The following full text is a publisher's version.

For additional information about this publication click this link.
https://hdl.handle.net/2066/226487

Please be advised that this information was generated on 2021-02-07 and may be subject to change.
Assessing differential item functioning for the Social Appearance Anxiety Scale: a Scleroderma Patient-centred Intervention Network (SPIN) Cohort Study

Sophia J Sommer,1,2 Daphna Harel,1,2 Linda Kwakkenbos,3 Marie-Eve Carrier,4 Shadi Gholizadeh,5,6 Karen Gottesman,7 Catarina Leite,8 Vanessa L Malcarne,5,9 Brett D Thombs,10,11,12,13

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-centred Intervention Network (SPIN).

Setting SPIN patients included in the present study were enrolled at 43 centres in Canada, USA, UK, France and Australia, with questionnaires completed in April 2014 to 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 >0.99), 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.

INTRODUCTION

A desire to improve the patient-centred 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 centres 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.  
Systemic sclerosis (SSc) is a rare, multisystem autoimmune disorder with heterogeneous symptomatology characterised by microvascular damage and fibrosis in multiple organs. Changes in appearance are common and can include telangiectasias, hypopigmentation and hyperpigmentation, loss of skin folds, loss of flexibility of the lips, digital ulcers and hand contractures. 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. 

The Social Appearance Anxiety Scale (SAAS) is a 16-item scale, which aims to measure patients’ fear of appearance-based evaluation. 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-centred 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. 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. The use of shortened versions, such as the SAAS-5, has the potential to decrease patient burden and increase data quality. 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 and July 2019. Participants in the SPIN Cohort were enrolled at 43 centres 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, 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 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 on enrollment and subsequently every 3 months.

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. The country of patient recruitment and language of assessment were also recorded.

Social Appearance Anxiety Scale
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). 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. 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. The SAAS was initially written in English. The French version used in this study was translated by SPIN investigators using the forward-backward method. 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. Scores on the SAAS-5 range from 5 to 25, with higher scores indicating higher levels of appearance anxiety.
Statistical analysis

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

A generalised 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 visualise 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 visualised as the intersections between consecutive curves; discrimination parameters are visualised 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 visualised as a horizontal shift of ICC for one demographic group compared with the other. Meanwhile, when the discrimination parameter varies across observed groups, items are said to display non-uniform DIF. Non-uniform DIF could be visualised as a change in the peaked-ness of the curves for one demographic group compared with the other.

The lordif package in R19 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

Figure 1 Three possible ICC for a five-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). ICC, item-characteristic curve.

Table 1 Demographic and disease characteristics by assessment language

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

Due to missing values.
* n=1036 for the English cohort.
† n=1038 for the English cohort.
SAAS, Social Appearance Anxiety Scale.
Table 2: SAAS items

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

SAAS, Social Appearance Anxiety Scale.

is accounted for by allowing item level parameters to vary across groups. The process stops once the same items are repeatedly flagged for DIF.

During the iterative search for items with DIF, items were flagged using a $\chi^2$ 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 $\chi^2$ 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 visualise and understand differences among the two groups on each item, item true score functions for English-speakers and French-speakers were compared, which showed expected responses for items with DIF as a function of estimated latent social appearance anxiety accounting for DIF.

The questionnaire-level impact of DIF on estimated latent factor scores was assessed by plotting test characteristic curves, which showed 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 SEs of initial scores. To assess whether the correlation significantly differed from 1, a randomisation 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. 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 coauthors of the present report.

RESULTS

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<0.01) 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.

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 with English speakers with the same level of appearance anxiety. The correlation between the two sets of GPCM scores was 0.99977 (95% CI 0.99975...
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 two 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% CI 0.99983 to 0.99986, p=0.003). At the individual level, the median score difference based on a GPCM was 0.0020, and differences in

---

**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 (ie, the true score functions for English and French speakers cross each other). DIF, differential item functioning.

**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. DIF, differential item functioning; SAAS, Social Appearance Anxiety Scale.
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 nine 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% CI 0.99996 to 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 SEs 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% CI 0.99995 to 0.99996, p=0.017) and 0.99971 for gender-based DIF (95% CI 0.99969 to 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 (eight uniform, one 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 to compare GPCM scores that do and do not account for DIF, other research has cautioned against this. 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 SEs. Thus, estimated scores accounting for DIF were all within the range of inherent uncertainty in naive 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 naive GPCM estimates (see figure 4 for language-based DIF) show that language-type, sex-type and disease-type-based DIF impact is not constant across the latent spectrum. Naïve GPCM estimates near values where GPCM score differences are larger (ie, 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 standardised 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 generalise 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 centre, and patients at these centres may differ from those in other settings, a comparison between SPIN Cohort participants and the European Scleroderma Trials and Research and Canadian Scleroderma Research Group cohorts showed broad comparability.27 This increases confidence that insights gained from the SPIN Cohort should be generalisable.

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.


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