicate number of participants with missing data for each variable of interest	n/a
Outcome data	15*	Report numbers of outcome events or summary measures	17
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	20

		(*b*) Report category boundaries when continuous variables were categorized	n/a
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	n/a
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	n/a
Discussion			
Key results	18	Summarise key results with reference to study objectives	21
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	23
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	22
Generalisability	21	Discuss the generalisability (external validity) of the study results	23
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	26

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.