Cross-Cultural Evaluation of the Dutch FACE-Q Rhinoplasty Questionnaires Using Rasch Analysis

Aesthetic Surgery Journal
2021, Vol 41(12) NP1916–NP1930
© The Author(s) 2021. Published by Oxford University Press on behalf of The Aesthetic Society. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
DOI: 10.1093/asj/sjab217
www.aestheticsurgeryjournal.com
OXFORD
UNIVERSITY PRESS

Frank Declau, MD, PhD; Laura Pingnet, MD; Valérie Verkest, MD; and Tina Hansen, MSc, OT, PhD

Abstract

Background: The English version of the FACE-Q rhinoplasty module, developed according to Rasch measurement theory, has recently been translated into Dutch. Before conclusions can be drawn from the Dutch version, this translation must also fit the item analysis by the Rasch model.

Objectives: The primary aim of this study was to evaluate cross-cultural equivalence between the Dutch and English versions of the FACE-Q rhinoplasty module by applying Rasch methodology.

Methods: Rasch analysis performed with Winsteps (Beaverton, OR) was used to evaluate the Dutch version of the FACE-Q nose and nostrils scales with data from a prospective consecutive cohort of 100 Dutch-speaking septorhinoplasty patients. New Dutch-related conversion tables were constructed for the FACE-Q nose and nostrils scales and compared to the original ones. Psychometric cross-validation was performed by receiver operating characteristics (ROC) analysis.

Results: Both questionnaires adequately met the requirement of invariance. Within an acceptable range, some issues with item and person fit were found, as well as some local item dependency and differential item functioning. However, comparison of the Dutch- and English-related conversion tables by ROC analysis demonstrated identical results for the FACE-Q nose and nostrils scales.

Conclusions: Item analysis by the Rasch model on the data of a Dutch-speaking population proved the conceptual correspondence with the original English version.

Editorial Decision date: April 19, 2021; online publish-ahead-of-print May 4, 2021.

Septorhinoplasty (SRP) is frequently performed by ear, nose, and throat (ENT) surgeons for both functional and aesthetic reasons. One of the most difficult aspects of SRP is measuring the outcomes after surgery.[1] During the past decade, considerable emphasis has been placed on health-related quality-of-life measurements. Generic or disease-specific instruments measure a patient's own perception of health and ability to function as a result of disease and health status. Patient-reported outcome measures (PROMs) are a recommended tool for evaluating postoperative results.[2]

Dr Declau is the head of the ENT Department, Sint-Vincentius Hospital, Antwerp, Belgium. Drs Pingnet and Verkest are residents, ENT Department, Sint-Vincentius Hospital, Antwerp, Belgium. Dr Hansen is a senior researcher, Department of Physical and Occupational Therapy, University Hospital Hvidovre-Amager, Hvidovre, Denmark.

Corresponding Author:
Dr Frank Declau, Sint-Vincentius Hospital, Sint-Vincentiusstraat 20, 2018-Antwerp, Belgium.
E-mail: frank.declau@gza.be

The FACE-Q rhinoplasty module is an instrument designed to evaluate patient-reported outcomes before and after undergoing rhinoplasty. The English version of FACE-Q was developed by Klassen et al in 2010 as a validated psychometric evaluation instrument for patients undergoing aesthetic surgery and is owned by Memorial Sloan Kettering Cancer Center (New York, NY), which holds the copyright of the original and all translated versions of the questionnaires.[3,4] FACE-Q was developed to measure the impact and effectiveness of facial aesthetic procedures from the patients' perspective and has the potential to support an evidence-based approach to facial aesthetic practice.[5] Two domains of the FACE-Q instrument are used to evaluate SRP: "satisfaction of the nose" contains 10 questions, and "satisfaction of the nostrils" contains 5 questions; both domains are graded on a 4-point Likert scale. These questionnaires have a Flesch-Kincaid grade of 0.8 for the nose and 1.4 for the nostrils.[5] The summation of the raw ordinal scores is converted into equivalent linear interval data from 0 to 100, generated by a Rasch transformation with higher scores indicating better outcomes.

For the English version of the FACE-Q questionnaires sufficient structural validity and internal consistency have been demonstrated.[6,7] The use of this rhinoplasty module requires translation into other languages. At this moment, the FACE-Q rhinoplasty module is available in English, French, Italian, Spanish, Portuguese, Norwegian, and Chinese.[4,7-11] Recently, a Dutch version of the same questionnaires has also been published and validated.[4,11] Before conclusions can be drawn from pooled data across different countries, the data from the specific languages have to fit the Rasch model.

The Rasch model is named after Georg Rasch, a Danish mathematician.[12] The model shows what should be expected in responses to items if measurement (at the metric level) is to be achieved.[13] The model assumes that the probability of a given respondent affirming an item is a logistic function of the relative distance between the item location and the respondent location on a linear scale. The Rasch model is distinct from most statistical modeling because the aim is not to describe a set of data, rather, it is an "ideal" that the data should meet in order to provide successful measurement.[14] In the 1-parameter Rasch model, estimates for person and item measures (called person ability and item difficulty, respectively) are independent and are displayed as logits (the natural log odds of success vs failure). Specifically, the probability of a response is modeled as a logistic function of the difference between the person's level (eg, aesthetic satisfaction) and the level of satisfaction expressed by the item.[15] If a person's ability is known, it is possible to predict how that person is likely to perform on a given item. The logit scale represents the latent trait.[16] In this way, item analysis by the Rasch model can provide questionnaire scores that lie on a true equidistant-interval scale. Dichotomous and polytomous versions of the Rasch model are available.[7,14] In contrast to item response theory, Rasch modeling also possesses the property of invariant comparison. However, a fundamental assumption of this approach is that the response probability of each subject to each item is a function of the ratio of a person's ability (person parameter) to the item difficulty (item parameter).[12] Within the framework of Rasch measurement, the scale should also work in the same way, irrespective of which group (eg, gender) is being assessed,[17] ie, the probability of affirming an item is conditioned on the trait, and nothing else. If for some reason one gender did not display the same probability of affirming the item, then this item would be deemed to display differential item functioning (DIF) and would violate the requirement of invariance.[18,19] Thus, item analysis by the Rasch model examines whether questionnaires can be applied for invariant comparisons across different sample groups of individuals that are to be compared. [20, 21] A more detailed introduction to the Rasch model can be found in Boone et al.[22]

The primary objective of this study was to use item analysis by the Rasch model to evaluate the measurement equivalence of the FACE-Q rhinoplasty module in the Dutch-speaking community. If not properly evaluated, any cross-cultural variability might reveal a common assessment bias in different countries with the same instruments.[23]

METHODS

Study Design and Population

The study was designed and conducted according to the Declaration of Helsinki.[24] This protocol has been approved by the ethics committee of GZA Hospitals. A prospective, consecutive cohort of 100 Dutch-speaking patients eligible for SRP (including primary and revision rhinoplasty) was recruited from the ENT outpatient department of St-Vincentius Hospital, Antwerp, Belgium. Participants had to be older than 18 years and be proficient in the Dutch language. Subsequently, all patients underwent an external SRP for functional and/or aesthetic purposes. These 100 patients were invited to fill in the FACE-Q questionnaires about their satisfaction with their nose and the nostrils preoperatively and 3 months postoperatively as part of standard clinical care (N = 200 assessments for each questionnaire). The questionnaire was distributed on paper by one of the employees of the ENT department when patients arrived at the waiting room. To eliminate the Hawthorne effect, all participants filled in their questionnaires unattended before seeing the surgeon. Afterwards, data were transferred to an anonymized Excel file. The raw scores from the FACE-Q nose and nostrils scales were summed over the 10 (A-J) or 5 questions (a-e), respectively, and then transformed by conversion tables into a linear measure from 0 to 100. All patients eligible and willing to take part signed an informed consent form. Participants were also asked to answer questions that allowed us to characterize the sample, including age, gender, previous nasal surgery, previous history of nasal trauma, and ethnicity.

Statistical Analysis

All data were entered into the statistics package IBM SPSS Statistics version 27 (IBM Inc., Armonk, NY). In order to determine whether the translated Dutch version fitted the mathematical model of the original English version, a Rasch rating scale model (in which all items share the same rating scale) analysis was undertaken.[14] Data were analysed with Winsteps version 4.4.7 software.[25] Rasch analysis was performed separately on the FACE-Q nose and nostrils scales. The analysis included evaluations of overall model fit, reliability, threshold ordering, individual item and person fits, unidimensionality, targeting, local item dependency (LID), and DIF.

The Rasch analysis was conducted on all assessments (N = 200) as well as separately on pre- and postoperative responses (N = 100). Overall model fit, reliability, threshold ordering, individual item fits, and person fits were studied in the whole group as well individually on the pre- and postoperative assessments. DIF, LID, targeting and unidimensionality were investigated in the complete assessment group (N = 200).

For both the FACE-Q nose and nostrils scales, the pre- and postoperative subgroups, separately as well as combined, were used to construct new conversion tables based on the assessments of the Dutch-speaking patients. Then, the equivalence of both the original and the newly constructed conversion tables was investigated by receiver operating characteristics (ROC).

Overall Model Fit

The overall fit to the model was assessed by evaluating 3 overall fit statistics, specifically 2 item-person interaction statistics and 1 item-trait interaction. The overall item and person fit were evaluated by inspecting the mean item-person standardized fit residuals. Gross departures from a mean of 0.0 and a population standard deviation (P.SD) of 1.0 indicate that the data do not conform to the basic Rasch model specifications.[15] The chi-square statistic of the total item-trait interaction should be nonsignificant ($\chi^2 > 0.05$) reflecting homogeneity of the items among the different class intervals.[13]

Reliability

The person reliability measure estimates how well we can discriminate people based on their estimated ability.[26] High reliability for persons assumes that persons with high measures actually do have higher measures than persons estimated to have low measures. Low values of person reliability might indicate a persons' narrow ability range or may be related to the small number of items on the test. Person reliability is conceptually similar to Cronbach's α and was also calculated.

The item reliability measure indicates how well items can be discriminated from one another on the basis of their difficulty. Rasch analysis also provides so-called person and item separation indices; the former (PSI) reveals how well a set of items separates persons measured, and the latter reveals how well a sample of people is able to separate the items.[27] The PSI is also an indicator of the power of the generated fit statistics.[28] The person (item) separation estimates measure the spread of persons (items) in standard error units and should have values >1.0 for a useful measurement instrument.[25] It is suggested that α/PSI ≥ 0.9 = excellent; 0.9 > α/PSI ≥ 0.8 = good; 0.8 > α/PSI ≥ 0.7 = acceptable; 0.7 > α/PSI ≥ 0.6 = questionable; 0.6 > α/PSI ≥ 0.5 = poor; and α/PSI < 0.5 = unacceptable.[29,30] If the PSI is not acceptable, the top measure cannot be statistically distinguished from the bottom measure with any confidence and the obtained fit statistics may not be reliable because of too large an error variance.[31]

Threshold Ordering

Threshold ordering is part of the analysis by the Rasch model: it is the required monotonicity that implies that the transition from one score to the next is consistent with the increase in the latent variable. This analysis also identifies how reliably respondents can discriminate between response categories. Categories that operate well should have ordered structure calibration thresholds.[32] In disordered thresholds, there is a failure of respondents to use the response categories in a manner consistent with the level of the trait being measured.[13] This indicates that every category has a distinct probability of being selected more than any other category for a particular person difficulty. Furthermore, the step difficulties should advance between the recommended values of 1.4 and 5.0 logits, ensuring measurement stability.[33] The ordering of thresholds, which are points of crossover between adjacent response categories, was examined by the construction of a Wright map for the FACE-Q nose and nostrils scales separately.[34] A Wright map presents both item difficulties and person abilities arranged along the same logit scale.

The item difficulties are usually demonstrated by mapping the Rasch-Thurstone thresholds (step values) which are parameters of the Rasch rating scale.[35,36] A respondent whose position on the latent variable aligns with a Rasch-Thurstone threshold for an item has a 50% probability of being observed in the categories above the threshold and a 50% probability of being observed in the categories below the threshold.[35]

Individual Item and Person Fit

A scale should map out a clinically important construct. Fit residuals provide a valuable source of information. The difference between the response expected by the model and the observed response is called a residual. Therefore, to check whether an item fitted the model adequately, we examined the χ^2 fit statistics based on the log residuals, generally expected to be nonsignificant. A significant χ^2 fit statistic ($P < 0.05$) would indicate misfit to the item. The fit residuals and item calibration measures are represented as mean-square residuals (MNSQs) across items for each person or across persons for each item. If the data fit the model, the mean square and standardized fit indices should be close to 1.0 and 0.0, respectively, and the standardized *z* score should be <2.[37] For clinical purposes, the MNSQ should have values between 0.5 and 1.7, although values between 0.5 and 2.0 are still acceptable.[38] High item fit residuals signify underdiscrimination and might reflect multidimensionality and can be considered as noticeable off-variable noise. Low-fit residuals signify overdiscrimination and might reflect potential redundancy or item dependency within the item set: these data would be too predictable and therefore misleading.[15,39] Item fit was also examined via visual inspection of graphs of observed item responses for each class interval plotted against the model expectations, which are displayed as an item characteristic curve (ICC).[31]

Unidimensionality

A principal-component analysis (PCA) of item-response or person-response residuals is performed. Sample data exist for the Rasch dimension (explained variance) and unexplained variance (residuals). If the Rasch dimension is correct, the differences between observed and expected variances should be small.[15] PCA further decomposes the unexplained variance into random noise and other kinds of variance such as a secondary dimension and strands. According to Linacre, eigenvalues >>2 for the first contrast typically indicate the presence of multiple dimensions and associations between data.[15] If this is the case, the disattenuated person measure correlations should be evaluated: >0.7 indicates 1 dimension in the data, whereas <0.3 indicates multidimensionality.[15] This is formally tested by allowing the factor loadings on the first residual to determine the most divergent items. The items with the greatest positive and negative loadings on their first residual factor (resulting from PCA) were used to create the 2 subsets and then tested, by a paired *t* test, to see if the person estimates (the logit of person “ability” or, in this case, “rhinoplasty satisfaction”) derived from these subsets differ significantly from each other.[40,15] Unidimensionality can be inferred if ≤5% of the *t* tests are significant, or the lower bound of a binomial 95% CI of the observed proportion overlaps by 5%. To be valid, there must be at least 12 item response thresholds in each subset.[41]

Targeting

It is important, particularly in clinical practice, that the measures used are appropriately targeted at the population being assessed.[13] Sample targeting is also shown in the Wright maps because pre- and postoperative samples were plotted against the items. In relation to the operative status, the person-item map displays the location of person abilities and item difficulties, respectively, along the same latent dimension.

LID

A requirement for objective measurement is local independence of items. That is, when items demonstrate LID, also referred to as statistical dependency, it suggests that a response to one item was directly influenced by a response to another item.[42] Items that are locally dependent typically cause participants to provide biased/inaccurate responses, and hence the implications regarding validity evidence are quite significant.[43] LID was investigated by considering the residual correlation matrix. A high correlation of residuals for 2 items (or persons) may indicate that either they duplicate some feature of each other (ie, when a person’s response to an item depends on the response to another item) or they both incorporate some other shared dimension (ie, multidimensionality).[42] Residual correlations between items should be below 0.20 as a benchmark.[44] Higher values are considered to exhibit LID.[34,19] According to Linacre, only high positive correlations would have consequences for the construction of the questionnaire leading to the use of testlets.[15] A testlet is a packet of test items that are administered statistically together.[45]

DIF

The Rasch model was applied to examine whether the FACE-Q nose and nostrils scales can be used to make invariant comparisons across different sample groups

Table 1. Patient Characteristics

Variable	Total (N = 100 patients)
Age (years), mean [SD] (range)	29.7 [10.45] (18-61)
Gender, male/female ratio	37/63
Race/ethnicity, n (%)	
North European	52 (52%)
Mediterranean	37 (37%)
Other	1 (1%)
History of nasal surgery, n (%)	25 (25%)
History of nasal trauma, n (%)	43 (43%)

SD, standard deviation.

Table 2. Rasch Fitting Parameters for the FACE-Q Nose Scale

FACE-Q nose	Preoperative (N = 100)	Postoperative (N = 100)	Total (N = 200)
Mean item	0.00	0.00	0.00
P.SD	1.01	0.99	1.00
Log-likelihood χ^2	0.475	0.4892	0.6121
Person separation	2.06	2.11	3.30
Person reliability	0.81	0.82	0.92
Item separation	4.35	2.26	4.05
Item reliability	0.95	0.84	0.94
Cronbach α	0.84	0.91	0.95

N, number of assessments in each group; P.SD, population standard deviation.

Table 3. Rasch Fitting Parameters for the FACE-Q Nostrils Scale

FACE-Q nostrils	Preoperative (N = 100)	Postoperative (N = 100)	Total (N = 200)
Mean item	0.00	–0.01	0.00
P.SD	0.99	0.92	0.97
Log-likelihood χ^2	0.5967	0.4584	0.5460
Person separation	2.11	2.02	2.44
Person reliability	0.82	0.80	0.86
Item separation	0.48	2.12	1.23
Item reliability	0.19	0.82	0.60
Cronbach α	0.88	0.93	0.93

N, number of assessments in each group; P.SD, population standard deviation.

of individuals that are to be compared; if present, different demographic predictors should be identified.[46] In the present study, Rasch analysis examined if the items worked invariantly across operative status, gender, ethnicity, age, and previous history of nasal trauma or surgery. Within Winsteps, 2 DIF detection methods are used: (1) the Welch *t* test and (2) the Mantel-Haenszel method.[15,47] The Educational Testing Service (ETS) classification system was assigned according to the DIF contrast and the significance of the DIF statistics. If |DIF| equals or exceeds 0.64 logits, with a Mantel chi square value that is statistically significant ($P < 0.05$), the item is assigned a C rating. Type C items are considered to have DIF with the most magnitude. An item is assigned a B rating if |DIF| has a magnitude that equals or exceeds 0.43 logits and is less than 0.64 logits, with a Mantel chi square value that is significant ($P < 0.05$). Type B items are considered to have slight to moderate DIF. Any other test combination of test result with neither a likelihood chi-square value that is significantly different from 0 nor a |DIF| greater than 0.43 is classified with an A rating. An item classified as Type A demonstrates no evidence of DIF.[48,49,15]

Two forms of DIF exist: item responses may differ uniformly across the measured variable or the item responses may show nonuniform DIF, where differences in item responses between subgroups vary across the measured variable. Theoretically, uniform DIF can be resolved by splitting the item into group-specific items. Nonuniform DIF is usually removed because it reflects misfit to the model.[31] In Winsteps, uniform or nonuniform DIF is graphically examined with the DIF ICC.[15]

Finally, to transform the original raw scores into a linear interval scale, Rasch conversion tables based on the Dutch population were constructed for the FACE-Q nose and nostrils scales. The Dutch-based conversion tables were then compared by ROC analysis to the original English reference tables published by Klassen et al.[5]

RESULTS

The sample characteristics are summarized in Table 1. Our population included 37 men and 63 women, aged between 18 and 61 years. Their mean age was 29.7 years. All patients completed their questionnaires preoperatively and 3 months postoperatively (mean follow-up, 97 [8.9] days).

Overall Model Fit

Mean standardized fit residuals and P.SD for the FACE-Q nose and nostrils scales were all within normal limits for both pre- and postoperative participant subgroups as well

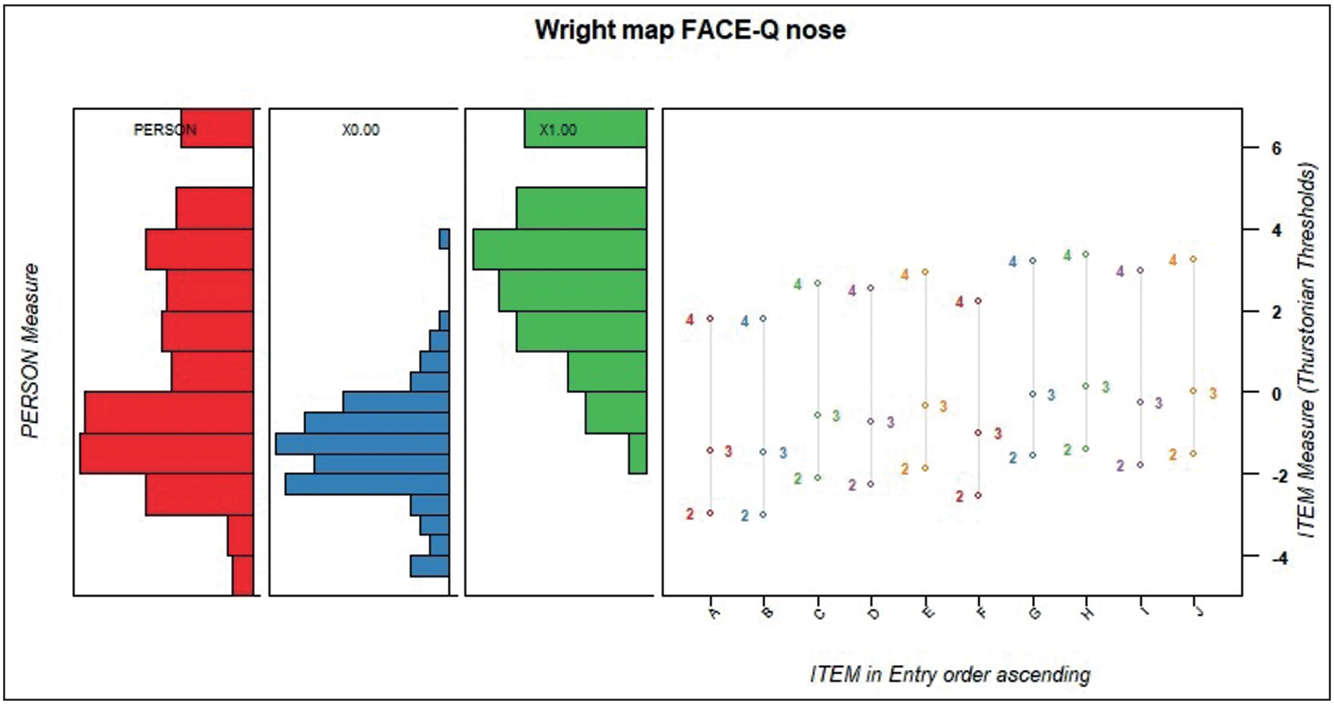

Figure 1. Person-item map for the FACE-Q nose questionnaire. Rasch-Thurstone thresholds are indicated by open circles and the cumulative probabilities are labeled as 2, 3, and 4. X0 represents the preoperative subgroup and X1 the postoperative subgroup.

as for the total group (Tables 2 and 3). The overall fit to the Rasch models was examined by looking at the log-likelihood χ^2. Nonsignificant values ($P > 0.05$) were found for both the FACE-Q nose and the FACE-Q nostrils scales.

Reliability

Tables 2 and 3 summarize the parameters determining the reliability of both questionnaires. Person reliability and separation indices as measured by Cronbach α levels and PSI demonstrated good to excellent values in all datasets.

All datasets also demonstrated good item reliability and meaningful item separation indices except in the preoperative population for the FACE-Q nostrils scale: the lower item separation index indicates a lower ability to consistently discriminate between persons, whereas a lower item reliability means that persons can discriminate less easily the items from one another on the basis of their difficulty.

Threshold Ordering

In the present evaluation, all items for both scales had ordered thresholds and categories (Figures 1 and 2). The analysis identified that the response categories from 0 to 4 were used as intended and displayed monotonicity. Regarding FACE-Q nose and nostrils items, there are 4 response categories, resulting in 3 Rasch-Thurstone thresholds. In all datasets, the step values are ordered in a convenient way over the measurement spectrum for all items. The Rasch person-item map also showed, for both questionnaires, no floor and ceiling effect. Step difficulties advanced between the recommended values of 1.4 and 5.0 logits for both questionnaires (FACE-Q nose, between 1.82 and 2.77; FACE-Q nostrils, between 3.81 and 4.07).

Individual Item and Person Fit

Tables 4 and 5 summarize the item fit for both questionnaires. Item difficulties change according to the pre- or postoperative status. For both questionnaires, the step difficulties advance between 1.1 and 5.44 logits, ensuring measurement stability.

The item location order along the trait in both questionnaires also changes as compared to the original English version. The summary fit statistics at the item and person level are called infit (inlier-pattern-sensitive) and outfit (outlier-sensitive) statistics and are reported as MNSQ or as standardized *z* score statistics. Infit is an

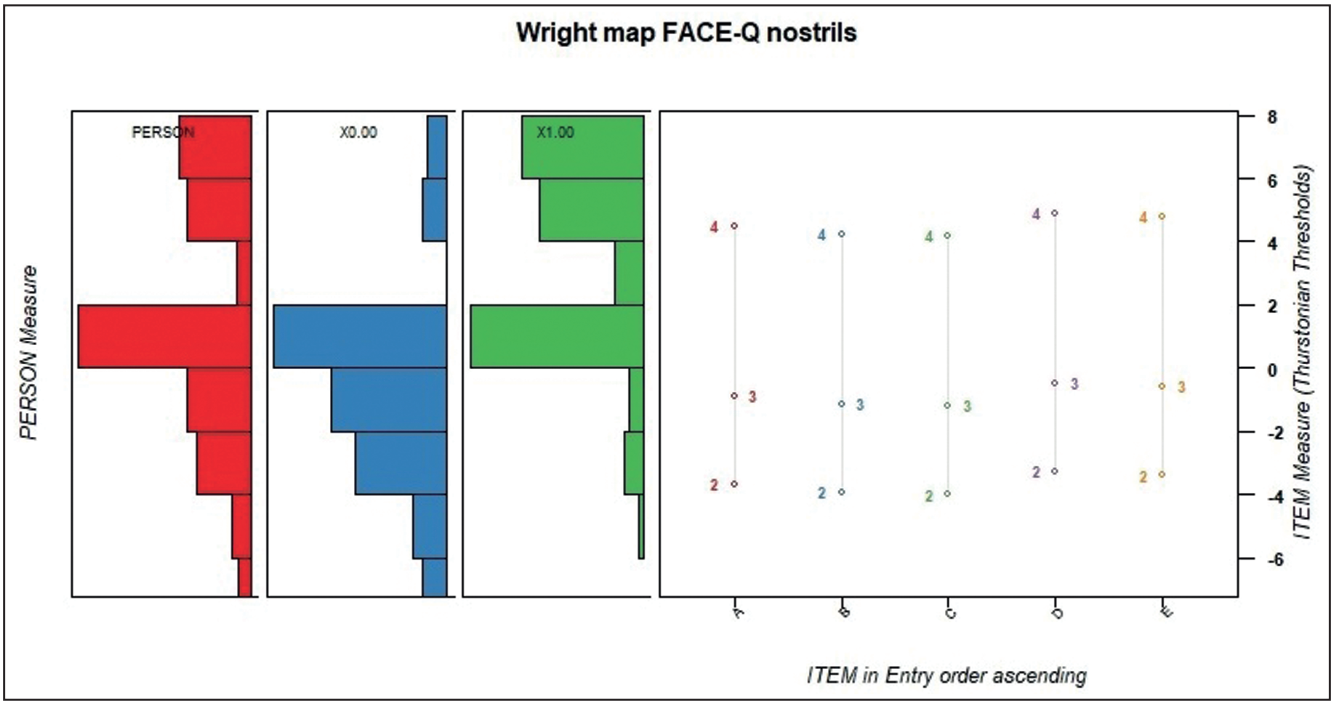

Figure 2. Person-item map for the FACE-Q nostrils questionnaire. Rasch-Thurstone thresholds are indicated by open circles and the cumulative probabilities are labeled as 2, 3, and 4. X0 represents the preoperative subgroup and X1 the postoperative subgroup.

information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's measure level. Outfit is an outlier-sensitive fit statistic, more sensitive to unexpected behavior by persons on items far from the person's measure level.[15] For the whole population, the outfit MNSQ for each item was reported to be within an acceptable range (between 0.5 and 1.7). Some more extreme misfits were found in the pre- and postoperative subgroup separately. In the preoperative subgroup, the MNSQ fit residual was 1.99 for item E of the FACE-Q nose scale and 0.48 for item b of the FACE-Q nostrils scale. High item fit residuals can be considered as noticeable off-variable noise. Low fit residuals signify overdiscrimination and might reflect potential redundancy or item dependency within the item set. In comparison with the pre- and postoperative datasets, the total group demonstrated the most statistically significant misfit items.

Unidimensionality

For each question of the FACE-Q nose and nostrils scales, at least 12 responses in each score category were found. PCA showed little evidence of multidimensionality. The differences between observed and expected variances were small, proving that the raw variance explained by measures of the Rasch dimension are correct. The eigenvalues of the first contrast were 2.0583 and 1.7458 for the FACE-Q nose and nostrils scales, respectively. In addition, the disattenuated person measure correlations between the first contrast clusters were within normal limits: 0.8687 for FACE-Q nose and 1.0000 for FACE-Q nostrils. The clusters with the greatest positive and negative loadings on their first residual factor (resulting from PCA) were used to create 2 subsets and then tested by a paired *t* test. For both questionnaires, no significant difference was found between these clusters, proving their merely unidimensional character ($P > 0.05$).

LID

Some positive residual correlations were present for the FACE-Q nose scale, but there were no significant correlations for the FACE-Q nostrils scale. Only the item pairs A/I (0.238), G/C (0.264), and B/F (0.265) on the FACE-Q nose scale were of statistical relevance (>0.20). Because these values were only slightly higher than the benchmark, no further adjustments with testlets were performed.

Targeting

The scale-to-sample targeting is shown in the Wright maps (Figures 1, 2). The results also provide evidence that the

Table 4. Rasch Measurement Theory Statistical Indicators of Fit for the FACE-Q Nose Scale

	FACE-Q nose	Preoperative group				Postoperative group				Total group			
Item	Questions	Item calibration	SE	Fit residual	*P* value	Item calibration	SE	Fit Residual	*P* value	Item calibration	SE	Fit Residual	*P* value
		(Average ability in logits)		Outfit MNSQ	χ^2	(Average ability in logits)		Outfit MNSQ	χ^2	(Average ability in logits)		Outfit MNSQ	χ^2
A	"Width at the bottom?"	–1.47	0.16	0.96	0.8110	0.29	0.21	1.13	0.4211	–0.86	0.13	1.53	**0.0015**
B	"Length?"	–0.89	0.15	1.05	0.7210	–1.10	0.25	0.981	**0.0010**	–0.90	0.13	0.97	0.851
C	"Bridge?"	0.15	0.16	1.3	0.7130	–0.35	0.23	1.4	**0.0614**	0.00	0.12	1.18	0.1312
D	"Suits your Face?"	–0.17	0.15	0.6	**0.0006**	–0.19	0.23	0.64	**0.0306**	–0.16	0.12	0.6	**0.0006**
E	"Straight?"	0.37	0.16	1.99	**0.0020**	0.00	0.22	1.49	**0.0115**	0.24	0.12	1.57	**0.0016**
F	"Overall size?"	–0.50	0.15	0.91	0.5609	–0.40	0.23	0.81	0.3408	–0.43	0.12	0.84	0.1808
G	"Shape in profile?"	1.11	0.18	0.7	0.1307	–0.40	0.23	0.8	0.3208	0.53	0.12	0.72	**0.0107**
H	"Looks in photographs?"	0.88	0.17	1.03	0.8110	0.55	0.21	0.7	**0.0407**	0.71	0.12	0.77	**0.0408**
I	"Tip?"	0.00	0.16	1.01	0.9110	0.89	0.20	1.24	0.1412	0.31	0.12	1.16	0.1712
J	"From every angle?"	0.52	0.16	0.61	0.1060	0.72	0.20	0.64	**0.0160**	0.57	0.12	0.59	**0.0006**

Items are in the original serial order of the English version. *P* values in bold are statistically significant ($P < 0.05$). MNSQ, mean square residual; SE, standard error.

Table 5. Rasch Measurement Theory Statistical Indicators of Fit for the FACE-Q Nose Scale

	FACE-Q nostrils	Preoperative group				Postoperative group				Total group			
Item	Questions	Item calibration	SE	Fit residual	*P* value	Item calibration	SE	Fit residual	*P* value	Item calibration	SE	Fit residual	*P* value
		(Average ability in logits)		Outfit MNSQ	χ^2	(Average ability in logits)		Outfit MNSQ	χ^2	(Average ability in logits)		Outfit MNSQ	χ^2
A	"Size?"	0.27	0.2	0.82	0.2908	–0.83	0.35	0.74	0.4907	–0.02	0.17	0.83	0.2608
B	"Shape?"	–0.18	0.2	0.74	0.1107	–0.59	0.34	0.48	0.1005	–0.28	0.17	0.69	**0.0207**
C	"Nostril show?"	–0.34	0.2	1.08	0.6311	–0.36	0.34	1.37	0.3314	–0.34	0.17	1.15	0.3311
D	"Well matched?"	–0.01	0.2	1.72	**0.0017**	1.46	0.32	1.06	0.7711	0.38	0.17	1.51	**0.0015**
E	"Look overall?"	0.27	0.2	0.54	**0.0005**	0.32	0.33	0.61	0.3006	0.27	0.17	0.51	**0.0005**

Items are in the original serial order of the English version. *P* values in bold are statistically significant ($P < 0.05$). MNSQ, mean square residual; SE, standard error.

2 scales define an adequate continuum for satisfaction with their appearance for pre- and postoperative patients undergoing rhinoplasty. Preoperative and postoperative patients (left-side histograms) are shown separately; the pattern of item-person scores of preoperative patients is clearly distinct from that of postoperative patients: the postoperative patients score considerably higher than the preoperative ones. For a good fitting model, we would expect that, for each of the items, respondents with high levels of the attribute being measured would endorse high-scoring responses, whereas individuals with low levels of the attribute would consistently endorse low-scoring responses.

DIF

In order to conduct an accurate comparison, a DIF analysis was performed on the complete dataset (N = 200) and stratified for age, operative status, gender, ethnicity, and history of trauma or previous surgery (Tables 6-9). A Mantel χ^2 and a Rasch-Welch test were performed for each demographic parameter. As the median age of this cohort of patients was

Table 6. DIF Contrasts by Gender, Age, and Nasal Trauma for the FACE-Q Nose Scale

	DIF contrast	Gender				Age	(<>27 yr)			Trauma			
	FACE-Q nose		Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test	
Item	Questions	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ2	Probability
A	"Width at the bottom?"	0.65*	0.0162*	1.4475*	0.2289*	–0.17	0.5200	1.2470	0.2641	–0.17	0.5154	0.0724	0.7879
B	"Length?"	0.12	0.6583	1.5434	0.2141	–0.37	0.1555	2.0272	0.1545	–0.09	0.7297	0.1674	0.6824
C	"Bridge?"	–0.45*	0.0872*	1.6297*	0.2017*	–0.04	0.8639	0.0390	0.8434	–0.54*	0.0321*	3.1637*	0.0753*
D	"Suits your Face?"	–0.22	0.4033	1.4184	0.2337	0.00	1.000	0.1236	0.7251	0.08	0.7628	0.5574	0.4553
E	"Straight?"	0.45*	0.0841*	0.3720*	0.5419*	0.20	0.4276	0.2706	0.6029	–0.19	0.4499	0.2484	0.6182
F	"Overall size?"	0.06	0.8136	0.4024	0.5259	–.017	0.4945	0.3462	0.5563	0.50*	0.0517*	3.6991*	0.0544*
G	"Shape in profile?"	–0.40	0.1263	0.1404	0.7079	0.00	1.000	0.0039	0.9501	0.11	0.6528	0.0575	0.8105
H	"Looks in photographs?"	0.03	0.9082	0.0114	0.9151	0.28	0.2649	3.0800	0.0793	–0.08	0.7504	0.5447	0.4605
I	"Tip?"	–0.47*	0.0745*	4.4421*	0.0351*	0.04	0.8628	0.0692	0.7925	0.19	0.4508	0.9955	0.3184
J	"From every angle?"	0.26	0.3136	4.0162	0.0451	0.19	0.4487	1.3924	0.2380	0.19	0.4479	1.0465	0.3063

DIF categories are staged according to ETS. The category into which a question will be placed depends on two factors: the absolute value of DIF contrast and whether or not the value is statistically significant (Mantel χ^2 and Rasch-Welch tests). *P* values <0.05 were considered as significant. ETS classification: category B: *slight to moderate DIF; all other items are classified as category A with negligible DIF. DIF, differential item functioning; ETS, Educational Testing Service.

Table 7. DIF Contrasts by Nasal Revisions, Ethnicity, and Operative Status for the FACE-Q Nose Scale

DIF	DIF contrast	Revisions				Ethnicity				Operative status			
	FACE-Q nose		Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test	
Item	Questions	DIF contrast	Probability	χ2	Probability	DIF contrast	Probability	χ2	Probability	DIF contrast	Probability	χ2	Probability
A	"Width at the bottom?"	0.00	1.000	0.0158	0.9001	–0.29	0.2707	1.1505	0.2834	–1.71*	0.0000*	3.3253*	0.0682*
B	"Length?"	0.20	0.5100	0.2000	0.6548	–0.29	0.2704	2.4777	0.1155	0.12	0.6781	0.0699	0.7914
C	"Bridge?"	0.31	0.2725	0.4619	0.4968	0.69**	0.0065**	5.2224**	0.0223**	0.47*	0.0797*	0.3250*	0.5686*
D	"Suits your Face?"	–0.21	0.4542	2.8667	0.0904	–0.03	0.8956	0.3779	0.5387	0.00	1.000	3.9359	0.0473
E	"Straight?"	–0.59*	0.0350*	1.1019*	0.2939*	–0.11	0.6720	0.0303	0.8617	0.37	0.1579	1.7494	0.1860
F	"Overall size?"	0.22	0.4575	0.6023	0.4377	–0.43*	0.0883*	2.7757*	0.0957*	–0.11	0.6834	1.1548	0.2825
G	"Shape in profile?"	0.02	0.9420	0.0706	0.7905	0.24	0.3326	0.6801	0.4096	1.45**	0.0000**	6.6022**	0.0102**
H	"Looks in photographs?"	–0.10	0.7109	1.1018	0.2939	0.23	0.3572	1.5993	0.2060	0.37	0.1548	0.0800	0.7773
I	"Tip?"	0.12	0.6614	0.4959	0.4813	–0.21	0.4071	0.6165	0.4324	–0.78*	0.0017*	2.6809*	0.1016*
J	"From every angle?"	0.09	0.7396	0.0686	0.7934	0.15	0.5607	0.9923	0.3192	–0.13	0.6170	0.9863	0.3207

DIF categories are staged according to ETS. The category into which a question will be placed depends on two factors: the absolute value of DIF contrast and whether or not the value is statistically significant (Mantel χ^2 and Rasch-Welch tests). *P* values <0.05 were considered as significant. ETS classification: category B: *slight to moderate DIF; category C: **moderate to large DIF; all other items are classified as category A with negligible DIF. DIF, differential item functioning; ETS, Educational Testing Service.

Table 8. DIF Contrasts by Gender, Age, and Nasal Trauma for the FACE-Q Nostrils Scale

						Age	(<>27 yr)			Trauma			
	FACE-Q nostrils		Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test	
Item	Questions	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ^2	Probability
A	"Size?"	0.09	0.8042	0.0042	0.9486	–0.09	0.7851	0.0008	0.9774	0.18	0.6129	0.0836	0.7725
B	"Shape?"	–0.64**	0.0787**	4.4423**	0.0351**	0.09	0.7998	0.1907	0.6623	0.20	0.5688	1.4831	0.2233
C	"Nostril show?"	0.87**	0.0160**	6.5169**	0.0107**	0.09	0.7974	0.0369	0.8477	0.10	0.7842	0.5445	0.4606
D	"Well matched?"	–0.18	0.6032	0.0840	0.7720	–0.21	0.5345	0.3363	0.5620	–0.75*	0.0275*	2.5852*	0.1079*
E	"Look overall?"	–0.12	0.7252	0.3772	0.5391	0.13	0.6989	0.6414	0.4232	0.31	.3569	3.9213	0.0477

DIF categories are staged according to ETS. The category into which a question will be placed depends on two factors: the absolute value of DIF contrast and whether or not the value is statistically significant (Mantel χ^2 and Rasch-Welch tests). *P* values <0.05 were considered as significant. ETS classification: category B: *slight to moderate DIF; category C: **moderate to large DIF; all other items are classified as category A with negligible DIF. DIF, differential item functioning; ETS, Educational Testing Service.

Table 9. DIF Contrasts by Nasal Revisions, Ethnicity, and Operative Status for the FACE-Q Nostrils Scale

		Revisions				Ethnicity				Operative status			
	FACE-Q nostrils		Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test			Rasch-Welch test	Mantel test	
Item	Questions	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ^2	Probability	DIF contrast	Probability	χ^2	Probability
A	"Size?"	–0.10	0.8094	0.2533	0.6147	–0.07	0.8428	0.0635	0.8011	0.88*	0.0134*	2.2858*	0.1306*
B	"Shape?"	0.37	0.3690	0.5244	0.4690	0.21	0.5498	0.7963	0.3722	0.21	0.5540	0.0189	0.8907
C	"Nostril show?"	–0.53*	0.1901*	0.0548*	0.8149*	–0.64*	0.0695*	1.0088*	0.3152*	–0.14	0.6965	0.0047	0.9456
D	"Well matched?"	0.27	0.4868	0.3663	0.5450	0.56*	0.0978*	0.2584*	0.6112*	–0.99*	0.0041*	1.5460*	0.2137*
E	"Look overall?"	–0.02	0.9579	0.8655	0.3522	–0.10	0.7589	0.0123	0.9117	0.11	0.7407	0.0589	0.8083

DIF categories are staged according to ETS. The category into which a question will be placed depends on two factors: the absolute value of DIF contrast and whether or not the value is statistically significant (Mantel χ^2 and Rasch-Welch tests). *P* values <0.05 were considered as significant. ETS classification: category B: *slight to moderate DIF; all other items are classified as category A with negligible DIF. DIF, differential item functioning; ETS, Educational Testing Service.

27 years, the data were dichotomized below and above the median. According to the ETS classification, moderate to high DIF values (category C) were detected for gender in the FACE-Q nostrils scale and for ethnicity and operative status in the FACE-Q nose scale; slight to moderate DIF values were found for all demographic indicators except for age.

For the FACE-Q nose scale, significant DIF contrasts were detected for ethnicity in item C ("Bridge?") and for operative status in item G ("Shape in profile?") (Tables 6 and 7). For the FACE-Q nostrils scale, significant DIF contrasts were seen for gender in item b "Shape?" and item c "Nostril show?" (Tables 8 and 9). All items with category C DIF contrasts showed at least some degree of nonuniform DIF in their ICCs (Figure 3).

ROC Analysis of Rasch Conversion Tables

Starting from the preoperative, postoperative, and combined datasets, new Rasch person score tables for the Dutch FACE-Q nose and nostrils scales were constructed. These newly created conversion tables were made so that the sum of the raw scores of patients is transformed into a logit scale from 0 to 100. Then, these person score tables were compared with each other and with the original English version.

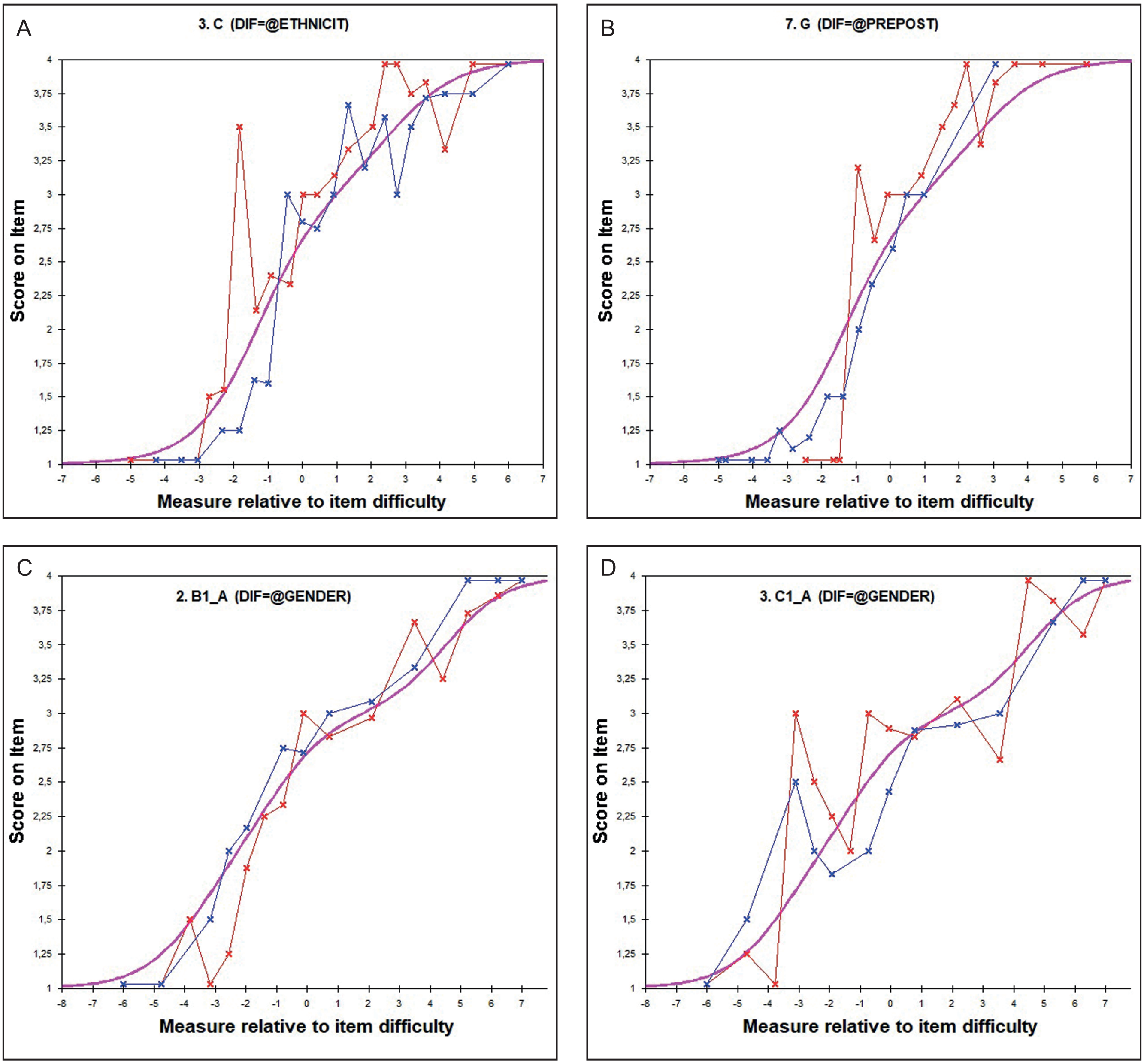

Figure 3. Item characteristic curves of items of Educational Testing Service differential item functioning category C. The pink curve represents the Rasch model for the item. (A) Item C: "Bridge?" (red curve, North European; blue curve, Mediterranean). (B) Item G: "Shape nose?" (red curve, postoperative; blue curve, preoperative). (C) Item b: "Shape nostrils?" (red curve, female; blue curve, male). (D) Item c: "Nostril show?" (red curve, female; blue curve, male).

The Pearson correlation coefficient between the conversion tables was highly significant (0.994-0.999 for the FACE-Q nose scale and 0.995-0.999 for the FACE-Q nostrils scale). Linear regression also demonstrated a very high correlation between the English Rasch and all 3 Dutch conversion tables. The determination coefficient r^2 was considered highly significant with values between 0.988 and 0.998 for the FACE-Q nose scale and between 0.994 and 0.997 for the FACE-Q nostrils scale. Analysis of variance for equal means revealed no significant difference (P = 0.945) between the English and Dutch conversion scales for the FACE-Q nose and nostrils scales.

Finally, the different versions of the Rasch transformation tables were applied to the present patient cohort of 100 patients and the resulting Rasch scores evaluated by ROC analysis (Figure 4A,B). In both languages, the FACE-Q nose and nostrils questionnaires had an excellent classifier output quality. The area under the curve (AUC) fitting for all conversion tables of the FACE-Q nose scale was exactly the same (P = 1). The AUC was 0 9592. The AUC fittings for

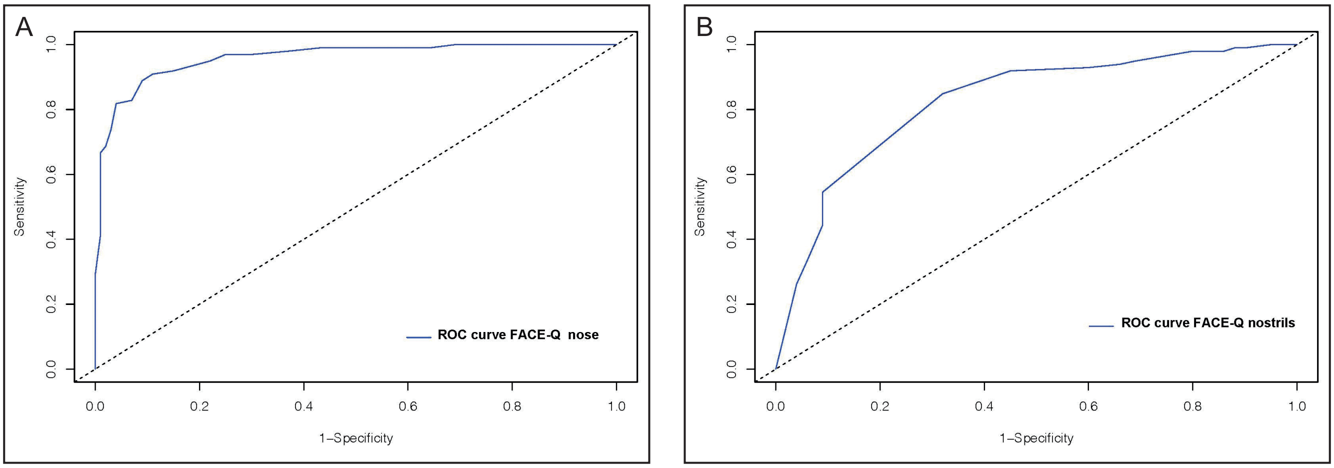

Figure 4. Receiver operating characteristics comparison of the FACE-Q nose (A) and nostrils (B) scales with different language versions of the Rasch transformation tables. The English and Dutch conversion tables were applied to the preoperative, postoperative and global datasets. As would expected from the Rasch theory, the receiver operating characteristics were exactly the same for all datasets (P = 1) with either the English or Dutch language version of the transformation tables.

all conversion tables of the FACE-Q nostrils scale proved to be exactly the same (P = 1). The AUC was 0.8253. In other words, the ROCs were independent of the datasets and of the languages.

DISCUSSION

PROMs are increasingly being used to incorporate the perspective of the patient into outcome assessments.[1,6,7] The FACE-Q rhinoplasty module is a rigorously developed PROM based on Rasch analysis in an English-speaking population. To the best of our knowledge, this is the first study that has used Rasch analysis to evaluate the equivalence of the FACE-Q rhinoplasty module in a different language. Our study aimed to examine how the structure of the FACE-Q nose and nostrils scales behaved when administered in this clinical Dutch-speaking sample who attended the public healthcare system in Belgium. In other words, this study empirically examined to what extent the item statistics from the Rasch model varied across a different culture and language for the same questions.

As the purpose of this study was essentially exploratory and not intended to modify the scale structure, sample size could be limited to 200 assessments. A sample size of at least 64 to 144 persons is required to achieve 95% confidence that the item calibration is within ±0.5 logits.[50] For definitive decisions about items, such as deleting items or collapsing response categories, several simulation studies have demonstrated that a sample size of around 250 to 500 will usually be needed to provide accurate and stable person and item estimates as well as a good balance for statistical interpretation of the fit statistics.[51, 52]

Validation of questionnaires in other languages is important to ensure their conceptual equivalence with the original questionnaire. The psychometric properties of a questionnaire can be affected by language transference and/or cultural differences by affecting the difficulty of the test, the range of abilities that it taps, and/or the respondents' ability to select appropriate answers. Thus, measurement error could potentially have serious implications in the context of cross-cultural research. Validation helps researchers to generalize their findings cross-culturally.

In the present study, Rasch analysis for the FACE-Q nose and nostrils scales in the Dutch-speaking population supported the responsiveness, reliability, and validity of these measures. Overall statistics indicated that both scales fit the Rasch model well: there is overall invariance of item difficulty across the scales, indicating Rasch modeling can be used in a proper way.[53,39] In-depth Rasch analysis showed minor differences in performance between the Dutch and English versions of the questionnaire. Regarding fit statistics, misfits were found in many questions for the FACE-Q nose and nostrils scales, but MNSQ fit residuals were still within the range of acceptability (MNSQ, 0.5-2.0).[15] The item reliability and item separation index were considerably lower for the FACE-Q nostrils scale in the preoperative population. Statistically, this may be due to the smaller number of questions in the FACE-Q nostrils scale compared with the FACE-Q nose scale. On clinical grounds, this finding can also be attributed to the preoperative unawareness of patients regarding nostril deformities. Often, once patients are operated, they start to look more closely to their nostrils in the mirror, or nowadays in selfies.

LID analysis demonstrated some correlation issues for the FACE-Q nose scale but had no consequential dependency because the values were just above the benchmark. According to Seyed and Soodeh, evidence of LID may not necessarily be attributed to substantive items that are similar/related, but instead can be attributed to the order and proximity in which they were presented to participants.[54] Item order should be randomized when possible because item order effects, often called assimilation effects or carry-over effects, can result in biased participant responses .[55]

DIF analysis demonstrated bias issues especially for items C ("Bridge?") and G ("Shape in profile?") of the FACE-Q nose scale, and items b ("Shape?") and c ("Nostril show?") of the FACE-Q nostrils scale. From a patient's perspective, item difficulty means that they will under- or overrate their response more than is expected.

In the lower item difficulty range of item C ("Bridge?"), patient scores of Mediterranean patients tended to be underrated whereas North European patient scores tended to be overrated. This kind of bias can be explained by patients having an ethnically based difference in awareness on their nasal hump. Item G functioned differentially depending on the pre- or postoperative status with more extreme item scores in the postoperative phase. Among Caucasians, for most aesthetic rhinoplasty patients, the dorsal profile is often one of their main concerns and the most important reason for demanding a nose correction.[56] Therefore, their responses may be exaggerated in both ways, good or bad, especially in the postoperative period. Due to DIF, item G performance does not remain stable across pre- and postoperative subgroups. Hence, the surgeon should be aware that, during the follow-up of SRP patients, the operative status may influence the test-retest reliability of the FACE-Q nose scale. Klassen et al found no significant DIF for age, gender, ethnicity, or country, but, in contrast to the present study, DIF was not examined for operative status.[5]

For both items b and c of the FACE-Q nostrils scale, women are more concerned about the unsightly aspect of their nostrils than men are, but for both genders it is not a factor of extreme beauty either. Consequently, these DIF contrasts help to understand how patients requesting a rhinoplasty experience their facial appearance, and more specifically their noses, differently. However, the effect of DIF in certain patient subgroups may bias the outcome measures and should always be kept in mind by the surgeon.

As demonstrated by the ROC analysis, Rasch modeling creates linear interval data that are independent of the datasets and of the languages. In the present study, raw ordinal data obtained by the Dutch version of the FACE-Q rhinoplasty module can safely be converted into linear measures by the use of the original English Rasch transformation tables developed by Klassen et al.[5] Therefore, fitting data to the Rasch model offers an elegant approach to addressing several key methodologic aspects associated with scale development and construct validation, as well as providing a linear transformation of the ordinal raw score.

Further studies with other language translations of FACE-Q are necessary to evaluate their multicultural equivalence with the original English version. Care should also be taken to generalize the DIF results in this study because the sample was drawn from a single facility. Further DIF evaluations in other cultures and with other language translations may reveal different population characteristics. As already mentioned, the psychometric properties of questionnaires can be seriously affected by measurement errors based on cultural differences. However, by means of Rasch analysis, the DIF among other cultures can be statistically evaluated. In this way, Rasch analysis can make multicultural comparisons of rhinoplasty satisfaction more meaningful.

CONCLUSIONS

The Dutch language version of the FACE-Q rhinoplasty module is an adequate instrument for determining successful aesthetic surgery based on patient satisfaction. This tool is able to measure the degree of success with respect to the patient and is a valuable assessment tool for the surgeon in both pre- and postoperative contexts. We hope that this construct validation will provide an additional aid to clinicians evaluating Dutch-speaking SRP patients.

Disclosures

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

REFERENCES

1. Saleh H, Apaydin F. Outcomes in rhinoplasty. *Facial Plast Surg.* 2019;35(1):47-52.
2. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol.* 1993;46(12):1417-1432.
3. Klassen AF, Cano SJ, Scott A, Snell L, Pusic AL. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. *Facial Plast Surg.* 2010;26(4):303-309.
4. Memorial Sloan Kettering Cancer Center. Q-Portfolio. http://qportfolio.org/FACE-q/aesthetics/. Accessed January 4, 2021.
5. Klassen A, Cano S, East C, et al. Development and psychometric evaluation of the FACE-Q scales for

patients undergoing rhinoplasty. *JAMA Facial Plast Surg.* 2016;18(1):27.

6. van Zijl F, Mokkink L, Haagsma J, Datema F. Evaluation of measurement properties of patient-reported outcome measures after rhinoplasty. *JAMA Facial Plast Surg.* 2019;21(2):152.
7. Barone M, Cogliandro A, Di Stefano N, Tambone V, Persichetti P. A systematic review of patient-reported outcome measures after rhinoplasty. *Eur Arch Otorhinolaryngol.* 2016;274(4):1807-1811.
8. Barone M, Cogliandro A, Di Stefano N, Aronica R, Tambone V, Persichetti P. Linguistic validation of the "FACE-Q rhinoplasty module" in Italian. *Eur Arch Otorhinolaryngol.* 2016;274(3):1771-1772.
9. Radulesco T, Penicaud M, Santini L, Graziani J, Dessi P, Michel J. French validation of the FACE-Q rhinoplasty module. *Clin Otolaryngol.* 2018;44(3):240-243.
10. Kalaaji A, Dreyer S, Schnegg J, Sanosyan L, Radovic T, Maric I. Assessment of rhinoplasty outcomes with FACE-Q rhinoplasty module: Norwegian linguistic validation and clinical application in 243 patients. *Plast Reconstr Surg Global Open.* 2019;7(9):e2448.
11. Pingnet L, Verkest V, Fransen E, Declau F. Validation of the Dutch FACE-Q rhinoplasty module. *Facial Plast Surg.* 2021. Doi:10.1055/s-0040-1721099.
12. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press; 1993.
13. Pallant J, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol.* 2007;46:1-18.
14. Andrich D. A rating formulation for ordered response categories. *Psychometrika.* 1978;43(4):561-573.
15. Linacre J. A User's Guide to Winsteps/Ministeps Rasch Model Computer Programs (Program Manual 4.4.7). http://www.winsteps.com/index.htm. Accessed April 12, 2020.
16. Tennant A, Conaghan P. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum.* 2007;57(8):1358-1362.
17. Rasch G. In: Rasch G, ed. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press; 1960.
18. Holland PW, Wainer H. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum; 1993.
19. Smith R. Fit analysis in latent trait measurement models. *J Appl Measure.* 2000;1(2):199-218.
20. Hagquist C, Andrich D. Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Pers Individ Dif.* 2004;36:955-968.
21. Andrich D. Rasch models for ordered response categories. In: Everitt B, Howell D, eds. *Encycopedia of Statistics in Behavioral Science,* vol 4. Chichester: John Wiley & Sons; 2005:1698–1707.
22. Boone W, Staver JR, Yale MS. *Rasch Analysis in the Human Sciences.* Dordrecht: Springer; 2014.
23. Tennant A, Pentra M, Tesio L. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care*. 2014;42(1):137-148.
24. World Medical Association Declaration of Helsinki. Wma.net. 1996. https://www.wma.net/wp-content/uploads/2016/11/DoH-Oct1996.pdf. Accessed January 4, 2021.
25. Linacre JM. Winsteps® (version 4.8.0) [computer software]. https://www.winsteps.com/. Beaverton, OR: JM Linacre; 2019. Accessed March 17, 2021.
26. Bond T, Fox C. Applying the Rasch model: fundamental measurement in the human sciences. In: Bond T, Fox C, eds. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum; 2001.
27. Wright B, Stone M. *Best Test Design*. Chicago: MESA Press; 1979.
28. Vincent J, MacDermid J, King G, Grewal R. Rasch analysis of the patient-rated elbow evaluation questionnaire. *Health Qual Life Outcomes*. 2015;13:84.
29. Cronbach L, Warrington W. Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika.* 1951;16(2):167-188.
30. Fisher W. Reliability statistics. *Rasch Meas Trans.* 1992;6:238.
31. Andrich D, Marais I. *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences.* Singapore: Springer; 2019.
32. Court H, Greenland Margrain TH. Measuring patient anxiety in primary care: Rasch analysis of the 6-item Spielberger State Anxiety Scale. *Value Health.* 2010;13(6):813-819.
33. Linacre J. *Midwest Objective Measurement Seminar, #2.* Chicago: Institute for Objective Measurement, Inc; 1997.
34. Wright B, Masters G. In: Wright B, Masters G, eds. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press; 1982.
35. Linacre J. Item discrimination and Rasch-Andrich thresholds. *Rash Meas Trans*. 2006;20(1):1054.
36. Simpelaere I, Vanderwegen J, Wouters K, De Bodt M, Van Nuffelen G. Feasibility and psychometric properties of the adjusted DSWAL-QoL questionnaire for dysphagic patients with additional language and/or cognitive impairment: part I. *Dysphagia.* 2017;32:401-419.
37. Frantom C, Green K, Lam T. Item grouping effects on invariance of attitude items. *J Appl Meas.* 2002;3(1):38-49.
38. Wright B, Linacre J. Reasonable mean-square fit values. *Rasch Meas Trans.* 1994;8(3):370.
39. Hansen T, Kjaersgaard A. Item analysis of the Eating Assessment Tool (EAT-10) by the Rasch model: a secondary analysis of cross-sectional survey data obtained among community-dwelling elders. *Health Qual Life Outcomes.* 2020;18:139.
40. Smith E. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas.* 2002;3(2):205-231.

41. Hagell P. Testing rating scale unidimensionality using the principal component analysis (PCA)/*t*-test protocol with the Rasch model: the primacy of theory over statistics. *Open J Stat.* 2014;4:456-465.
42. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas.* 2008;9(3):200-215.
43. Royal K. The impact of item sequence order on local item dependence: an item response theory perspective. *Surv Pract.* 2016;9(5):1-7.
44. Christensen K, Makransky G, Horton M. Critical values for Yen's Q 3: identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas.* 2017;41(3):178-194.
45. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: a case for testlets. *J Educ Meas.* 1987;24:185-201.
46. Schwitzer J, Albino F, Matthis R, Scott A, Gamble L, Baker S. Assessing demographic differences in patient-perceived improvement in facial appearance and quality of life following rhinoplasty. *Aesthet Surg J.* 2015;35(7):784-793.
47. Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J Amer Stat Assoc.* 1963;58:690-700.
48. Zwick R. When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *J Educ Stat.* 1990;15(3):185-197.
49. Zwick R. A review of ETS differential item functioning assessment procedures: flagging rules, minimum sample size requirements, and criterion refinement. 2012. https://www.ets.org/Media/Research/pdf/RR-12-08.pdf. Accessed January 4, 2021.
50. Linacre J. Sample size and item calibration stability. *Rasch Meas Trans.* 1994;7(4):328.
51. Hagell P, Westergren A. Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (RUMM) program in health outcome measurement. *J Appl Meas.* 2016;7(4):416-431.
52. Chen W, Lenderking W, Jin Y, Wyrwich W, Gelhorn H, Revicki D. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Qual Life Res.* 2014;23(2):485-489.
53. Kukull W, Larson E, Teri L, Bowen J, McCormick W, Pfanschmidt M. The Mini-Mental State Examination Score and the clinical diagnosis of dementia. *J Clin Epidemiol.* 1994;47(9):061-067.
54. Seyed M, Soodeh B. Differential item functioning analysis of high-stakes test in terms of gender: a Rasch model approach. *Malays Online J Educ Sci.* 2017;5(1):10-24.
55. Heiman G. *Research Methods in Psychology.* Boston, MA: Houghton Mifflin; 2002.
56. Rowe-Jones J, van Wyk FC. Special considerations in Northern European primary aesthetic rhinoplasty. In: Sclafani AP, ed. *Rhinoplasty. The Experts' Reference.* New York: Thieme; 2015:538-546.