Development, validation, application and abbreviation of an international thyroid-related quality of life patient-reported outcome measure

Torquil Watt
In loving memory of Lars Thorgaard
Development, validation, application and abbreviation of an international thyroid-related quality of life patient-reported outcome measure

Doctoral Thesis
Torquil Watt
This thesis consists of the following 8 papers:


#6 Watt T, Hegedüs L, Bjorner JB, Groenvold M, Bonnema SJ, Rasmussen ÅK, Feldt-Rasmussen U. Is Thyroid Autoimmunity per se a Determinant of Quality of Life in Patients with Autoimmune Hypothyroidism? European Thyroid Journal 2012; 1: 186-192.


Part of the introduction is based on


and

The work on which this thesis is built was carried out during my occupation as post doc at Department of Medical Endocrinology, Rigshospitalet, led by professor Ulla Feldt-Rasmussen and Head of Department Åse Krogh Rasmussen. Thank you for your support and guidance during all the years we have shared.

My path as a scientist was laid down many years earlier, when I first came across Jakob Bjørner and Mogens Grønvold in the early 90's, and has been lid ever since by your indigenous scientific insight, thoughtfulness and creativity. Thank you for your inspiration and friendship.

Companion inspirators on the latter part of this voyage have been Laszlo Hegedüs and Steen Bonnema. Thank you for your engagement and impressive workability during these many years of scientific joint venture.

Many scientific knots within this thesis were tied during my time in Massachusetts with the group of John Ware, including Barbara Gandek. Thank you for having me there and for your inspiration.

I also wish to thank those who helped making it possible to write this thesis during my specialty training; thank you Tina Vilsbøll, Henrik Ullits, Ebbe Winther and Finn Rønholt.

Please also see Acknowledgements on page 64.

Finally I want to thank my wife and children, Line, Selma, Naja and Sylvester for bearing with my absence; and moreso with my absentmindedness, which has at times of difficult analyses or demanding phases been debilitating, I am afraid. I know you acknowledge that my work as a scientist is borne by a deeply felt conviction, that acquiring greater scientific insight, as an act in itself, is at the core of being a human being, as is performing music or writing poetry.
LIST OF ABBREVIATIONS

CAS  Clinical activity score
CFA  Confirmatory factor analysis
CI   Confidence Interval
CNS  Central nervous system
CTQ  Chronic thyroid questionnaire
DIF  Differential item functioning
GO   Graves' orbitopathy
GO-QLS Graves' ophthalmopathy quality of life scale
GOQOL Graves' orbitopathy quality of life instrument
GRASS Graves' disease selenium supplementation trial
HCQ  Hyperthyroid complaints questionnaire
HSS  Hyperthyroid symptoms scale
IRT  Item response theory
MID  Minimal important difference
MR   Magnetic resonance
NOSPECS A clinical grading system for Graves' orbitopathy
PRO  Patient-reported outcome
QoL  Quality of life
RCT  Randomized clinical trial
SF-36 Short Form (36) Health Survey
TED-QOL Thyroid eye disease quality of life scale
ThyDQoL Underactive thyroid-dependent quality of life questionnaire
ThyPRO Thyroid-related patient reported outcome
ThyPRO-39 Thyroid-related patient reported outcome, 39-item version
ThyQoL Thyroid quality of life project
ThySRQ Thyroid symptom rating questionnaire
ThyTSQ Thyroid treatment satisfaction questionnaire
TPO-Ab Thyroid peroxidase antibodies
TSH  Thyroid stimulating hormone
TSQ  Thyroid symptoms questionnaire
INTRODUCTION

The thyroid is a gland located in the anterior, lower part of the neck. It produces thyroid hormones, essential for normal cell function and metabolism in all human tissues. Diseases related to the thyroid gland are common, affecting around 10-15% of the adult population in most countries (1-3). The spectrum of benign thyroid diseases relates to the structure and/or function of the gland. Structural changes involve enlargement (goiter) and functional changes involve either under- or over-functioning of the gland, i.e. hypo- or hyperthyroidism, respectively.

In recent years, a paradigm shift has occurred in medical care. Patient-centered care has become the goal in most clinical settings and the value of an intervention is increasingly judged by its ability to maximize health, with an emphasis on the degree to which longevity and health-related quality of life (QoL) are impacted. QoL assesses the patient’s point of view concerning how they feel and what they are able to do in everyday life. It has been defined as the subjective assessment of the impact of disease and its treatment across the physical, psychological, social, and somatic domains of functioning and well-being (4, 5). QoL is measured by standardized questionnaires (patient-reported outcomes (PRO)), where responses to questions are converted to and summarized in numeric values that quantify the attributes being measured. Typically, multi-item scales are used, where several items measuring the same construct are summarized into one scale, to reduce random measurement error and to simplify reporting. Findings are generally ordered within three main domains: physical, mental and social/participation, in accordance with the definition of health-related quality of life (6).

For QoL assessments to be valid, it is essential to apply tools with appropriate measurement properties for the population and research question under study. The International Society of Quality of Life Research recently recommended minimum requirements for patient-reported outcomes used in patient-centred outcomes and comparative effectiveness research (7). The definitions and recommendations are presented in Table 1, with permission.

QoL is impacted by many factors. In addition to the underlying disease pathology and symptoms, individual patient characteristics and environmental factors interact with patients' functional status and perceptions of their health, to influence the degree to which diseases and treatment affect QoL (8).
<table>
<thead>
<tr>
<th>Measurement property</th>
<th>Definition</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual and measurement model</td>
<td>The conceptual model provides a description and framework for the targeted construct(s) to be included in a PRO measure. The measurement model maps the individual items in the PRO measure to the construct</td>
<td>A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use. In addition, there should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts included in the PRO measure</td>
</tr>
<tr>
<td>Reliability</td>
<td>The degree to which a PRO measure is free from measurement error</td>
<td>The reliability of a PRO measure should preferably be at or above 0.70 for group-level comparisons, but may be lower if appropriately justified. Reliability can be estimated using a variety of methods including internal consistency reliability, test–retest reliability, or item response theory. Each method should be justified</td>
</tr>
<tr>
<td>Internal consistency reliability</td>
<td>The degree of the interrelatedness among the items in a multi-item PRO measure</td>
<td></td>
</tr>
<tr>
<td>Test–retest reliability</td>
<td>A measure of the reproducibility of the scale, that is, the ability to provide consistent scores over time in a stable population [2]</td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>The degree to which a PRO instrument measures the PRO concept it purports to measure</td>
<td>A PRO measure should have evidence supporting its content validity, including evidence that patients and experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application.</td>
</tr>
<tr>
<td>Content validity</td>
<td>The extent to which the PRO measure includes the most relevant and important aspects of a concept in the context of a given measurement application</td>
<td>A PRO measure should have evidence supporting its construct validity, including documentation of empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO</td>
</tr>
<tr>
<td>Construct validity</td>
<td>The degree to which scores on the PRO measure relate to other measures (e.g., patient-reported or clinical indicators) in a manner that is consistent with theoretically derived a priori hypotheses concerning the concepts that are being measured</td>
<td></td>
</tr>
<tr>
<td>Criterion validity</td>
<td>The degree to which the scores of a PRO measure are an adequate reflection of a “gold standard.”</td>
<td>A PRO measure for use in longitudinal research study should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the measured PRO in the target population for the research application</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The extent to which a PRO measure can detect changes in the construct being measured over time</td>
<td>A PRO measure should have documentation to support interpretation of scores, including what low and high scores represent for the measured concept</td>
</tr>
<tr>
<td>Interpretability of scores</td>
<td>The degree to which one can assign easily understood meaning to a PRO measure’s scores</td>
<td></td>
</tr>
<tr>
<td>Minimal important difference (MID)</td>
<td>The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the management</td>
<td>A PRO measure must not be overly burdensome for patients or investigators. The length of the PRO measure should be considered in the context of other PRO measures included in the assessment, the frequency of PRO data collection, and the characteristics of the study population.</td>
</tr>
<tr>
<td>Burden</td>
<td>The time, effort, and other demands placed on those to whom the instrument is administered (respondent burden) or on those who administer the instrument (investigator or administrative burden)</td>
<td>A PRO measure translated to one or more languages should have documentation of the methods used to translate and evaluate the PRO measure in each language. Studies should at least include evidence from qualitative methods (e.g., cognitive testing) to evaluate the translations</td>
</tr>
<tr>
<td>Translation of the PRO measure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
There are two broad categories of QoL measures, disease specific and generic measures. Disease specific QoL measures typically emphasize evaluation of symptoms, functioning and patient perceptions that relate to a narrowly defined disease or condition. Generic measures assess broad categories of functioning and well-being that can be affected by a multitude of conditions, and enable comparisons between the relative burdens of diseases and their treatment benefits. Both types of measures have been widely used to quantify QoL. However, because generic QoL measures favour broad functional health and well-being concepts over symptoms and problems that are more specific, they are sometimes found to be less sensitive to differences among clinically relevant groups and to be less responsive to small changes in health (9, 10). Hundreds of disease specific QoL measures exist, while a more limited number of generic measures have been widely adopted (11).

Parallel to the increased focus on patient-centered outcomes, the use of QoL measures has grown exponentially in clinical trials of therapies for chronic diseases (12, 13). Although the widespread use of QoL measures within routine clinical care and health systems as a whole, these tools are increasingly being applied in clinical settings to both monitor patients’ conditions and to help in clinical decision-making (14). For example, QoL measures have been used internationally to monitor patients with low-back pain and spinal conditions (15), to track patients with chronic diseases in primary care (16), and as part of the assessment for the need for a total joint replacement (17). In the UK QoL measures are already being used to provide public comparisons of provider performance, while in the U.S. current plans are to incorporate these measures into evaluations of health insurer and health care provider performance (14, 16). Such routine consideration of QoL provides the opportunity to monitor the impact of treatments on the outcomes most meaningful for patients and an opportunity to shape the delivery of value based care (18).

While the majority of QoL surveys were developed as simple paper-pencil questionnaires, using classical psychometric methods, the field has advanced greatly in standardizing metrics for quantifying the major domains of functional health and well-being, and in applying modern psychometric techniques to ease the burden of using these tools in practice (19, 20). Modern psychometric techniques have allowed for the application of adaptive testing to reduce respondent burden and increase precision over a wider range of QoL scores. Web and mobile technologies have further eased the ability to deliver surveys to patients in the clinic as well as outside of the context of the traditional clinical visit to monitor patient progress with greater continuity (21). Some QoL tools are now as reliable
as clinical measures like diastolic blood pressure or glucose (14, 22). Together, these advances in measurement techniques, combined with the growing applications for QoL in clinical settings, make it important for clinicians to become familiar with QoL concepts and the tools designed for their specific area of clinical practice.

Thyroid diseases affect QoL (23-28), work role function (29, 30), as well as morbidity and mortality (31, 32). Nonetheless, research focusing on QoL among thyroid patients has been scarce and until recently, no validated thyroid-specific patient-reported outcome measuring thyroid-related QoL has been available. This has probably in part been due to lack of pharmaceutical interest, which has fuelled PRO research in many other fields of medicine (33).

The ThyQoL project was launched to address this shortage (34). The ThyPRO instrument was developed as a comprehensive thyroid-related stand-alone PRO for patients with any benign thyroid disease (35-37). It was crucial that it covered all benign thyroid diseases, in order for it to maintain content validity, when patients 'converted' from one diagnosis to the other, as a consequence of treatment (e.g. patients with non-toxic goiter becoming hypothyroid and requiring thyroid hormonal substitution after thyroidectomy).

The development and initial psychometric validation of the ThyPRO in Danish, was described in my PhD-thesis (24, 34-36). It was conducted within a classical QoL measurement theoretical framework (6, 8, 38), adapted to thyroid diseases, as outlined in Figure 1.
The ThyPRO covers both physical symptoms specifically relevant to thyroid diseases, e.g. symptoms of hyperthyroidism and goiter, but also non-specific aspects of high importance to patients with thyroid diseases, e.g. fatigue. It consists of 85 items summarized in 13 multi-item scales as well as a single item measuring impact of the thyroid disease on overall QoL. Each item is rated on a 0-4 Likert scale from no symptoms / problems =0 to severe symptoms/problems =4. The average score of items in a scale is divided by four and multiplied by 100 to yield thirteen 0-100 scales, with higher scores indicating worse health status.
Figure 2: Danish/English parallel development process

1. Literature review
2. Interviews with 15 expert clinicians
   -> List of relevant issues
3. Interviews with 80 patients
   -> First Danish version of the questionnaire
   -> Iterative rounds of cognitive interviews
   -> "Second" reconciled Danish version of the questionnaire
   -> Scale validation
     - Clinical validation
     - Reliability testing (i.e. internal consistency and test-retest)
     - Resulting item reduction
4. First English version of the questionnaire
   -> Standard (forward-backward) translation procedure
5. Iterative rounds of cognitive interviews
6. Harmonization
   -> Second" reconciled English version of the questionnaire
   -> Item reduction reflected in English version
7. Third Danish version: the Danish ThyPRO
8. Third English version: the English ThyPRO
An important way of assessing validity, especially for measures attempted for clinical use, is known-groups validity. In this approach, clinically based criteria are used to classify patients into groups with expected high or low scores on a questionnaire and then test whether these expected differences are found in patient samples. Another is test-retest reproducibility, where duplicate measurements are obtained by collecting two responses separated by two to three weeks from stable respondents (39). A further very important aspect of the validity of a PRO is whether the instrument can detect relevant clinical changes over time, i.e. its responsiveness (40, 41).

The initial psychometric validation employed classical psychometric methodology. However, within such a framework, it is not possible to test the overall fit of a model (42), nor can misfit of items be modeled specifically. Further, we were interested in developing a short version of the instrument as well as versions applicable to ecological momentary assessments. Development of such versions can be informed by the application of item response theory (IRT) models. However, IRT models require additional, more detailed examinations of the dimensionality of the ThyPRO scales. Structural equation models provide a latent variable modeling framework that is useful in detailed examinations of dimensionality. The measurement part of structural equation models can be used to assess the dimensionality of measured variables such as questionnaire items, using confirmatory factor analysis (CFA) for categorical data. Structural equation modeling can also test relationships among modeled latent variables (i.e., structural part of the models) (43-47).

A PRO instrument and its scales may be subject to differential item functioning (DIF) (48-52). An item in a multi-item scale is subject to DIF if different groups of respondents with the same level of the measured attribute, score differently on that item (53). A PRO measurement instrument may display DIF also according to culture. In order to increase translatability, the ThyPRO was developed in a parallel multi-language process (Figure 2) and is available in several languages. Each translation was conducted using stringent criteria, including cognitive interviewing of patients. However, some items may have low translatability, and they can be identified using DIF analyses across samples from different countries. (33, 54, 55)

Currently, the ThyPRO is or has been in use in more than 40 clinical studies world-wide. This original version is rather long (85 items) and is reported in numerous (i.e. 14) scales. A shorter version for use as e.g., secondary outcome in clinical trials and in daily clinical practice would further advance its applicability (56).
The purpose of the work included in this thesis was to develop, validate and apply an international patient-reported outcomes measurement system, measuring thyroid-related quality of life. The work is subdivided into five phases: 1) clinical validation, 2) psychometric validation, 3) cross-cultural validation, 4) abbreviation and 5) application.

METHODS

STUDY POPULATIONS

CROSS-SECTIONAL SAMPLE
Nine hundred and seven patients were recruited cross-sectionally during 2007-2008 from the endocrinological outpatient clinics at Copenhagen University Hospital Rigshospitalet and Odense University Hospital. Questionnaires were sent to the patients about three weeks prior to an appointment in the clinic.

TEST-RETEST SUBSAMPLE
A subset (n=87) of the patients enrolled in the cross-sectional study was asked to complete another questionnaire about two weeks after their first response. In addition, they rated any change occurring since the initial response.

LONGITUDINAL SAMPLE
From 2008 to 2013, 435 patients undergoing treatment for benign thyroid disease at the two abovementioned sites, completed ThyPRO and SF-36 Health Surveys prior to and at six weeks and six months after treatment.

INTERNATIONAL SAMPLES
For the cross-cultural validation study, data came from 7 samples in 7 languages: English (Scotland and Ireland, n=166), Dutch (n=147), Serbian (n=150), Italian (n=110), Hindi (India, n=148), Denmark (the ‘Cross-sectional sample’) and Swedish (n=187).
PHASE 1: CLINICAL VALIDATION

For the known-groups comparisons, groups of patients within the ‘Cross-sectional sample’ with expected high and low scores were defined a priori by a panel of four thyroid experts. Test-retest reliability was evaluated in the ‘Test-retest subsample’ by intra-class correlations between the two measurements (57-60).

For evaluation of responsiveness, three patient groups from the ‘Longitudinal sample’, undergoing clinically relevant change were defined: patients with 1) hyperthyroidism rendered euthyroid by treatment six months after treatment initiation; 2) patients with autoimmune hypothyroidism rendered euthyroid six months after treatment initiation, and 3) patients with a clinically noticeable goiter treated with ablative therapy.

PHASE 2: PSYCHOMETRIC VALIDATION

DIMENSIONALITY

A one-factor confirmatory model for ordinal data was fitted to each individual ThyPRO scale (61, 62), using Mplus (version 7.11) (63), in the ‘Cross-sectional sample’. Appropriateness of this initial one-factor model for each scale was assessed by: 1) overall goodness-of-fit statistics (52, 64-67); 2) magnitude of factor loadings; 3) model residual correlations and 4) modification indices (63, 68). Revisions to improve model fit were based on both confirmatory factor modeling and item/scale content. For scales where secondary factors seem plausible, a bi-factor model was fitted to evaluate the dominance of the primary factor when secondary factors were modeled (52, 69-72). In an attempt to understand any possible item misfit identified through individual scale analyses, hypotheses which could explain the misfit were evaluated in a combined, investigational multidimensional model.

DIFFERENTIAL ITEM FUNCTIONING

Uniform and non-uniform DIF was investigated using ordinal logistic regression (73) (p<0.05 with Bonferroni correction) (74, 75). In addition, any DIF should induce a change in $R^2 > 0.02$ (73, 76), i.e. it should explain more than 2% of the variation in item scores. Scale level was estimated by the sum score, including also the item being tested (73, 77, 78). Other items with DIF were excluded, i.e. the sum score was purified (73, 79).
Impact of DIF on scale score was evaluated by plotting the expected mean item responses for each DIF grouping, as estimated by the regression model, against the level of the respective purified scales.

**PHASE 3: CROSS-CULTURAL VALIDATION**

In the analyses of DIF according to language, each of the six non-English versions was tested against the English version. DIF was investigated using ordinal logistic regression (73, 77-79). All instances of DIF were reviewed by language and clinical experts in order to identify possible linguistic, cultural or clinical explanations for the DIF.

**PHASE 4: ABBREVIATION**

The abbreviation was conducted in three separate steps: 1) selection of items for the short form, including selection of scales for a composite score, 2) scoring of short scales and 3) validation of the short form.

**ITEM SELECTION**

The Hypothyroid Symptoms scale was retained in full length and the Impaired Sex Life scale was excluded from the abbreviated version. For each of the remaining 11 scales, items previously shown to fit a unidimensional factor model, (80) were analyzed using Samejima’s graded IRT model (81, 82) in the ‘Cross-sectional sample’, until a model without misfit was identified (83-86). The best items were selected, aiming at three-item scales.

**SCORING COMPOSITE AND SHORT-FORM SCALES**

The ThyPRO mental and social well-being and function scales are highly correlated (36, 80) and it was decided to summarize them in a supplementary composite score. The individual short-form scales were scored, using the Orlando and Thissen IRT-based summed-score linking (87).

**SCALE VALIDATION**

Agreement was estimated by plots, mean score levels and by intra-class correlation (58, 88) in the ‘Cross-sectional sample’. Responsiveness was evaluated by effect sizes (89, 90) and relative validity (91) in the ‘Longitudinal sample’. In addition, evaluation of known-groups validity and test-retest reliability (see above) was repeated for the short-form scales.
Analyses of the relationship between clinical variables and ThyPRO scales, among patients with autoimmune hypothyroidism in the ‘Cross-sectional sample’, were conducted within the theoretical ThyQoL model. Clinical variables were thyroid volume, thyroid hormones and thyroid antibodies. Initially, pairwise relationships between all variables in the theoretical model were analyzed univariately, controlling for age, gender and educational level. Subsequently, statistically significant associations (p<0.05) were joined in one overall path model, using Mplus.(63)

In order to investigate beneficial effects of selenium supplementation to patients with Graves’ hyperthyroidism, ThyPRO was implemented in the GRASS (GRAves’ disease Selenium Supplementation trial) randomized clinical multi-center trial. In GRASS, the experimental intervention group will receive 200 μg selenium once daily for the 24-30 months intervention period. The control intervention group will receive an identical placebo tablet once daily for the 24-30 months intervention period. The trial will include 492 participants (2 x 296 participants). Inclusion criteria are: active Graves’ hyperthyroidism within the last two months and written informed consent. Exclusion criteria are: major co-morbidity; previous radioactive iodine treatment; current anti-thyroid drug treatment for more than 8 weeks; treatment with immunomodulatory drugs; allergy towards the pill-components, pregnancy or breast-feeding; intake of selenium supplementation above 70 μg per day; inability to read and understand Danish. The primary outcome is the proportion of participants with ‘anti-thyroid treatment failure’ at end of the intervention period (24-30 months). Secondary outcomes are: thyroid-specific quality of life during the first year after randomisation; level of TSH-receptor antibodies at 18 months after randomisation, and at end of the intervention period (24-30 months); hyperthyroid symptoms during first year after randomisation; eye symptoms during first year after randomisation, and at end of intervention period (24-30 months); adverse reactions during the intervention period; and serious adverse events during the intervention period.

As part of the trial, a full-fledged trial data-management IT-system for pragmatic trials with focus on high-quality PRO data was developed (Figure 3) (21).
Figure 3 The trial management system PROgmatic

PROgmatic: A clinical trial management system, optimized for PROs

- Patients
  - Electronic surveys [SurveyXact]
  - EMA (mobile phone app)
  - SMS gateway

- Trial personnel
  - Emails linking to electronic surveys

- Researchers
  - Portal: trial data and trial progress

- Electronic CRFs
  - Emails
RESULTS

PHASE 1: CLINICAL VALIDATION

As shown in Figure 4, the expected high level groups had substantially higher mean scores compared with the expected low level groups on all ThyPRO scales. In the test-retest analyses, all intra-class correlations were above 0.70 and for all but two scales it was above the “almost perfect” concordance (i.e. > 0.81) (92).

Figure 4 Mean scale scores for the expected high score groups vs. the low score groups for the 13 ThyPRO scales. Vertical lines indicate 95% confidence intervals

Of the 435 respondents in the longitudinal study, a total of 212 could be classified into one of the three pre-defined clinically respondent groups six months after treatment initiation. As shown in Table 2, statistically significant changes were observed in all ThyPRO and SF-36 scales in the hyperthyroid response sample. In the hypothyroid response sample, all ThyPRO scales except the Hypothyroid Symptoms scale measuring physical symptoms, were significantly better at follow-up, as were five of the eight SF-36 scales. In the goiter response sample, statistically significant improvements were observed in four of the ThyPRO scales, i.e. Goiter Symptoms, Anxiety, Tiredness and Emotional Susceptibility. In this group, none of the SF-36 scales improved significantly.
Table 2 Mean (SD) ThyPRO and SF-36 scale scores at baseline and at follow-up six months after treatment. Statistically significant differences (p<0.05 after Hochberg) in bold. ThyPRO scales which the four thyroid experts predicted to change are marked with *. One • for each expert expecting the scale to change. Differences are in absolute values.

<table>
<thead>
<tr>
<th>ThyPRO scale</th>
<th>Hyperthyroidism treated to become euthyroid at follow-up (n=66)</th>
<th>Autoimmune hypothyroidism treated to become euthyroid (n=84)</th>
<th>Non-toxic goiter treated with surgery or radio-iodine for volume reduction (n=62)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline At six months Change</td>
<td>Baseline At six months Change</td>
<td>Baseline At six months Change</td>
</tr>
<tr>
<td>Goiter Symptoms</td>
<td>19 (19) 7 (10) 12 (17)</td>
<td>12 (15) 8 (10) 4 (14)</td>
<td>36 (21) 14 (14) 21 (24)</td>
</tr>
<tr>
<td>Hyperthyroid Symptoms</td>
<td>35 (21) 14 (13) 20 (22)</td>
<td>19 (15) 15 (14) 4 (13)</td>
<td>22 (18) 16 (16) 6 (18)</td>
</tr>
<tr>
<td>Hypothyroid Symptoms</td>
<td>24 (21) 16 (17) 9 (18)</td>
<td>25 (22) 23 (20) 2 (16)</td>
<td>20 (19) 18 (19) 2 (19)</td>
</tr>
<tr>
<td>Eye Symptoms</td>
<td>19 (22) 13 (15) 6 (17)</td>
<td>13 (14) 9 (10) 4 (11)</td>
<td>15 (16) 10 (11) 5 (13)</td>
</tr>
<tr>
<td>Tiredness</td>
<td>63 (27) 39 (25) 24 (29)</td>
<td>61 (28) 44 (27) 17 (26)</td>
<td>51 (24) 41 (25) 10 (23)</td>
</tr>
<tr>
<td>Cognition</td>
<td>21 (22) 13 (17) 8 (16)</td>
<td>27 (25) 21 (22) 6 (19)</td>
<td>21 (23) 18 (23) 3 (19)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>31 (23) 13 (19) 18 (23)</td>
<td>20 (18) 7 (10) 1 (15)</td>
<td>32 (23) 17 (18) 16 (23)</td>
</tr>
<tr>
<td>Depressivity</td>
<td>33 (21) 23 (18) 10 (21)</td>
<td>33 (22) 27 (19) 6 (21)</td>
<td>29 (19) 27 (21) 1 (17)</td>
</tr>
<tr>
<td>Emotional Susceptibility</td>
<td>41 (26) 24 (23) 16 (23)</td>
<td>40 (24) 29 (21) 11 (19)</td>
<td>33 (22) 27 (22) 6 (15)</td>
</tr>
<tr>
<td>Impaired Social Life</td>
<td>16 (23) 8 (13) 7 (17)</td>
<td>15 (20) 9 (16) 5 (18)</td>
<td>9 (15) 7 (13) 2 (11)</td>
</tr>
<tr>
<td>Impaired Daily Life</td>
<td>30 (28) 12 (20) 18 (28)</td>
<td>25 (28) 13 (22) 12 (22)</td>
<td>14 (19) 9 (15) 5 (17)</td>
</tr>
<tr>
<td>Impaired Sexlife</td>
<td>31 (35) 17 (28) 14 (30)</td>
<td>25 (30) 17 (22) 8 (29)</td>
<td>20 (26) 15 (26) 5 (31)</td>
</tr>
<tr>
<td>Cosmetic Complaints</td>
<td>18 (22) 14 (19) 4 (16)</td>
<td>18 (18) 13 (16) 5 (15)</td>
<td>17 (16) 17 (19) 0 (19)</td>
</tr>
<tr>
<td>Overall quality of life</td>
<td>43 (35) 17 (26) 26 (34)</td>
<td>40 (36) 19 (26) 21 (33)</td>
<td>32 (29) 23 (28) 8 (32)</td>
</tr>
<tr>
<td>SF-36 scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Function</td>
<td>67 (27) 82 (25) 15 (25)</td>
<td>81 (22) 84 (23) 2 (14)</td>
<td>84 (20) 84 (22) 0 (15)</td>
</tr>
<tr>
<td>Role Physical</td>
<td>59 (32) 77 (29) 16 (31)</td>
<td>70 (29) 79 (27) 10 (22)</td>
<td>78 (26) 81 (24) 3 (27)</td>
</tr>
<tr>
<td>Bodily Pain</td>
<td>72 (29) 83 (23) 11 (25)</td>
<td>74 (30) 82 (24) 7 (21)</td>
<td>74 (26) 82 (23) 8 (22)</td>
</tr>
<tr>
<td>General Health</td>
<td>62 (19) 70 (19) 7 (18)</td>
<td>68 (22) 68 (22) 0 (13)</td>
<td>68 (19) 70 (20) 1 (16)</td>
</tr>
<tr>
<td>Vitality</td>
<td>40 (27) 58 (25) 18 (25)</td>
<td>42 (26) 54 (25) 13 (22)</td>
<td>55 (23) 60 (26) 5 (19)</td>
</tr>
<tr>
<td>Social Function</td>
<td>75 (27) 87 (20) 12 (26)</td>
<td>76 (29) 83 (26) 7 (19)</td>
<td>83 (23) 89 (17) 6 (20)</td>
</tr>
<tr>
<td>Role Emotional</td>
<td>71 (29) 80 (29) 8 (30)</td>
<td>75 (25) 81 (25) 5 (23)</td>
<td>80 (27) 85 (20) 5 (28)</td>
</tr>
<tr>
<td>Mental Health</td>
<td>62 (20) 74 (18) 13 (21)</td>
<td>68 (16) 74 (17) 6 (15)</td>
<td>68 (19) 75 (19) 6 (17)</td>
</tr>
</tbody>
</table>
In the hyperthyroid response sample, a statistically significant change was observed on all scales expected by clinicians to change (Table 2). In the hypothyroid sample, a change was observed on all except one (Hypothyroid Symptoms) of the expected scales. In the goiter response sample, significant changes were observed in two of the scales expected to improve, whereas no change was found on the other two expected to change, i.e. Impaired Social Life and Cosmetic Complaints.

Relative responsiveness of similar scales could be compared in five instances: ThyPRO Fatigue vs. SF-36 Vitality, ThyPRO Anxiety, Depression and Emotional Susceptibility vs. SF-36 Mental Health, ThyPRO Impaired Social Life vs. SF-36 Social Function, ThyPRO Impaired Daily Life vs. SF-36 Role Physical and Role Emotional and ThyPRO overall QoL impact vs. SF-36 General Health. For all comparisons, except for social function, the ThyPRO had better responsiveness than SF-36.

**PHASE 2: PSYCHOMETRIC VALIDATION**

**DIMENSIONALITY**

In the confirmatory factor analyses, factor loadings were high in all scales and confirmatory fit index was also high for the vast majority of scales. In contrast, for most scales, root mean square error of approximation was not below the 0.08 threshold for appropriate fit. The consequential remodeling resulted in the revised scales presented in Figure 5.

The investigative modeling of possible item misfit within one combined multidimensional model is presented in Table 3. As shown, the hypothesized explanations for the apparent misfit were confirmed in most cases.
Figure 5 Parameter estimates of the unidimensional confirmatory factor analyses of the revised ThyPRO scales. Overall goodness-of-fit of the models are provided in the text. Grayed out items were omitted during model revision. The two-item Impaired Sexlife scale was not estimated.
<table>
<thead>
<tr>
<th>Item</th>
<th>Hypothesized reason for misfit</th>
<th>Investigative modeling of the hypothesized reason for misfit</th>
<th>Results of the investigative modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visible swelling on neck from the Goiter Symptoms scale</td>
<td>May relate to cosmetic concerns, rather than being a symptom</td>
<td>Item was allowed to cross-load on the Cosmetic Complaints factor</td>
<td>Loaded -0.23 on the Cosmetic Complaints factor. Loading on own factor: 0.68</td>
</tr>
<tr>
<td>Throat pain felt in ears from the Goiter Symptoms scale</td>
<td>May be relevant only for patients with subacute thyroiditis, during the acute inflammatory phase.</td>
<td>No marker of acute inflammation is available in the clinical database describing the patients. Only 9 patients in this sample had subacute thyroiditis</td>
<td>Extraneous modeling not possible. Loading on own factor in the full model: 0.75</td>
</tr>
<tr>
<td>Hoarseness from the Goiter Symptoms scale</td>
<td>Hoarseness is also a classical symptom of hypothyroidism. Might relate more to hypothyroidism than to goiter.</td>
<td>Item was allowed to cross-load on the Hypothyroid Symptoms factor</td>
<td>Loaded 0.22 on Hypothyroid Symptoms factor. Loading on own factor: 0.46</td>
</tr>
<tr>
<td>Loose stools from the Hyperthyroid Symptoms scale</td>
<td>Might be a non-specific physical symptom</td>
<td>Item was allowed to load on the other physical symptoms factors, except for Eye Symptoms</td>
<td>Loaded -0.15 on Goiter Symptoms factor and -0.55 on Hypothyroid Symptoms. Loading on own factor: 1.20</td>
</tr>
<tr>
<td>Afraid of being seriously ill from the Anxiety scale</td>
<td>May be related to not being fully examined yet, and thus an initial fear of e.g. cancer has not yet been ruled out completely</td>
<td>Item was regressed on time since diagnosis.</td>
<td>A significant negative association with time since diagnosis was found</td>
</tr>
<tr>
<td>Other people lack understanding from the Impaired Social Life scale</td>
<td>May relate more to depressive mood and emotional distress than the other items in the Social Life scale</td>
<td>Item was allowed to cross-load on the Depressivity and the Emotional Susceptibility factor</td>
<td>No significant loading on Depressivity or Emotional Susceptibility was found. Loading on own factor: 1.08</td>
</tr>
<tr>
<td>Felt too fat from the Cosmetic Complaints scale</td>
<td>Weight gain is often experienced during hypothyroidism. Feeling too fat may also relate more to a negative self-esteem aspect of depressive mood</td>
<td>Item was allowed to cross-load on the Hypothyroid Symptoms and Depressivity and Anxiety factors</td>
<td>Loaded -0.16 on Hypothyroid Symptoms factor, -0.22 on Anxiety and 0.15 on Depressivity factor. Loading on own factor: 0.53</td>
</tr>
</tbody>
</table>
DIFFERENTIAL ITEM FUNCTIONING

In total, among the 84 items being tested, 20 instances of DIF were found, involving 17 items within 8 of the 13 ThyPRO scales. Eight instances of DIF according to diagnosis were found, all of which involved symptom scales and all but one of these were physical symptoms scales. The Goiter Symptoms scale was most affected with 4 instances of DIF, mainly due to patients with goiter scoring higher on the involved items for a given level of goiter symptoms, compared to patients with the autoimmune thyroid diseases.

DIF according to diagnosis was also found for the symptom scales Hyperthyroid Symptoms, Eye Symptoms, and Anxiety. For the Hyperthyroid Symptoms scale, patients with Graves' hyperthyroidism had a higher probability of reporting 'Trembling hands' than other disease groups. For the Eye Symptoms scale, patients with Graves' orbitopathy were less likely to report 'Grittiness in eyes', for a given level of Eye Symptoms and patients with goiter were less likely to report 'Reduced sight'. For the Anxiety scale, patients with a non-toxic goiter, reported more 'Concern about being seriously ill' (most often implying fear of cancer), for a given level of Anxiety.

Eight instances of DIF according to age were found (Table 4). Five of the eight DIFs were in positively worded items, and for all of these, younger patients were more likely to endorse positively on the positively worded items than older patients at the same level of the involved scales.

One instance of DIF according to gender (women more likely to respond positively to the item about crying easily, provided the same level of depressivity, compared to men) and three instances according to educational level was found (Table 5). In terms of the influence of the identified DIF on the scale score, the effects were small.
## Table 4 Differential item functioning according to age

<table>
<thead>
<tr>
<th>Abbreviated items wording</th>
<th>Significance of DIF</th>
<th>Variance explained by DIF</th>
<th>Direction of DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tiredness scale (Threshold for significance level with Bonferroni: 0.0071)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full of life</td>
<td>&lt;0.0001</td>
<td>2.7%</td>
<td>Younger patients more full of life</td>
</tr>
<tr>
<td>Energetic*</td>
<td>&lt;0.0001</td>
<td>3.5%</td>
<td>Younger patients more energetic</td>
</tr>
<tr>
<td>Able to cope with life</td>
<td>&lt;0.0001</td>
<td>3.8%</td>
<td>Younger patients more able to cope with life</td>
</tr>
<tr>
<td><strong>Depressivity scale (Threshold for significance level with Bonferroni: 0.0071)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crying easily</td>
<td>&lt;0.0001</td>
<td>2.6%</td>
<td>Younger patients cry more easily</td>
</tr>
<tr>
<td>Happy</td>
<td>&lt;0.0001</td>
<td>2.1%</td>
<td>Younger patients more happy</td>
</tr>
<tr>
<td>Self-confident*</td>
<td>&lt;0.0001</td>
<td>2.1%</td>
<td>Younger patients more self-confident</td>
</tr>
<tr>
<td><strong>Impaired Social Life scale (Threshold for significance level with Bonferroni: 0.0125)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult being with other people</td>
<td>0.0033</td>
<td>3.4%</td>
<td>Younger patients less difficulty being with other people</td>
</tr>
<tr>
<td>A burden to other people</td>
<td>0.0020</td>
<td>2.3%</td>
<td>Younger patients less a burden to other people</td>
</tr>
</tbody>
</table>
### Table 5 DIF according to gender and educational level.

Educational levels were: No education, currently studying, apprenticeship education, short theoretical education (1-3 years), medium theoretical education (3-5 years) and long theoretical education (5+ years)

<table>
<thead>
<tr>
<th>Scale and significance level with Bonferroni</th>
<th>Abbreviated items wording</th>
<th>Significance of DIF</th>
<th>Variance explained by DIF</th>
<th>Direction of DIF</th>
<th>DIF reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressivity (0.0071)</td>
<td>Crying easily</td>
<td>&lt;0.0001</td>
<td>2.2%</td>
<td>Women cried more easily</td>
<td>DIF17</td>
</tr>
<tr>
<td>Depressivity (0.0071)</td>
<td>Self-confident</td>
<td>&lt;0.0001</td>
<td>2.4%</td>
<td>Patients with shorter education were less self-confident</td>
<td>DIF18</td>
</tr>
<tr>
<td>Impaired Social Life (0.0125)</td>
<td>Difficult being with other people*</td>
<td>0.0100</td>
<td>3.5%</td>
<td>Patients with No education and Currently studying had less difficulty (especially with higher scale score). Patients with an Apprenticeship and Medium theoretical education had more difficulty</td>
<td>DIF19</td>
</tr>
<tr>
<td>Cosmetic Complaints (0.0083)</td>
<td>Disease affect appearance*</td>
<td>0.0060</td>
<td>2.0%</td>
<td>Patient currently studying and patients with an apprenticeship education had more affected appearance (especially with higher scale score). Patients with a medium long theoretical education had less affected appearance</td>
<td>DIF20</td>
</tr>
</tbody>
</table>
PHASE 3: CROSS-CULTURAL VALIDATION

There were seven items with DIF in Italian and Hindi respectively, six in Serbian and three in Swedish (Table 6). Most instances of DIF were small with ΔR² below 5%, but seven were larger: “Sense of suffocating” in Serbian, “Felt happy” and “Conflicts with other people” in Italian, “Full of life”, “Concerned being seriously ill” and “Difficulty managing job” in Hindi and “Upset stomach” in Swedish. When reviewed by in-country clinical experts, these DIFs could be explained in three of these seven instances: In Hindi, “Concerned being seriously ill” could be explained by the fact that more patients in this sample had goiter, and that these were relatively large goiters. “Difficulty managing job” in Hindi may reflect that the jobs held by most Indians are more dependent on a good health, compared to the other countries. The Serbian sample also comprised many patients with large goiters (like the Indian sample, these patients were referred to thyroid surgery), which may explain the DIF in “Sense of suffocating” from the Goiter Symptoms scale.

In twelve instances, DIF was identified in more than one country. The magnitude of most of these instances of DIF were small, but for “Visible swelling in front of neck”, “Felt too fat”, “Been confused” and “Feeling not like yourself “ the results were implying larger differences across the countries for patients with the same level of the attribute measured for those items.
Table 6 DIF in single languages. For each instance of DIF which was only present in one of the six tested languages, the English reference wording, the tested language wording, direction of DIF, scale to which the item belongs and the magnitude of DIF, in terms of the variance explained (ΔR² in percent), is presented. In case of non-linearity, the direction of DIF is ambiguous and cannot be presented briefly.

<table>
<thead>
<tr>
<th>Item #</th>
<th>Serbian wording</th>
<th>Direction of DIF</th>
<th>English wording</th>
<th>Scale</th>
<th>Italian wording</th>
<th>Direction of DIF</th>
<th>Hindi (India) wording</th>
<th>Scale</th>
<th>Swedish wording</th>
<th>Direction of DIF</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1i</td>
<td>- had difficulty swallowing?</td>
<td>Goiter</td>
<td>- had difficulty swallowing?</td>
<td>Goiter</td>
<td>- had difficulty swallowing?</td>
<td>Goiter</td>
<td>- had difficulty swallowing?</td>
<td>Goiter</td>
<td>- had difficulty swallowing?</td>
<td>Goiter</td>
<td>- had difficulty swallowing?</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 2.8</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.8</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.8</td>
<td>Item English wording</td>
</tr>
<tr>
<td>1j</td>
<td>- had the sensation of suffocating?</td>
<td>Goiter</td>
<td>- had the sensation of suffocating?</td>
<td>Goiter</td>
<td>- had the sensation of suffocating?</td>
<td>Goiter</td>
<td>- had the sensation of suffocating?</td>
<td>Goiter</td>
<td>- had the sensation of suffocating?</td>
<td>Goiter</td>
<td>- had the sensation of suffocating?</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 7.6</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 7.6</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 7.6</td>
<td>Item English wording</td>
</tr>
<tr>
<td>1cc</td>
<td>- had swollen hands or feet?</td>
<td>Hypo</td>
<td>- had swollen hands or feet?</td>
<td>Hypo</td>
<td>- had swollen hands or feet?</td>
<td>Hypo</td>
<td>- had swollen hands or feet?</td>
<td>Hypo</td>
<td>- had swollen hands or feet?</td>
<td>Hypo</td>
<td>- had swollen hands or feet?</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 4.6</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 4.6</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 4.6</td>
<td>Item English wording</td>
</tr>
<tr>
<td>1v</td>
<td>- had bags under the eyes or swollen eyelids?</td>
<td>Hypo</td>
<td>- had bags under the eyes or swollen eyelids?</td>
<td>Hypo</td>
<td>- had bags under the eyes or swollen eyelids?</td>
<td>Hypo</td>
<td>- had bags under the eyes or swollen eyelids?</td>
<td>Hypo</td>
<td>- had bags under the eyes or swollen eyelids?</td>
<td>Hypo</td>
<td>- had bags under the eyes or swollen eyelids?</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 3.3</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 3.3</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 3.3</td>
<td>Item English wording</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 2.0</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.0</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.0</td>
<td>Item English wording</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 2.9</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.9</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 2.9</td>
<td>Item English wording</td>
</tr>
<tr>
<td>11a</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic</td>
<td>- has your thyroid disease affected your appearance?</td>
</tr>
<tr>
<td></td>
<td>Serbian wording</td>
<td>ΔR² 4.4</td>
<td>English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 4.4</td>
<td>Item English wording</td>
<td>Scale</td>
<td>Item Serbian wording</td>
<td>ΔR² 4.4</td>
<td>Item English wording</td>
</tr>
</tbody>
</table>
The Impaired Social Life scale only included three unidimensional items, and was thus not abbreviated further. In the original long-form scales: Goiter Symptoms, Hyperthyroid Symptoms, Cognitive Complaints, Anxiety, Social Life, and the Cosmetic Complaints scale, all items had good fit to the IRT model. For the Eye Symptoms, Tiredness, Depressivity, Emotional Susceptibility and the Daily Life scale, item-level misfit was eliminated through reduction of the scales. Abbreviated scales with 3 items each were obtained for 10 of the 11 scales, whereas the Hyperthyroid Symptoms scale had 4 items.

Agreement plots of the new, rescaled short-form scales vs. the original long-form scales are presented in Figure 7. Good and uniform agreement was shown across the entire range of scores. Only the Tiredness short-form mean score was outside the confidence interval for the long-form scale. However, the difference between the two mean levels was only 3 points on the 0-100 scale.

With specification of a method-factor comprising the positively worded items, a bi-factor model with acceptable overall fit (comparative fit index =0.97, root mean square error of approximation =0.085) was fitted to the composite score (Figure 6).

Effect sizes and responsiveness in groups of patients undergoing treatment were preserved in the short-form scales. Only the short-form Anxiety scale had smaller effect size and slightly lower responsiveness than the original long-form; conversely, the short-form Cosmetic Complaints scale had slightly higher effect size and responsiveness, compared with the long-form. Test-retest reliability was similarly preserved in the short-form scales, with only the Cosmetic Complaints scale having significantly, but marginally lower reliability. Very high short- vs. long-form intra-class correlations were found, 0.89 to 0.98. Further, the previously shown discriminant validity was reproduced.
Figure 6 Bi-factor modeling evaluating the composite scale, summarizing the seven short-form mental and social well-being and function scales (left hand side). Factor loadings are presented at the relevant arrows. All items except one had higher loading on the general factor representing the Composite score.
Figure 7 Agreement plots of short-form (horizontal axis) vs. original long-form (vertical axis) scale levels with regression lines. For each scale, mean score (95%CI) for the long-form is presented in the upper left corner and mean score for the short-form in the lower right. Short-form mean outside long-form 95%CI is marked by *.

Long-form mean=17 (15-18)

Short-form mean=16

Goiter Symptoms Hyperthyroid Symptoms Eye Symptoms

49 (47-51) 22 (20-23) 23 (21-24)

*46 22 23

Tiredness Cognitive impairment Anxiety

29 (28-31) 33 (32-35) 11 (10-12)

30 33 10

Depressivity Emotional Susceptibility Impaired Social Life

18 (16-20) 18 (17-19) 17 (15-18)

17 18

Impaired Daily Life Appearance Short form vs.
PHASE 5: APPLICATION

In the univariate, controlled comparisons of the relationship between thyroid clinical variables and QoL, thyroid volume, dysfunction or eye involvement (i.e. CAS and NOSPECS) were not related to QoL-scores. Thyroid peroxidase antibodies (TPOAb) levels were statistically significantly associated with several QoL dimensions: Goiter symptoms (p=0.024), Depressivity (p=0.004), Anxiety (p=0.004), Emotional susceptibility (p=0.005) and Impaired Social Life (p=0.047). In the path analysis multivariate model, TPOAb was related to the symptom scales (Goiter (p=0.019), Depressivity (p=0.001) and Anxiety (p=0.002), but no longer to Emotional Susceptibility or Impaired Social Life (Figure 8).

**Figure 8 Results of the multivariate path analysis.** ThyPRO scales are ordered in accordance with the theoretical quality of life model. Standardized regression coefficients estimating the strength of the association is presented along each association arrow.
DISCUSSION

PHASE 1: CLINICAL VALIDATION

All scales of the ThyPRO detected clinically relevant differences among thyroid patients had adequate test-retest reliability and the instrument was responsive to relevant changes over time.

Since this is the first QoL measure for use in thyroid patients covering the majority of benign thyroid disorders, no previous studies evaluating measurement properties of such measures exist. Regarding individual diagnoses, test-retest reliability and clinical known groups validity have been evaluated (93) for a disease-specific questionnaire for patients with Graves’ orbitopathy. Test-retest reliability estimates were comparable to the ones found in the present study. However, their clinical validity analyses did not entirely support the validity of the questionnaire, since some of the expected associations with clinical variables were not found (93). For another Graves’ orbitopathy-questionnaire, no associations with clinical variables were found at all (94). To our knowledge, no other study has evaluated clinical validity or test-retest reliability of any of the other existing thyroid-specific (i.e. for individual diagnoses) questionnaires (reviewed in (24)).

The test-retest reliabilities found in the study were a little lower than the coefficients found by internal consistency reliability (36), and thus may in fact underestimate true reliability slightly, possibly due to less than perfect clinically stable conditions of the patients in the test interval. In fact, the mean score was indeed a little lower at time of re-evaluation, for all scales.

Significant longitudinal changes were observed in all instances expected by experts, with three exceptions: 1) Physical symptoms of hypothyroidism did not improve significantly with treatment in the hypothyroid group. One possible explanation for this could be that currently, the most prominent features of hypothyroidism are not physical symptoms such as dry skin or swollen hands or feet, but rather diffuse symptoms such as fatigue and reduced mental well-being. Alternatively, it may take longer than six months for physical consequences of hypothyroidism to resolve after euthyroidism has been reached. Another explanation could be the relatively limited severity of hypothyroidism in this sample. 2) In the goi-

34
ter group, expected improvement of Social Life Impairment was not observed. According to the low baseline level, this is probably due to lack of impact on social function of non-toxic goiter. 3) In the same group, Cosmetic Complaints did not improve either. This is more surprising, since cosmetic concerns would be expected to be the indication for treatment in a substantial proportion of these patients. However, at baseline, the non-toxic goiter group had in fact lower mean scores on the Cosmetic Complaints scale than the two groups with thyroid dysfunction, indicating that cosmetic issues may be of little concern. This is consistent with a qualitative study comparing patients’ and clinicians’ rating of the importance of QoL issues possibly relevant for patients with non-toxic goiter (34). In that study, the clinicians rated dissatisfaction with appearance as the sixth most important issue, whereas the patients rated it as the 39th most important. Masking visible signs of goiter was rated as the seventh most important by the clinicians whereas the patients rated it the 112th most important issue, i.e. not important at all.

The comparison of ThyPRO and SF-36 showed that ThyPRO was more responsive than SF-36 to treatment effect. For seven of the ThyPRO scales, there was no equivalent SF-36 scale: The four physical symptom scales (i.e. Goiter, Hyperthyroid, Hypothyroid and Eye Symptoms), Cognitive Problems, Impaired Sexlife and Cosmetic Complaints. These aspects of QoL were identified through patient interviews and found important by these patients and thus contributes importantly to content validity of the measurement. The two largest effect sizes were, indeed, found among these scales. The SF-36 scales Physical Function and Bodily Pain does not have equivalent ThyPRO scales, since they were not rated important in the patient interviews (34). For all other scales except one, Social Function, the ThyPRO had the highest relative validity. In five of the 12 comparisons where ThyPRO had best responsiveness, the relative validity index of the corresponding SF-36 scales included the value of one, indicating that a larger sample size is warranted for a firm establishment of a larger responsiveness of the ThyPRO on these scales. This was also the case in all three comparisons, where the SF-36 scale (Social Function) had the highest F-value. The high responsiveness demonstrated here is an important verification of the validity of the ThyPRO.

The results concerning responsiveness were in line with most previous studies in other patient groups, where higher responsiveness for disease specific compared to generic instruments has also been demonstrated (41, 95, 96), although there are exceptions. In some studies the generic instrument had the highest responsiveness (97-99). Responsiveness has
previously been demonstrated for a PRO developed specifically for patients with Graves’ orbitopathy (the GOQOL questionnaire) (100), but our study is the first to evaluate the responsiveness of a QoL instrument in the most common thyroid diseases, i.e. patients with hyperthyroidism, hypothyroidism and goiter.

A limitation of the cross-sectional study relates to the clinician-evaluation. Although it was of value that these were available for each individual patient at time of survey completion, the fact that they were also based on patient history, implies that they were not entirely external criteria, as would for example thyroid function tests, etc. be. Still, they do represent evaluations external to the ThyPRO, and clinician evaluations are often used in studies of criterion validity (101, 102). Although the patients were clinically well-characterized and a large number of clinical descriptors were available at time of completion, the timing was not perfect; some patients may have had blood-samples drawn up to about three weeks away from their questionnaire response. However, this would weaken the associations between e.g. thyroid function tests and QOL-data, and yet positive association despite such hypothetical bias was found.

It is a limitation of the longitudinal part of this phase, that the sample size did not allow further sub-division of the effect of various treatments, e.g. anti-thyroid drug treatment, thyroidectomy, and radioactive iodine for hyperthyroidism, which would have profoundly improved our current understanding of evidence-based medicine (103, 104). Such data could have provided information for interpretability, i.e. what magnitude of change to expect as a consequence of a specific treatment, better guide power calculations for future studies, and most importantly help prioritize choice of therapy.

**PHASE 2: PSYCHOMETRIC VALIDATION**

**DIMENSIONALITY**

In the confirmatory factor analyses, each of the ThyPRO scales could be appropriately represented by a unidimensional model after minor revisions. Eleven items were identified in the unidimensional models as potentially misfitting and understood further by multidimensional modeling. Thus, overall the previous initial examinations of the construct validity of the scales (36) were corroborated using a more elaborate technique.

In general, items had high loadings on their own factors and the comparative fit indices were high, but for the majority of the scales, the root means
square error of approximation indicated that a simple unidimensional model was not fitting the data sufficiently well. Based on prior expectations informed by content analyses, modeling results (model inter-item correlations and model residual correlations) and on model modification indices, the models were adjusted in order to reduce the overall misfit. For all scales, an appropriate fit according to the overall goodness-of-fit indices could be reached. During this process, a total of 11 items were left out of the models and 18 residual correlations indicating local dependence were specified.

In most instances, the magnitude of the residual correlations representing local dependencies was small, and the loading on the relevant general factor was still high. Most of the residual correlations were among very similarly worded items. Such local dependencies are not problematic for the current scoring of the ThyPRO, but may lead researchers to overestimate the precision gained by the instrument, because locally dependent items provide less measurement precision than assumed by standard psychometric analyses (105).

Although positively worded items did tend to exhibit residual correlations, we found no consistent evidence of a method factor among the positively worded items. Similar studies with other outcome measures have previously found substantial influence of the value of the wording (69, 106-108), whereas other studies either did not identify such an effect (109) or the identified effect had only minor influence on the results regarding the substantive factor (110).

An attempt to model potential item misfit identified during the dimensionality analyses of the existing ThyPRO scales was made. This was done within a model including all scales, which were allowed to correlate, in order to allow for cross-loadings of items to be examined and in order to evaluate if possible misfit identified during individual scale analyses was due to interrelation with other factors. In doing so, the hypothesized reason for misfit was confirmed in five of seven items: Visibility of the goiter, cross-loaded on Cosmetic Complaints. Loose stools, had a large negative loading on Hypothyroid Symptoms, as had the Hoarseness item. Both constipation and hoarseness are indeed salient and classical features of hypothyroidism (111). The rather non-specific item concerning feeling too fat, which is a common complaint among hypothyroid patients and among hyperthyroid patients after treatment, had cross-loadings on several other scales and low loading on its own factor, also when modeled multi-dimensionally. Thus, these four items are very strong candidates for item
reduction when developing abbreviated and focused versions of the scales or when fitting models where uni-dimensionality is a strong assumption, for example as in unidimensional IRT models.

The use of theoretically driven analyses within a clinically well-described and relatively (for thyroid diseases) large sample was a strength. However, the analyses were carried out in one sample and should ideally be confirmed in a new independent sample. Furthermore, although the present sample comprised patients in all stages of disease and treatment, stability of the factor structure across time could not be evaluated, since the data did not contain longitudinal measurements.

**Differential Item Functioning**

DIF was found in 17 of the 84 tested items. Eight DIFs according to diagnosis were found, eight according to age, one to gender and three according to education.

Most DIFs were small, but two items had larger DIF: 12% of additional variance in the item ‘Visible swelling in front of neck’ and 5.3% for ‘Sensation of fullness in the neck (both from the Goiter Symptoms scale) was explained when adding diagnosis to the regression equation.

Two general patterns of DIF were observed: 1) Only symptom scales were affected by DIF according to diagnosis and 2) Positively worded items were endorsed more readily by younger patients compared to older. Gender DIF was observed only for the item about crying. There was no obvious pattern in the DIFs according to education.

The fact that some items in the symptom scales relate slightly differently to the overall scale in some patient-groups compared to others seems intuitively reasonable. For example, most DIFs in the Goiter Symptoms scale were due to patients with nodular goiter having greater likelihood of endorsing the items for a given level of Goiter Symptoms, compared to patients with the autoimmune diseases. An explanation for this could be that the sensation of the goiter experienced by patients with nodular goiter (i.e. a swelling of the thyroid without accompanying autoimmunity) is different from the sensation of neck symptoms experienced by patients with thyroid autoimmunity. The symptoms in patients with a nodular goiter may be more directly related to the increased size of the goiter (visible swelling, sensation of fullness, pressure), whereas patients with autoimmune diseases experience more diffuse, irritative symptoms (hoarseness,
pain, globulus, swallowing problems, and the need to clear the throat more often).

Also the DIF according to diagnosis regarding the Anxiety scale is clinically interpretable. Patients with non-toxic goiter have a higher probability of a positive response to the item concerning fear of being seriously ill, conditioned on their level of anxiety, than the other patient groups. It is indeed often such a concern that leads these patients to their physician: they discover a swelling of their neck, and fear that it is cancer. So it is a relevant concern, independent of their level of anxiety in general, in contrast to e.g. patients with hyperthyroidism, where anxiety as an affective state is part of the clinical picture.

The findings regarding DIF according to age in the positively worded items raise the question as to whether this is due to a method factor, to which younger patients are more sensitive, or whether positive well-being and energy, are separate dimensions. As mentioned, previous factor analyses supported a one-factor solution for this scale, yielding the former explanation the most likely: younger patients are more influenced by a 'positive wording'-method factor. For the same level of tiredness or depressivity, they are more likely to report positively on positively worded items than older patients are.

Although the concept of DIF concerns group comparisons, and thus is cross-sectional in nature, it is a limitation that the clinical significance of the DIFs found here cannot be tested in a longitudinal setting, since only cross-sectional datasets of a magnitude allowing for DIF analyses (112) were available. If longitudinal data from patients undergoing relevant changes were analyzed, the influence of DIF on the responsiveness to these changes could be tested. Further, it is possible that some items are prone to DIF across time, i.e. that an item functions differentially at baseline compared to follow-up (113).

The overall sample size is fairly large for DIF analyses using ordinal logistic regression. All subgroups had acceptable sample sizes (well above 100 and most around 150, for all groups), except for patients with Graves' orbitopathy (the rarest of the examined diagnoses, n=91) and students (n=12) (73, 114), although one simulation study has suggested a minimum of n=200 in each group (115). However, those analyses did not include any purification, which has been shown to increase detection rates (116).
Previous studies have used different levels of cut-off for the effect size (i.e. \( \Delta R^2 \)); typically higher than the one used here (2\%). Applying a higher cut-off, e.g. 3.5\%, as previously suggested (76), would reduce the number of instances of DIFs to only six. However, simulation studies have indicated that a level of 3.5\% leads to relatively high levels of Type II errors and substantial instances of DIF may thus be missed (117, 118). Moreover, the fact that a well-known DIF, as the gender/crying DIF (119-122) is identified only if the 2\% level is maintained, corroborates the appropriateness of this level. Another limitation was the fact that analyses were only performed in one sample; ideally, they should be re-tested in an independent sample. However, this was somewhat compensated for by the fact that the sample spanned a broad range of thyroid diagnoses, allowing for comparison across all the relevant diagnoses.

**PHASE 3: CROSS-CULTURAL VALIDATION**

Very good cross-cultural validity was found, as might also be expected due to the rigorous translation methodology applied. Hardly any difference was identified between the English, Danish, Dutch and Swedish versions of the questionnaire. Small differences were found in the Italian, Serbian and Indian versions, which were most likely caused by differences in thyroid sub-diagnoses. Other differences may reflect cultural differences in conceptualizations of specific item content in relation to the concept measured by the scale. However, none of these instances of DIF led to substantial change in the summary scale score, as estimated by the expected mean scores modeling the DIFs.

Twelve items had differential item functioning in more than one country, indicating possible low translatability or cross-cultural equality of these items. However, differential item functioning is only potentially harmful to the validity of a measurement, if it has an impact on the scale score. Although the DIFs identified in this study are above the explained variance cut-off level for substantiality, they have little impact on the ThyPRO scale scores. The reason for the limited effect on scale score is that the items are part of multi-item scales of at least four items. So for “Visible swelling in neck”, for example, the difference in mean item scores between the two extremes (Hindi and Swedish) are 1.7 item score levels. With 11 five-point items in the Goiter Symptoms scale this difference corresponds to 3.9\% of the 44 raw score range of the scale, i.e. a difference of 3.9 points on the transformed 0-100 range scale.
The cross-cultural equivalence was higher than expected, since the study encompassed patients from as different cultures as e.g. the Danish (the majority in this study) and the Indian. A possible reason for this may be the universality of the concepts measured, embedded in a unifying clinical context (thyroid disease). This was also reflected in the early, qualitative phases of the project, where the rephrasing of item content by patients were very much in line with the pre-specified and rephrased content of the originals.

A limitation was the clinical sample dissimilarity, due to different sampling strategies, which was a consequence of practical issues within each participating site. Since different diagnoses, as shown above, did indeed cause DIF in the ThyPRO, one might question whether the identified DIF's are consequences of the differences in sample distributions, rather than cross-cultural differences (123). However, against such a concern speaks the fact that some of the samples which were clinically most dissimilar to the English-speaking sample (i.e. Dutch, Danish and Swedish) had virtually no DIF. Thus, the DIF identified in this study are likely to reflect cross-cultural and linguistic differences, rather than clinical differences.

Another concern could be that the determination of DIF based on its presence in more than one country may be overly simplistic and potentially misleading. However, if the present results are used as guidance to select items for a shorter version, avoiding these items could optimize the translatability of such a shorter version.

These results could have implications at various levels. It could have implications for the level of comparability of studies conducted in different countries; in terms of analysis strategies for multi-language clinical studies (e.g. international multi-center clinical trials) and in terms of revised or abbreviated versions of the ThyPRO. Further, the results could document the extent to which it is justified to extrapolate validation studies conducted in one language to other languages. In terms of the former, the degree of DIF among these language versions are of such small magnitude, that it does not hamper comparisons of studies conducted in different countries. In international multi-center studies, one should consider taking the four major instances of DIF into account in the analysis strategy, by conducting sensitivity analyses of the results when the scales are scored without the items displaying DIF. Alternatively, one could adjust the scale scores by the magnitude and direction of the DIF (113). Howev
er, in clinical trials this will seldom be a significant problem, since participants are usually stratified by center in the randomization procedure.

**PHASE 4: ABBREVIATION**

Based on previous validation studies and IRT-modeling, an abbreviated version of the ThyPRO was developed containing 1) four physical symptom scales, two of which with three (Goiter and Eye Symptoms) and two with four items (Hypo- and Hyperthyroid physical symptoms), 2) seven three-item scales about physical, mental and social well-being and function, 3) one three-item scale concerning appearance and 4) one single item about impact on overall quality of life. Thus, the abbreviated version consists of thirty-nine items, if all physical symptom scales are administered. Each of the twelve short-form scales and the single QoL item can be reported separately, but the seven well-being and function scales can also be summarized in one single composite score.

The subsequent validation analyses showed that the abbreviated scales had very high agreement with the original long-form ones, including roughly similar mean levels, and comparable measurement quality. Thus, good test-retest reliability, responsiveness to clinical change and sensitivity to relevant clinical differences were demonstrated. The preservation of good measurement properties in scales with much fewer items is considered a result of selection of items with best measurement properties, under consideration of the conceptual model and content validity, thereby reducing random and systematic measurement error.

It is a limitation, that the short-form has not yet been tested as a standalone form in an independent, novel clinical sample. Further, although the aim was to develop and test the instrument in a broad, heterogeneous sample, as specified in the introduction, and although the cross-sectional sample size was fairly large, it was not large enough to permit multi-group analyses (61) according to diagnosis. Application of a short-form may lead to loss of content validity. The extent to which this has occurred can only be evaluated in qualitative studies (124). However, since the individual scales were found to be uni-dimensional, the potential loss should theoretically be minimal. As evident from the agreement plots, the short-form scales have fewer measurement points along the entire spectrum and application of the short-forms may also lead to poorer discrimination at the extremes. Another potential weakness was the fact that five of the scales were slightly modified (some items were omitted) to avoid item level misfit in the IRT-model. Since the rescaling was based on these IRT ana-
lyses, this may lead to weaker linking between the two versions. On the other hand, as mentioned, the correlation, agreement and mean levels among the two versions of each scale, all supported the appropriateness of the present linking.

The applied approach is in line with recent recommendations for item reduction (56). When reviewing the available item-reduction literature, the authors found, that 55% of the studies had preserved scale structure and the median proportion of reduction was 57% (range: 21-88%). The present study was close to this median reduction (from 85 to 39 items, i.e. 54% reduction). In 62% of the studies, only the long-form was administered. Use of IRT-methods was recommended as advantageous in the suggested guidelines, but was only applied in 11% of the studies.

The two-level scale-scoring approach, where both a composite score and the underlying more detailed sub-scales can be scored for the well-being and function scales, has also been adopted in previous studies. The most prominent is the most widely used short-form measure, the SF-36 Health Survey (125). Based on SF-36, eight domain scores as well as two Component Summaries can be derived (126), depending on the level of detail required in reporting. The scoring of the SF-36 summaries is based on results from principal component analyses, in contrast to the present study, where a simple summation approach was adopted for ease of scoring and reporting.

PHASE 5: APPLICATION

When evaluating the relationship between clinical variables and QoL, TPOAb level, but not thyroid function, was related to thyroid-specific QoL. This raises the hypothesis that thyroid autoimmunity may play a role, independent of thyroid dysfunction, for the QoL impairment associated with autoimmune hypothyroidism. It is surprising, that no relationship between thyroid function and QoL was found. However, this is probably due to the fact that the vast majority of patients were only marginally dysthyroid at time of participation, as a consequence of the sampling procedure spanning all disease-stages. We would expect this to be different, if more patients with current overt hypothyroidism had been included. Indeed, Canaris et al. (127) and others (128, 129) have demonstrated a relationship between hypothyroid symptoms and thyroid function, when investigating patients with overt hypothyroidism and euthyroid controls. Carle et al (130) have recently shown, that symptoms associated with hypothyroidism are indeed only weakly associated with a new diagnosis of
overt hypothyroidism. Another recent study among 597 patients receiving L-thyroxine (73% of whom had autoimmune hypothyroidism), found an association between psychological well-being and fT4 and serum TSH, but not fT3 or TPOAb-positivity (131). In a recent Danish study, no association between TPOAb-positivity and symptoms of depression was found (132) either. However, in both of these latter two studies, TPOAb were analyzed as a dichotomized, not as a continuous variable as in our study, which could explain this difference. In accordance with the present findings, Ott et al. demonstrated a correlation between symptoms and level of TPOAb, but not thyroid function tests, in patients thyroidectomized for benign goiter (133). Another field of research has focused on the role of thyroid autoimmunity in musculoskeletal complaints, particularly fibromyalgia (134, 135). For example, Bazzichi et al. found more symptoms among patients with fibromyalgia, who also had autoimmune thyroiditis (134). Additionally, in a population-based study, higher prevalence of TPO-Ab positivity was found in respondents with musculoskeletal complaints, compared to those without (136). Finally, microvascular alterations have been found in skeletal muscle from patients with autoimmune thyroiditis, independent of thyroid function (137), suggesting a possible mechanism of action.

Several studies have indicated a link between the thyroid gland and mental diseases (138). As for autoimmune thyroid diseases, Pop et al. (139) found a three times higher risk of current depression (according to the Edinburgh Depression Scale) in individuals with positive TPOAb in a large population-based study, and presence of TPOAb during pregnancy has been identified as a risk factor for postpartum depression (140). A direct relationship between thyroid autoimmunity and QoL has also been hypothesized to be at play in the rare and controversial diagnosis of Hashimoto’s encephalopathy, where a variety of neurological symptoms have been linked to the presence of TPOAb (141).

The main limitation of this part of the project was its cross-sectional design, the relatively limited range in thyroid dysfunction, and the few severely hypothyroid individuals. Future hypothesis testing studies should adopt a longitudinal design, compare changes in all available thyroid autoantibodies, or even better various epitopes of these antibodies (142-146), and in thyroid function, with changes in QoL. Also, the relationship between such indicators of thyroid autoimmunity and QoL could be evaluated in a sample of patients with autoimmune thyroiditis and normal thyroid function. Applying diagnostic imaging, such as e.g. MR-spectroscopy (147), may offer further insight into the possible CNS-effect and thereby
the link with affected QoL measures. Future interventional research may attempt to target the autoimmune component of the disease more directly, as in the GRASS trial. Future studies focusing on QoL methodology could further develop and investigate the theoretical framework and its relationship with clinical variables.

OVERALL PROCESS AND FUTURE DIRECTIONS

At an overall level, several other paths towards an international thyroid-related PRO could have been followed.

First of all, two decisions were made at an early stage, which had major implications for the outcome: the instrument should not be specific to individual thyroid diagnoses, as argued for in the introduction, and it should not measure only disease-specific issues of relevance, but all issues of relevance, including generic ones. Alternatively, the instrument could have been limited to only thyroid-specific dimensions (and one instrument per diagnosis could have been made) and could be administered along with the generic instrument which best covered the relevant issues. With the present approach, the instrument can stand alone, independent of a particular generic measure with potential irrelevant items, and researchers can be assured that relevant issues are addressed. In addition, it permitted flexibility to adapt the instrument and e.g. develop an abbreviated measure with only 39 items, still being comprehensive: standard generic instruments are not eligible for such modifications. However, it is at the expense of comparability with other diseases: direct comparability would require an additional process, such as linking the ThyPRO scales measuring generic aspects to e.g. the generic SF-36 scales.

Although the development ran partially in parallel for some languages, not all languages were included from the beginning. This reflects the dynamic process during which the project grew with time and the limited sources initially. A more elaborate set-up could ideally have developed all language versions in parallel, which would probably have increased the cross-cultural validity and exploited the resources better (55, 148).

The ThyPRO was developed for fixed-length, summated scoring. A more elaborate method would be to develop item banks for each dimension and score the scales using computerized adaptive testing (149, 150). With such an approach, a variable number of items are administered and the selection of items is based on the responses to previously administered items.
Scoring is based on IRT-parameters and thus a more precise measurement can be obtained with fewer items. Further, the process can be preset to a specific level of precision, depending on the purpose and dimension of a study. However, computerized adaptive testing requires a set-up much larger than what was obtainable for our purpose. IRT-based scoring for the fixed-format version could have been adopted, but a simple summation approach was selected due to ease of scoring and reporting.

**OTHER THYROID-SPECIFIC PRO MEASURES**

A number of alternative instruments assessing patient-experiences of thyroid diseases are available (24, 151, 152). Many of these are not patient-reported, but scoring systems for clinicians, and focus on specific, physical symptoms. These instruments, including the physician-administered symptom scores for completeness, are reviewed below.

**PHYSICIAN-ADMINISTRATED SYMPTOM SCORES**

*Clinical ratings scales for patients with non-toxic goiter*

No rating scale for patients with non-toxic goiter has been identified.

*Clinical ratings scales for patients with hyperthyroidism*

In the 1950’s a Scottish group, comprising among others Crooks, Wayne, Murray and Billewicz, worked with clinical scoring systems for thyroid patients. First, they described their development of a diagnostic index of hyperthyroidism (153-156), subsequently referred to as either the ‘Crooks index’ (157-161) or the ‘Wayne index’ (162-164). It included 12 symptoms and 11 signs, recorded by the physician, in a weighted scoring system (Table 7). Later, the index was modified and shortened (to 8 symptoms and 4 signs), in order to make it more sensitive to the effect of treatment (159, 165, 166). Evaluation of the instrument focused on its diagnostic rather than its psychometric characteristics (154, 155). Apparently, it is the original index, rather than the modified one, which has been applied in subsequent trials.

The index was modified by Gurney and colleagues (164) to yield the *Newcastle Index*. They included only one of the original symptoms and seven of the signs and added three psychological symptoms and age of onset, in order to make the index more applicable to patients with psychiatric comorbidity. Another modification was made by Benvenga and colleagues (161), who changed the scoring method from the original summation of
weighted item-scores to a 5-point score given for each symptom/sign. Further, they reduced the original number of symptoms from 12 to 5 and added insomnia, reduced the signs from 11 to 2 and added knee reflexes, yielding 6 symptoms (asthenia, dyspnea, insomnia, nervousness, palpitations, weight change) and 3 signs (tremor, knee reflex, heart rate).

No subsequent studies seem to have applied any of these modified indices.

**Table 7 Wayne/Crooks Diagnostic Index of Thyrotoxicosis** (from (156)).

Scores of +20 or more indicate thyrotoxicosis; scores of +10 or less indicate euthyroidism.

<table>
<thead>
<tr>
<th>Symptoms of Recent Onset and/or Increased Severity</th>
<th>Present Score</th>
<th>Absent Score</th>
<th>Signs</th>
<th>Present Score</th>
<th>Absent Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyspnoea on effort</td>
<td>+1</td>
<td></td>
<td>Palpable thyroid</td>
<td>+3</td>
<td>-3</td>
</tr>
<tr>
<td>Palpitations</td>
<td>+2</td>
<td></td>
<td>Bruit over thyroid</td>
<td>+2</td>
<td>-2</td>
</tr>
<tr>
<td>Tiredness</td>
<td>+2</td>
<td></td>
<td>Exophtalmus</td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>Preference for heat</td>
<td>+5</td>
<td>-5</td>
<td>Lid retraction</td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>Preference for cold</td>
<td></td>
<td></td>
<td>Lid lag</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>Indifference to temperature</td>
<td>0</td>
<td></td>
<td>Hyperkinetic movements</td>
<td>+4</td>
<td>-2</td>
</tr>
<tr>
<td>Excessive sweating</td>
<td>+3</td>
<td></td>
<td>Finger tremor</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>Nervousness</td>
<td>+2</td>
<td></td>
<td>Hands: Hot</td>
<td>+2</td>
<td>-2</td>
</tr>
<tr>
<td>Appetite increased</td>
<td>+3</td>
<td></td>
<td>Moist</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>Appetite decreased</td>
<td>-3</td>
<td></td>
<td>Casual pulse-rate:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight increased</td>
<td></td>
<td></td>
<td>Atrial fibrillation</td>
<td>+4</td>
<td></td>
</tr>
<tr>
<td>Weight decreased</td>
<td>+3</td>
<td></td>
<td>Regular rates:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;80 per minute</td>
<td>-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80 to 90 per minute</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&gt;90 per minute</td>
<td>+3</td>
<td></td>
</tr>
</tbody>
</table>

The Crooks’ index was used by Paschke and colleagues in a study of the psychological performance of 15 Graves' patients in the course of treatment (157). No correlations with an array of psychological parameters were found, apart from the anxiety-scale STAI-X1. No explicit scores were reported and no validity considerations presented. Schlote and colleagues used a German translation of the index and found successively increasing scores among overtly and subclinically hyperthyroid patients compared to euthyroid controls (158). Further, moderate correlations between thyroid hormones and the index were found. Likewise, Stott and colleagues found higher scores in 15 patients with subclinical/mild hyperthyroidism and 10 overtly hyperthyroid patients, compared to 10 euthyroid controls (163). Sgarbi and colleagues used the index in their study of methimazole treatment of 10 patients with subclinical/mild hyperthyroidism (162). Significantly higher baseline scores were found among patients, compared to 10
matched controls. The Wayne index correlated with TSH, but not with peripheral thyroid hormones.

Klein and colleagues developed the 10-item Hyperthyroidism Symptom Scale (HSS) on the basis of Crooks' index (167). It consists of seven items about classical hyperthyroid symptoms, one item specific to limitations in daily function and two physical signs. The scoring was changed from Crooks' dichotomous response options to ratings on a five-point scale and all items are summed to provide a single overall score. In a longitudinal study of ten initially untreated hyperthyroid patients, HSS scores declined significantly as a result of treatment. Further, patients initially scored significantly higher on the HSS than ten control patients with diabetes, indicating sensitivity of the scale as well. Additionally, high inter-rater and test-retest reliability (intraclass correlation coefficients 0.94 and 0.90 respectively) was demonstrated. The HSS has been used in a number of subsequent studies (168-177). Results from several of these studies confirm that the index is sensitive to relevant clinical differences between patients and responsive to relevant clinical changes within patients over time.

Clinical ratings scales for patients with Graves’ orbitopathy
A number of clinical grading systems for Graves’ orbitopathy exists, e.g., CAS, NOSPECS, VISA, EUGOGO system (178), but they are clinical grading systems for use in management, including decision making, rather than symptom assessments, and are not reviewed here.
Clinical ratings scales for patients with hypothyroidism

The Scottish group developing the Crook's/Wayne index of hyperthyroidism also developed a clinical scoring system to diagnose hypothyroidism (128, 155, 156). It has subsequently been referred to as the Billewicz index and consisted of 8 symptoms and 6 sign resulting in a weighted total score ranging -53 to 72 (Table 8).

Table 8 Billewicz' Diagnostic Index of Hypothyroidism (from (128)). Scores of +25 or more indicate hypothyroidism; scores of -30 or less indicate euthyroidism.

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Present Score</th>
<th>Absent Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diminished sweating</td>
<td>+6</td>
<td>-2</td>
</tr>
<tr>
<td>Dry skin</td>
<td>+3</td>
<td>-6</td>
</tr>
<tr>
<td>Cold intolerance</td>
<td>+4</td>
<td>-5</td>
</tr>
<tr>
<td>Weight increase</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>Constipation</td>
<td>+2</td>
<td>-1</td>
</tr>
<tr>
<td>Hoarseness</td>
<td>+5</td>
<td>-6</td>
</tr>
<tr>
<td>Paraeasthesiae</td>
<td>+5</td>
<td>-4</td>
</tr>
<tr>
<td>Deafness</td>
<td>+2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signs</th>
<th>Present Score</th>
<th>Absent Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow movements</td>
<td>+11</td>
<td>-3</td>
</tr>
<tr>
<td>Coarse skin</td>
<td>+7</td>
<td>-7</td>
</tr>
<tr>
<td>Cold skin</td>
<td>+3</td>
<td>-2</td>
</tr>
<tr>
<td>Periorbital puffiness</td>
<td>+4</td>
<td>-6</td>
</tr>
<tr>
<td>Pulse rate less than 75 per min.</td>
<td>+4</td>
<td>-4</td>
</tr>
<tr>
<td>Ankle jerk</td>
<td>+15</td>
<td>-6</td>
</tr>
</tbody>
</table>

The diagnostic performance of the index was subsequently evaluated by an Indian group (179). They found diagnostic properties inferior to the ones demonstrated by the developers. However, the 'Billewicz index' which they examined, included three symptoms that were excluded by the Scottish group in their original work, based on the diagnostic inefficiency of those items.

The Billewicz' index has been used in a number of subsequent studies (175, 180-195).

Zulewski and colleagues simplified the scoring of the index by replacing the weights with simply 1/0 for absence of symptoms (196). This scoring method, named the 'Zulewski score', has been adopted in several subsequent studies (197) (182, 184, 198-203).

In light of the widespread use of the Billewics/Zulewski score, it is surprising, that the content validity or other aspects of validity have never been formally examined. Further, the appropriateness of using it as a measure of symptom severity and to evaluate treatment effect has not been examined either.

Cooper and colleagues used a physician-rated symptom score, which they modelled after the Billewicz index, in a randomized clinical trial (RCT) of
L-thyroxine treatment of subclinical hypothyroidism (204). It comprised six symptoms, of which three was also included in the Billewicz index (dry skin, cold intolerance, constipation), and six additional, which were not (muscle cramps, poor energy, easy fatigability). A baseline score was derived by summation of present symptoms (i.e. range 0-6) and a follow-up or "transitional score" was constructed by summation transition scores for each symptom, ranging -2 for worsening of symptom to +2 for resolution of that symptom. Initial scores were higher among patients, compared to controls, and scores improved with active vs. placebo treatment. Twelve years later, the questionnaire was used in a similar study, without reproduction of the results (205).

**PATIENT-REPORTED THYROID-SPECIFIC INSTRUMENTS**

*Instruments for patients with non-toxic goiter*
No goiter-specific instruments apart from ThyPRO have been identified.

*Instruments for patients with hyperthyroidism*
The Hyperthyroidism Complaint Questionnaire (HCQ) measures “somatic and mental discomfort…..typical for patients suffering from hyperthyroidism” (206) and consists of 31 dichotomous items requiring a dual response: currently present and formerly present; two overall summary measures are generated: Past and Present Complaints, respectively. Eleven items concern physical symptoms, 6 are about emotional distress, 6 evaluate fatigue, 3 concern cognitive function whereas existential problems, sleeping problems, anxiety, sexual function and social function were covered by one item each. The development was based on interviews with a small sample of patients (not documented). In a study including 303 patients previously treated for hyperthyroidism, Cronbach’s α was 0.93 for both scales. Item-total correlations were generally low, some were 0.21, questioning the justification of the overall summary score. A relationship between overall score and degree of self-reported thyroid dysfunction was found, but no further description of the validity of the instrument has been provided. The HCQ has not been used in any subsequent study.
Instruments for patients with Graves’ orbitopathy

The 

Graves’ Ophthalmopathy Quality of Life Questionnaire (GOQOL) is a disease specific QOL-instrument for patients with Graves’ orbitopathy (93, 100, 207). The development was based on a review of existing eye QOL-measures, as well as open-ended questionnaires from 24 patients and was pretested in 8 patients. Detailed description of these content validity studies have not been published. The GOQOL consists of 16 items summarized in two scales: ’Visual Functioning’ and ’Cosmetic Complaints’; i.e. two distal QoL-concepts, albeit not as distal as “overall quality of life”. No proximal concepts such as physical symptoms are measured. Subsequent studies comprising 70-164 well-described patients have shown excellent test-retest and internal consistency reliability (93, 207) and have supported its construct validity (93, 100, 207) in terms of exploratory factor analyses and examinations of con- and divergent validity. Further, good responsiveness was found (100). Modern psychometric methods (DIF, IRT) have not been applied. According to the European Group of Graves’ Orbitopathy (EUGOGO) website, the GOQOL is available in 15 languages (www.eugogo.eu). The Korean version has been clinically (not cross-culturally) validated (208). GOQOL has been applied in numerous subsequent clinical studies (209-223). Cross-cultural validity does not seem to have been quantitatively assessed, but a forthcoming validation of the English version is described in a published trial protocol (224). A non-validated modified version has also been applied in one clinical study (225)

In an attempt to develop a measure with high correlation with clinical variables, Yeatts developed the nine-item GO-QLS questionnaire (226). The development was based on analyses of responses from 256 patients to 105 existing generic and eye-specific items. The best discriminating items fitting a one-factor analysis were selected. The resulting nine-item measure does not seem to have been implemented in subsequent clinical studies yet.

Recently, a three-item Graves’ orbitopathy-specific PRO has been developed, the TED-QOL (227). It consists of three visual analogue scale-like single items measuring overall QoL impact, interference with daily activities and satisfaction with appearance, respectively. The development was very briefly described, test-retest reliability was found to be good, convergent and discriminant validity was confirmed, compared with similar or dissimilar aspects from GO-QOL and GO-QLS and meaningful correlation with clinical variables were found. It has been translated and validated (not cross-culturally though) in Korean (228), but clinical studies applying
the measure or studies reporting other important aspects of validity such as responsiveness, have not yet been published.

**Instruments for patients with hypothyroidism**

For the purpose of detecting patients developing hypothyroidism in a follow-up system after radioiodine-treatment, Barker and colleagues used a patient-completed questionnaire about nine classical symptoms of hypothyroidism (129, 229). The evaluation of the instrument focused on its screening abilities, i.e. ability to detect biochemical hypothyroidism. As with the Zulewski index, a simple summation of the number of symptoms were as effective as more complex weighting and scoring systems and if patients with five or more symptoms was referred to further diagnostic procedures, this would include 70% of the hypothyroid and 40% of the euthyroid patients.

For the analysis of the relationship between symptoms and biochemical thyroid status, Canaris and colleagues developed a 16-17 item hypothyroid symptom index (127), sometimes referred to as the **Colorado Thyroid Symptom Survey**. The questionnaire was developed on the basis of a combination of a literature study and expert evaluation. It was tested for readability and with regards to its ability to discriminate between hypothyroid and euthyroid patients. A modest correlation with TSH was found. A modified version of the questionnaire was used in the large Colorado Thyroid Disease Prevalence Study (230) and in a subsequent screening study (231). Generally, low discriminating properties of individual symptoms were found. The proportion of hypothyroid patients increased with increasing number of symptoms present and with increasing weighted score; the total score had moderate discriminative power.

The **Chronic Thyroid Questionnaire (CTQ)** is a hypothyroidism and patient-specific QOL-questionnaire. It consists of 104 items, each representing a specific complaint, covering four domains: 'Physical Complaints', 'Mood and Emotions', 'Energy and General Well-Being', and 'Cognitive Complaints' (205, 232). The development of the CTQ was quite thorough. Based on a literature review, a list of symptoms or problems related to hypothyroidism, potentially responsive to treatment and likely to influence the quality of life of the patients, was generated (232). This list was expanded through interviews with endocrinologists and patients. The scoring of the CTQ is unusual: Of the 104 complaints, each patient identifies applicant items and rates the degree of discomfort represented by these items. Thus, for a patient with two of the 104 complaints, the instrument consists of two items, whereas a patient with 22 complaints
rates 22 items. This approach increases the potential sensitivity of the measure to improvements in the individual patient, but it makes between-patient comparisons and interpretations of what is actually measured difficult and new complaints arising from intervention are ignored in longitudinal studies. The CTQ has not been validated in any subsequent studies, but applied in one subsequent study (233) and in an unclearly modified version in two other studies ((234, 235).

The **Thyroid Symptom Questionnaire (TSQ)** consists of twelve items: six items on cognitive complaints, five items on physical symptoms and one item on fatigue, summarized in one overall score (236). The items were selected on the basis of patient-responses to a notice in the British Thyroid Foundation newsletter, inviting patients to tell about persisting complaints despite replacement therapy with L-thyroxine. Moderate correlations with the generic QOL questionnaire General Health Questionnaire (GHQ-12) were found, but no other evidence of validity has been presented. The measure does not seem to have been adopted in subsequent clinical studies.

The **Underactive Thyroid-Dependent Quality of Life Questionnaire (ThyDQoL)** (237). ThyDQoL is a 20-item questionnaire measuring impact of hypothyroidism on various domains of QoL: overall QoL (two items), limitations in usual activities (six items), social function (four items), fatigue (two items), emotional well-being (two items), sexual function, cosmetic complaints, weight problems, and bodily discomfort (one item each). Items are scored individually in a two-step procedure: both impact and importance of the items are rated, and the item score is derived by multiplication of these two ratings. An average weighted impact score is obtained by summing all applicable weighted domain scores (i.e. not the two overall QoL items), divided by number of domains applicable to the individual. Content validity was ensured through interviews with 38 hypothyroid patients, as part of the development process (237). In a subsequent validation study, appropriate internal consistency reliability was found. Construct validity in terms of scale validity as evaluated by exploratory factor analyses was also examined. Although the authors conclude that an identified two-factor solution was un-interpretable, it could be argued, that a picture of one set of items clustering around an aspect of participation and another around symptoms and well-being within the patient, was indeed drawn (Table 3 in (238)); but a one-factor solution also seems well justified. Criterion validity was not explicitly addressed, but a tendency towards worse perceived negative impact on QoL was observed among patients with overt hypothyroidism, compared to patients with subclinical
disease. One problem with the two-step importance rating approach is the reduced inter-individual comparability of the measure and the susceptibility to a confounding effect of coping. These results regarding validity have been reproduced for the German version (239). In that study, also criterion validity was evaluated (correlation with clinical variables), and only partial support for criterion validity was found: of note, however, only 25 patients and 27 healthy controls were included in the criterion validation study. ThyDQoL has been applied in a clinical study evaluating effect of L-thyroxine on subclinical hypothyroidism (240) and in a study of impact of transient hypothyroidism among patients with thyroid cancer (195). In the latter, scores worsened during hypothyroidism, supporting the clinical validity of the measure. Also, it is specified as an outcome in a published trial protocol (241).

During the development of the ThyDQoL, symptoms related to hypothyroidism were extracted into a separate instrument, the Thyroid Symptom Rating Questionnaire, ThySRQ (237). The motivation was the observation, that the symptoms were "too specific in nature to be important for many aspects of life, e.g., voice problems, or because some patients were unsure whether they were attributable to hypothyroidism". ThySRQ consists of 15 items, covering both physical, psychological and functional (tiredness, cognition) symptoms. Items are reported separately, since factor analyses could not support summation (238). Similarly to ThyDQoL, items are completed in a two-step procedure: first, whether it is experienced, then, the extent to which it causes distress. To the authors' surprise, high internal consistency reliability was found. Along with ThyDQoL, it was applied in patients with subclinical hypothyroidism (240) and is included as outcome in a published trial protocol (241). A simplified version was applied in a study in Brazil; no information regarding method of translation or validation thereof was presented (242). In a third study applying the measure, an average sum-score was analyzed and reported, despite the evidence against the validity of such an approach (195).

In parallel with ThyDQoL and ThySRQ, a Thyroid Treatment Satisfaction Questionnaire, the ThyTSQ, was developed (237). The questionnaire has two parts: seven items measuring satisfaction with present treatment and four items with past treatment, each rated from 6 (very satisfied) to 0 (very dissatisfied). In a validation study, excellent internal consistency reliability, good acceptance and evidence of unidimensionality, justifying reporting as a single summated scale, for the two multi-item scales (Past
and Present Satisfaction) (243). It has been applied in the abovementioned L-thyroxine study (240) and in a published study protocol (241).

**THYPRO AND THYPRO-39 IN COMPARISON**

Of the abovementioned instruments, only few have been developed in accordance with current standards, as shown in Table 9: Three instruments for patients with Graves` orbitopathy: GOQOL, GO-QLS and TED-QOL and one triplet of instruments for patients with hypothyroidism, ThyDQoL, ThySRQ and ThyTSQ. Thus, ThyPRO (and ThyPRO-39) is the only PRO validated for use among patients with non-toxic goiter and hyperthyroidism.

Compared to the abovementioned, ThyPRO differs in a number of perspectives:

*Development strategy.* The aim of developing an instrument for all benign thyroid diseases is unique to ThyPRO. Further, it is the only instrument developed without a pre-existing starting point. The development of the other instruments was based on existing questionnaires, either in full format (ThyDQoL instruments were modelled over a diabetes instrument (ADDQoL)) or by sampling and subsequently testing and revising items from existing instruments (GOQOL and TED-QLS). Development of ThyPRO was based only on a theoretical measurement framework and input from literature, clinicians and patients. Input from patients has been very substantive, not only in establishing content validity, but also in elaborated rounds of cognitive interviews, revising questionnaire format and content iteratively. Further, a parallel multi-language development process is has also only been adopted for ThyPRO.

*Level of validation.* Compared to the other relevant instruments, ThyPRO is by far the most intensively validated, as evidenced by the much larger number of studies reporting on aspects of its development and validity (15, compared to maximum 4, Table 9). Also, although construct validity has been evaluated to some extent for other instruments, as shown in Table 9, this was limited to one principal component analysis or simple item-total correlation, whereas for ThyPRO several full-focused scale validation studies have been published separately. Also unique to ThyPRO is the systematic and full criterion validation, i.e. confirmation of all hypothesized clinical differences for every individual scale. No other instrument has been validated using IRT-methods.
Table 9 Existing thyroid-related PRO instruments. Description of development and validation, including number of associated studies published per January 2016, in accordance with the terminology presented in Table 1

<table>
<thead>
<tr>
<th>Measurement property documented</th>
<th>Thy PRO 39</th>
<th>Thy PRO 39</th>
<th>GO-QOL</th>
<th>GO-QLS</th>
<th>TED-QOL</th>
<th>Bar-</th>
<th>Can-</th>
<th>CTQ</th>
<th>Thy D QoL</th>
<th>Thy SRQ</th>
<th>Thy TSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual model</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Reliability:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intern. consist. reliability</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Test–retest reliability</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Validity:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>(+)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Construct validity</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>(+)</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>(+) (+)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(±) (±)</td>
<td>+</td>
</tr>
<tr>
<td>IRT models applied</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Interpretability of scores</td>
<td>(+) ±</td>
<td>(+) ±</td>
<td>-</td>
<td>(+)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MID</td>
<td>(+) ±</td>
<td>(+) ±</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Burden (n of items)</td>
<td>85</td>
<td>39</td>
<td>31</td>
<td>16</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>16</td>
<td>104</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>Language versions</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Target population:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-toxic goiter</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Toxic nodular goiter</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Graves’ hyperthyroidism</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Graves’ orbitopathy</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Domains covered:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical symptoms</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Psychological symptoms</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Well-being</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Function</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sexlife</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Participation</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Overall QoL</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Treatment satisfaction</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Methodological publications</td>
<td>15</td>
<td>15 h</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Application publications</td>
<td>11</td>
<td>11</td>
<td>(1)</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1 (+2)</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

aA total score was not fully justified by factor analyses
bNo significant relationship with clinical variables, but study design not specifically aimed at criterion validity
cPRO scores correlated in a meaningful direction with some clinical variables. Criterion validity only partially supported.
dOnly applicable to patients currently receiving treatment
eMID has been published as a conference paper and responsiveness data as well as general population norms offer interpretability-information
fCross-cultural validity not yet quantitatively assessed
gNot based on patient-interviews
hHigh level of agreement, repetition and nestedness allows extrapolation from the full version validation studies
Cross-cultural validity. No other thyroid-related PRO has quantitatively evaluated cross-cultural validity.

Concepts measured. As evident from Table 9, some instruments focus on very specific manifestations of disease impact, e.g. physical symptoms (Barker and Canaris indices for example), whereas other focus on broader aspects of QoL, e.g. impact on social functioning (TED-QLS for example). These broader aspects may also be termed “distal” concepts (corresponding to the right, “downstream” part of the QOL-model (Figure 1)), whereas aspects more closely associated with a particular condition may be referred to as “proximal” aspects (corresponding to the left part of the QOL-model). Disease- or condition-specificity of an instrument may in fact be obtained with both categories. Proximal outcomes may do so by selecting symptoms specific for, or very relevant to, a specific condition. Distal outcomes may acquire specificity by attributing the concept in focus to the relevant condition. As an example of the former approach, the Canaris index measures symptoms characteristic of hypothyroidism. In contrast, the TED-QLS asks about impact of eye-disease on overall QoL, daily function and well-being, using only one item for each of these. This latter approach is tempting, due to the ease of administration, and because it can easily be applied to a wide variety of diseases, using the same format. However, with this approach, precision and reliability is a concern. Moreover, the standard approach for ensuring content validity may be less suited in this situation: In a study comparing TED-QLS with GOQOL and GO-QLS, patients were asked if the items covered their situation. Since the questions are so broad, it is difficult to bring up an issue, which cannot be categorized under one of these aspects. In such a situation, a more elaborate approach is needed, to evaluate content validity. For example, in-depth cognitive interviewing, exploring whether patients truly consider all relevant sub-aspects, when they formulate their replies to such broadly formulated questions. Proximal items offer the advantage of providing cues for comprehension and retrieval to the respondent. As described previously, ThyPRO includes both proximal and distal aspects, reported separately. Proximal aspects are administered first, to stimulate retrieval, and distal aspects are administered in the end of the questionnaire. Aspects measured by other instruments not covered by ThyPRO are “treatment satisfaction” and “positive effects of the disease”, measured in ThyDQoL and ThyTSQ, respectively. Thus, studies focusing on treatment satisfaction among patients with hypothyroidism could use the ThyTSQ.
Length. As a consequence of the abovementioned development strategy, the ThyPRO is long. Even the abbreviated version is rather long, compared to other instruments (39 items vs. maximum 20). Nonetheless, very high completion rates are observed in ongoing clinical trials (>90%) probably reflecting the uniform design of the ThyPRO items, elaborate cognitive testing and revising as well as the use of mixed modes of administration (21). Still, the length is a concern for many scientists interested in using the ThyPRO (personal communications).

**FUTURE DIRECTIONS**

**FUTURE CLINICAL TRIALS**

Availability of a validated PRO for patients with thyroid diseases enables studies evaluating currently unresolved clinical questions. For example, studies evaluating if treatment of mild or subclinical thyroid disease improves QoL, including the question of whether certain subgroups (e.g. various strata of TSH-elevation) benefit more than others.

It could also be argued, that the role of thyroxine/thyronine combination therapy for patients not satisfactorily treated with standard thyroxine supplementation alone is not clarified yet. A pragmatic RCT, testing the treatment approach recently suggested by the European Thyroid Association (244), or a new slow-release combination tablet, could be set up similarly to the GRASS trial, including clinical and genetic biomarkers (e.g. deiodinase gene polymorphisms), preferably in a multi-cultural setting.

A study evaluating if various TSH-targets lead to differences in QoL would also provide important clinical insights into a previously debated and studied issue (245-248).

Ideally, QoL-outcome of the various treatment modalities available for Graves’ disease (including titration vs. block-replacement (249)) should also be evaluated in RCTs in analogy to a previous unique Swedish study (250)

**IMPLEMENTATION IN ROUTINE CLINICAL PRACTICE**

As described in the introduction, PROs are being implemented in routine clinical practice across the world (14, 251-255), in individual clinics, but also across entire health care systems, e.g. in UK, Sweden, The Netherlands, Australia, Canada. A centrally driven implementation has not been adopted in Denmark, but applications are emerging (256-258) and a cross-national expert board has been set (http://vibis.dk/ekspertgruppe), with
the purpose of identifying best practice in the field. Two perspectives and purposes are in play in this process: use of PROs in the direct interaction with individual patients and use of PROs at aggregate levels, for quality improvement and health system surveillance purposes. Here, focus is on the use in individual patients.

When reading and communicating with health care executives and researcher advocating the routine use of PROs in clinical practice, its implementation is anticipated with enthusiasm and expected to improve several aspects of patient care (259, 260), by:

- Identifying treatable problems
- Highlighting previously unrecognized health problems
- Assessing the effectiveness of different treatments
- Detecting adverse effects
- Monitoring disease progress, including changes not evident via clinical testing
- Improving communication
- Promoting shared decision making and patient empowerment

which may lead to better patient outcomes in the form of reduced symptoms, better QoL and enhanced satisfaction (261).

According to systematic reviews, RCTs evaluating effect of implementation in clinical practice have not unequivocally supported the abovementioned expectations (261-270). Generally, they raise methodological concerns and have difficulty aggregating the results. However, they generally do find strong evidence that the well-implemented PROs improve patient-provider communication and patient satisfaction (263). They also find growing evidence of improved monitoring of treatment response and detection of unrecognized problems. However, only weak evidence exists regarding change of patient management and health outcomes (262). Improvement in health outcomes were mostly on symptoms, side effects, toxicity, less on emotional well-being and least on social well-being and overall QoL. Concerning the latter though, a very recent and thus not yet reviewed, large, well-conducted trial evaluating use of symptom self-reporting found strong effect on overall QoL. According to the reviews, patients, as well as health professionals, generally value the PROs (259) and finds them useful. For example, a review by Chen (263) and colleagues found, that 21/23 studies evaluating patient-provider communication reported a positive effect, 11/11 studies reported strong or modest positive effect on treatment response monitoring, 15/16 studies found strong or moderate positive impact on detecting unrecognized problems, 13/17 reported strong or modest effect on changes to patient management,
very strong to moderate positive effect on patient satisfaction, 13/15 reported a modest to strong improvement in health outcomes.

The ambiguity of the conclusions drawn by the systematic reviews may in part a consequence of the methodology of the reviewed RCTs. For example, due to inappropriate measurement of the potential effect (271) or bias (272). Or it may in part be due to challenges the reviews themselves face; the RCT are rather complex and it may thus be difficult to draw overall conclusions using traditional review methodology (273).

Nonetheless, important issues facilitating successful implementation are currently being identified (251, 273-278). Among these are:

1. A sound theoretical framework concerning an understanding of how the implementations are expected to have an effect, thereby guiding RCT study design (273).
2. Clearly stated purposes of the feedback.
3. Application of outcomes which are relevant to patients and clinicians. This may lead to selection of more specific outcomes (276), conflicting potentially with an interest also in using the PROs at aggregate levels (253).
4. Brief, clear, psychometrically sound outcomes.
5. Education of staff to administrate, understand, evaluate and use PROs.
6. Interpretability. PRO results should read easy and fast and should be easily understood. For example by providing reference scores, alert thresholds (similar to marks on lab tests), highlighting of relevant changes, etc. In case results are to be fed back to patients, they should also be interpretable to lay persons. Of note, the best way to convey PRO information to professionals and patients has yet to be determined (279-281).
7. Technology and technical support in place. Ideally, a multi-modal PRO-system, fully integrated into the electronic health record, should be used.
8. Actionability. Clinicians will be less inclined to raise issues, if they cannot do anything about the problem (259, 273, 277), which may explain why PRO implementation increases discussion of symptom more often than of function, or other more distal aspects of QoL (282). Action guidance may be developed using a systematic approach, involving a multi-phase, multidisciplinary process (275).

Another reason why some studies fail to show an effect could be that good doctors already address these issues, and do so in the optimal way: by
skillful, empathic, clinical conversation with the patient. No PRO can replace this. However, often clinical daily life is very busy, and this ideal interaction becomes challenged. Routine use of PROs may be regarded as a mean to alleviate this. For example, patients doing well on current treatment, according to automated PRO surveillance, may visit a clinic more seldom, making room for patients whose PRO responses are indicative of need of additional attention. On the other hand, if PRO implementation in clinical practice fail to be relevant and meaningful, it may continue to be just one more bureaucratic burden and may end up doing more harm than good (283).

Relevance for thyroid diseases and ThyPRO
No studies on routine clinical use of PRO assessments among patients with thyroid diseases have been published.

Several features of thyroid diseases could make routine PRO assessments particularly well suited for these patients: 1) Thyroid hormones affect many organ systems and have profound effect on the central nervous system and thereby mental health. 2) QoL-issues have major influence on choice of intervention (e.g. physical symptoms and cosmetic concern in non-toxic goiter). 3) Some patients experience reduced QoL despite adequate thyroid function on treatment and may experience lack of congruence between the focus of the endocrinologist and themselves (30). 4) The diseases are often chronic and occur in all ages, including during working life. 5) Thyroid diseases are rarely life-threatening, and thus focus on and relevant addressment of QoL-issues may have relatively large impact. In contrast, most research on routine use of PRO have been among patients with cancer, who face a life-threatening disease, serious adverse effects, potential disease progression and treatment failure.

From a scientific point of view, thyroid diseases may also be an advantageous group for studies of the effect of PRO-implementation in clinical practice. For example, independent markers of management (e.g. TSH) exist and can be used as independent effect markers. Also, the existence of different, yet linked, thyroid diseases allows for comparison of effects among the different diagnostic groups.

As regards ThyPRO, several features position it as a good candidate for a PRO-implementation trial. First, it is relevant to patients and clinicians (34), easy to understand and complete (35), framed within a clinical and theoretical understanding (80, 284) and with good measurement properties, including responsiveness to treatment. Second, it encompasses both
proximal (symptoms) and distal (participation, overall QoL) aspects, ena-
bling analyses of differing effects and uses of such. Third, a standard ad-
ministration system, integrated within the electronic health record will
become available (ThyPRO·39 has been integrated into the forth-coming
regional electronic health record system “Sundshedsplatformen”) as well
as a multi-modal, flexible administration system (PROgmatic) are availa-
ble for its administration. Fourth, despite being disease-specific, general
population norms are available for interpretation-optimization. Fifth, a
network of research sites across Denmark are already applying the in-
strument for research purposes, facilitating e.g. cluster randomization
among clinics already familiar with the instrument.

OTHER FUTURE DIRECTIONS

PROs have become a natural, integral part of clinical research. Although
the optimal use in routine clinical practice still has to be identified, cli-
nical implementation for descriptive purposes seems well justified at pre-
sent. Thereby data regarding usual course of disease and treatment can
be described and possibly predictors for reduced QoL after treatment can
be identified. Such data could improve other patient-engaging activities,
such as shared decision-making, for which a tool has recently been de-
veloped (285). More detailed information about impact on QoL can improve
such tools.

Further patient-engagement, also in design of clinical trials, is recom-
mended.

In future studies, ThyPRO scales measuring generic dimensions may be
linked to standard generic instruments, e.g. SF-36, to improve compara-
bility and interpretability. For example, a study could be set up, where the
two Component Summaries (Physical and Mental) could be predicted from
the ThyPRO scales and agreement between SF-36 derived and ThyPRO
derived scales compared. MID’s could also improve interpretability and
inform power/sample size calculations for studies applying the instru-
ment.

The ThyPRO was developed for benign thyroid diseases. Future projects
could evaluate validity and usefulness of this instrument for use in thy-
roid cancer, e.g. by conducting qualitative interviews identifying relevant
issues and then evaluate coverage of these by ThyPRO.
Agreement between a paper- and an electronic version of the ThyPRO, to increase the versatility and flexibility of the instrument, has very recently been established (286).
CONCLUSION

It was possible to develop an international, valid PRO measurement system with high levels of psychometric, cross-cultural and clinical validity and reliability. An integrative approach applying both classical and modern psychometric and clinimetric as well as qualitative methods proved useful. The fact that ThyPRO is being increasingly used world-wide, indicates, that it has addressed an unmet clinical and scientific need.

ACKNOWLEDGEMENTS

I am deeply indebted to the six “ThyQoL seniors” Ulla Feldt-Rasmussen, Åse Krogh Rasmussen, Laszlo Hegedüs, Steen Bonnema, Mogens Grønvold and Jakob Bjørner. Warm thanks to Ulla, Åse and Mogens for comments and guidance on previous versions of this thesis. Thank also to John Ware, for well-coming me as a visiting scholar at University of Massachusetts and to Region Hovedstaden for financial support thereof. A warm thank also to the “ThyQoL juniors” Sofie Rasmussen, Kim Æbelø, Emilie Birch, Thea Christophersen, Victor Boesen, Stine Nissen, Mette Andersen Nexo, Kristian Winther and Per Cramon. Special thanks to Per, who’s unselfish, engaged, hard work and good spirit have availed many of us.

This work has been supported by Danish Agency for Science, Technology and Innovation: Council for Strategic Research and Council for Independent Research and by unrestricted research grants from Genzyme Corporation and Agnes and Knut Mørk’s Foundation.
REFERENCES


30. Nexo MA, Watt T, Cleal B, Hagedus L, Bonnema SJ, Rasmussen AK, Feldt-


60. **Fayers PM, Machin D**: Quality of Life: Assessment, Analysis and Interpretation. Chichester: John Wiley and Sons; 2000.


65. **Bentler PM**: Comparative fit indexes in structural models. *Psychological Bulletin* 1990, **107**:238-246.


119. Gelin MN, Zumbo BD: Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. Educational and Psychological Measurement 2003, 63(1):65-74.

120. Yang FM, Jones RN: Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. J Clin Epidemiol 2007, 60(11):1195-1200.


128. Billewicz WZ, Chapman RS, Crooks J, Day ME, Gossage J, Wayne E, Young JA:


74


163. Stott DJ, McLellan AR, Finlayson J, Chu P, Alexander WD: Elderly patients with suppressed serum TSH but normal free thyroid hormone levels usually have mild thyroid overactivity and are at increased risk of developing overt hyperthyroidism. *Quarterly Journal of Medicine* 1991, 78(285):77-84.


188. Stuckey BG, Kent GN, Allen JR: The biochemical and clinical course of


201. Slawik M, Klawitter B, Meiser E, Schories M, Zwermann O, Borm K, Peper M, Lubrich B, Hug MJ, Nauck M, Olszewski M, Beuschlein F, Reincke M: Thyroid hormone replacement for central hypothyroidism: a randomized controlled trial comparing two doses of thyroxine (T4) with a combination of T4 and...


228. Son BJ, Lee SY, Yoon JS: Evaluation of thyroid eye disease: quality-of-life questionnaire (TED-QOL) in Korean


231. **Canaris GJ, Tape TG, Wigton RS**: Thyroid disease awareness is associated with high rates of identifying subjects with previously undiagnosed thyroid dysfunction. *BMC Public Health* 2013, 13:351.


243. **McMillan C, Bradley C, Razvi S, Weaver J**: Psychometric evaluation of a new questionnaire measuring treatment


259. Boyce MB, Browne JP, Greenhalgh J: The experiences of professionals with using information from patient-reported outcome measures to improve the quality


263. Chen J, Ou L, Hollis SJ: A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013, 13:211.


274. Boyce MB, Browne JP: The effectiveness of providing peer benchmarked feedback to hip replacement surgeons based on patient-reported outcome measures: results from the PROFILE (Patient-Reported


Paper #1
CLINICAL STUDY

Validity and reliability of the novel thyroid-specific quality of life questionnaire, ThyPRO

Torquil Watt1,2, Laszlo Hegedüs3, Mogens Groenvold2,4, Jakob Bue Bjorner2,5, Åse Krogh Rasmussen1, Steen Joop Bonnema1 and Ulla Feldt-Rasmussen1

1Department of Endocrinology, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark, 2Institute of Public Health, University of Copenhagen, DK-1014 Copenhagen, Denmark, 3Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark, 4Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark and 5National Research Centre for the Working Environment, Copenhagen, Denmark

(Correspondence should be addressed to T Watt at Department of Endocrinology, Copenhagen University Hospital Rigshospitalet; Email: t.watt@rh.dk)

Abstract

Background: Appropriate scale validity and internal consistency reliability have recently been documented for the new thyroid-specific quality of life (QoL) patient-reported outcome (PRO) measure for benign thyroid disorders, the ThyPRO. However, before clinical use, clinical validity and test–retest reliability should be evaluated.

Aim: To investigate clinical (‘known-groups’) validity and test–retest reliability of the Danish version of the ThyPRO.

Methods: For each of the 13 ThyPRO scales, we defined groups expected to have high versus low scores (‘known-groups’). The clinical validity (known-groups validity) was evaluated by whether the ThyPRO scales could detect expected differences in a cross-sectional study of 907 thyroid patients. Test–retest reliability was evaluated by intra-class correlations of two responses to the ThyPRO 2 weeks apart in a subsample of 87 stable patients.

Results: On all 13 ThyPRO scales, we found substantial and significant differences between the groups expected to have high versus low scores. Test–retest reliability was evaluated by intra-class correlations of two responses to the ThyPRO 2 weeks apart in a subsample of 87 stable patients.

Conclusion: We found support for the clinical validity of the new thyroid-specific QoL questionnaire, ThyPRO, and evidence of good test–retest reliability. The questionnaire is now ready for use in clinical studies of patients with thyroid diseases.

European Journal of Endocrinology 162 161–167

Introduction

Measurements applying standardized self-reports to capture the impact of health on patients’ lives are termed health-related quality of life (HRQL) measurements (1). They usually conceptualize HRQL as a multidimensional concept encompassing various aspects of physical, mental, and social functioning and well-being. To an increasing extent, the broader, but also more neutral term ‘patient-reported outcomes (PROs)’ is replacing HRQL. Today, PROs or HRQL measurements are recognized as inevitable and important outcomes in high quality clinical studies. Further, they can provide important documentation for evidence-based patient information and may even be implemented in clinical management of the individual patient, as has been done within, e.g. oncology (2), where randomized trials have documented significant improvement of patient–clinician interaction, without prolonging consultations, and impact on patient management (3). HRQL measurements may be either generic, i.e. applicable to any patient group regardless of diagnosis, or specific, i.e. targeted to a specific disease group. Specific HRQL measurements are usually more sensitive than generic, which on the other hand have the advantage of allowing comparisons across dissimilar populations.

Some questionnaires have been developed for specific thyroid diseases (4–12). However, a thoroughly validated questionnaire only exists for thyroid-associated ophthalmopathy (TAO) patients (5–7). Another TAO questionnaire has been developed, but has not been validated (8). One questionnaire for patients with hyperthyroidism was developed, but has never been validated (4). Three questionnaires for hypothyroid patients have been developed (9–12), but studies
evaluating the validity of these measures are still awaited. Most importantly, no validated, thyroid-specific PRO instrument is available for use across different thyroid diseases (13). This is a major deficiency, because benign thyroid diseases are characterized by a substantial overlap between various disease entities (e.g. coexistence of goitre and hyperthyroidism) and a shift between diseases (e.g. hyperthyroid patients becoming hypothyroid through ablative therapy). Therefore, optimally an HRQL outcome measure for thyroid patients should encompass all thyroid diseases in order to have content validity (i.e. capture HRQL issues of relevance to the patients). If not, the results of longitudinal studies may be misleading, because important HRQL aspects are not measured at follow-up (e.g. impact of hypothyroidism after ablative therapy).

We have recently developed a quality of life (QoL) questionnaire for patients with benign thyroid diseases, called the ThyPRO (13–15), and evaluated important aspects of its measurement properties (16). In HRQL terminology, the measurement property termed accuracy within biochemical assay methodology (i.e. degree of systematic bias) is called validity and what endocrinologists may refer to as precision (or reproducibility) is termed reliability. In the just mentioned validation study (16), support for appropriate validity and reliability was found, in terms of a valid scale structure (an important aspect of what is termed construct validity) and very good internal consistency reliability. However, for HRQL measurements, evaluation of accuracy, or validity, is a more complex task than with most other measurement fields, since no gold standard against which other measures can be tested is available. Thus, the validation of an HRQL measure is an iterative process where evidence for or against the validity of a measure is usually gathered by several studies taking different approaches to evaluating validity, one of which is the above mentioned scale validation. Another important way of assessing validity, especially for measures attempted for clinical use, is termed known-groups validity. In this approach, clinically based criteria are used to classify patients into groups with expected high or low scores on a questionnaire and then test whether these expected differences are found in patient samples. Precision, or reliability, can also be evaluated by several techniques. One approach is ‘internal consistency reliability’ (Cronbach’s alpha), which has been used in the initial analyses of the ThyPRO (16). Another approach is test–retest reproducibility where duplicate measurements are obtained by collecting two responses from stable respondents separated by 2–3 weeks (17).

The purpose of the present study was to investigate accuracy, in terms of clinical/known-groups validity, and precision, in terms of test–retest reliability, of the Danish version of the thyroid-specific QoL questionnaire, the ThyPRO.

### Material and methods

#### Patients and clinical characterization

Patients were recruited from the endocrinological outpatient clinics at two university hospitals in Denmark: Copenhagen University Hospital Rigshospitalet (RH) and Odense University Hospital (OUH). At RH, the sampling strategy was cross-sectional: all thyroid patients born within the first 20 days of each month (to limit running sample size) were invited during February–June 2007 by mail 3 weeks prior to their appointment in the clinic. At OUH, all eligible patients referred to the thyroid unit of the endocrine outpatient clinic during May–November 2007 were recruited. The questionnaire was sent about 3 weeks prior to the appointment in the clinic. Blood samples were drawn the week prior to their appointment, and the participants were instructed to complete the questionnaire at about that time. Questionnaires were either returned by mail or delivered by hand in the laboratory or at the clinic on the day of appointment. Exclusion criteria were absence of any thyroid disorder, thyroid cancer, age <15, and inability to complete a questionnaire due to communication problems (non-Danish speaking, blindness, etc). One reminder was sent after 2 weeks to non-responders, and all participants gave signed informed consent.

Socio-demographic data and information about co-morbidity and non-thyroid medication were self-reported. Laboratory data, diagnostic imaging results, exact diagnosis, previous and current treatment and time of diagnosis among respondents were obtained by chart review. Biochemical thyroid tests were TSH, total thyroxine (T4), total triiodothyronine (T3), free T4 (fT4), free T3 (fT3), resin-T3 test (only OUH), thyroid peroxidase antibodies (TPOAb), thyroglobulin antibodies (only in patients diagnosed with hypothyroidism and negative TPOAb) and TSH receptor antibodies (only in patients with a diagnosis of hyperthyroidism). All analyses were performed using the standard methods at the laboratories of the participating hospitals. In patients with thyroid eye disease TAO, NOSPECS (18) classification was performed by an ophthalmologist and clinical activity scoring (19) was performed by a physician. Thyroid volume was determined by ultrasound, using the ellipsoid method (20). All examiners were blinded to the ThyPRO results. The patients were classified according to primary diagnosis, i.e. their initial diagnosis, prior to treatment. For example, patients with a non-toxic goitre, who had their thyroid removed and thus have hypothyroidism and receive thyroid hormone replacement, were classified with a diagnosis of non-toxic goitre. Self-completed data were entered using optical scanning. All clinical data were entered via SPSS Data Entry Builder 4.0 (SPSS Inc, Chicago, IL, USA) by medical staff. Data were converted into
SAS datasets, and all analyses were performed with SAS 9.1 (SAS Institute, Cary, NC, USA). The project was approved by the local ethical committee (KF01 2006-1579) and the Danish data protection agency and registered at ClinicalTrials.gov (NCT00150033).

Outcome

The ThyPRO questionnaire is self-administered and measures QoL with 13 scales (see Table 1), covering physical and mental symptoms, well-being and function as well as impact of thyroid disease on participation (i.e. social and daily life) and overall QoL. It consists of 84 items and, on average, takes 14 min to complete. Each scale ranges 0–100 with increasing scores indicating decreasing QoL (i.e. more symptoms or greater impact of disease) (16).

Known groups

Groups with expected high and low scores were defined *a priori* by a panel of four thyroid experts. The criteria used for this classification is outlined in Table 1. They include clinical data, physician ratings of overall clinical condition, and depression and anxiety status according to Hospital Anxiety and Depression Scale (HADS). The overall physician ratings were obtained at the outpatient visit, where the consulting physician, who was blinded to the ThyPRO results, rated the overall clinical condition of the patient on a five-point scale ranging from ‘very bad’ to ‘excellent’, based on all available clinical information, including patient history. Depression or anxiety according to HADS was scored using its standard thresholds: a score <8 on the depression scale indicates absence of depression, and a score above 10 indicates depression, and likewise for the anxiety scale. In addition, we evaluated the correlation between the ThyPRO depressivity and anxiety scales and the HADS depression and anxiety scales respectively using both parametric (Pearson) and non-parametric (Spearman) correlation.

Test–retest data

Since fewer patients were required for test–retest analyses than for the other validation analyses, only patients enrolled at RH during a limited period (mid-April to mid-June 2007) were included in the test–retest study. The only difference compared to the main study was the fact that they were asked to complete another questionnaire about 2 weeks after their first response. In addition, they rated any change occurring since the initial response on a seven-point scale (‘compared to when you answered the first time, would you say your overall state is much worse/somewhat worse/a little worse/more or less the same/a little better/somewhat better’.

Table 1 Description of the expected high versus low score groups used in the known groups comparisons.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Expected high score group</th>
<th>Expected low score group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goitre symptoms</td>
<td>Patients with untreated non-toxic diffuse or multinodular goitre (n=105)</td>
<td>Patients with non-goitrous autoimmune hypothyroidism treated with L-thyroxine for at least 3 months (n=107)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Untreated non-toxic diffuse or nodular goitre (n=161)</td>
</tr>
<tr>
<td>Hyperthyroid symptoms</td>
<td>Patients with hyperthyroid Graves’ disease or nodular goitre and overt hyperthyroidism (n=70)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=88)</td>
</tr>
<tr>
<td></td>
<td>within the last 3 months (n=20)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=87)</td>
</tr>
<tr>
<td>Hypothyroid symptoms</td>
<td>Patients diagnosed with overt hypothyroidism (i.e. worse than ‘Only signs’) (n=16)</td>
<td>Patients with a HADS anxiety score indicating no anxiety (i.e. score &lt;8) (n=577)</td>
</tr>
<tr>
<td></td>
<td>rated by physician as ‘very bad’ (n=12)</td>
<td>Patients with a HADS depression score indicating no depression (i.e. score &lt;8) (n=703)</td>
</tr>
<tr>
<td>Eye symptoms</td>
<td>Patients with TAO and NOSPECS &gt;1 (i.e. worse than ‘Only signs’) (n=16)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=87)</td>
</tr>
<tr>
<td></td>
<td>Tiredness Patients with overall clinical condition rated by physician as ‘very bad’ (n=12)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=84)</td>
</tr>
<tr>
<td>Cognitive impairment</td>
<td></td>
<td>Patients with overall clinical condition rated by physicians as ‘excellent’ (n=85)</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=85)</td>
</tr>
<tr>
<td>Depressivity</td>
<td>Patients with a HADS depression score indicating depression (i.e. score &gt;10) (n=64)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=87)</td>
</tr>
<tr>
<td>Emotional susceptibility</td>
<td></td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=87)</td>
</tr>
<tr>
<td>Impaired social life</td>
<td>Patients with overall clinical condition rated by physician as ‘very bad’ (n=12)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=84)</td>
</tr>
<tr>
<td>Impaired daily life</td>
<td>Patients with overall clinical condition rated by physician as ‘very bad’ (n=12)</td>
<td>Patients with overall clinical condition rated by physicians as ‘excellent’ (n=85)</td>
</tr>
<tr>
<td>Impaired sex life</td>
<td>Patients with overall clinical condition rated by physician as ‘very bad’ (n=9)</td>
<td>Patients with overall clinical condition rated by physician as ‘excellent’ (n=85)</td>
</tr>
<tr>
<td>Cosmetic complaints</td>
<td>Patients with TAO and NOSPECS above 1 and patients with non-toxic diffuse or multinodular goitre with a volume greater than 150 ml (n=21)</td>
<td>Patients with autoimmune hypothyroidism and an overall clinical condition rated by physician as ‘excellent’ (n=22)</td>
</tr>
</tbody>
</table>

TAO, thyroid-associated ophthalmopathy; NOSPECS, a classification system for defining severity of TAO; HADS, Hospital Anxiety and Depression Scale.
better/much better'). Patients who completed the second questionnaire between 10 and 24 days after the first, and who rated themselves as stable were included in test–retest analyses.

**Statistical analyses**

Differences in mean scale scores between the expected high versus low level groups were analysed with Student’s unpaired t-test, with Satterthwaites correction in case of unequal variances according to the folded F test, using SAS PROC TTEST.

Test–retest reliability was evaluated by intra-class correlations between the two measurements (21–24). Correlations were calculated using SAS PROC GLM, and 95% confidence intervals were estimated by empirical bootstrap (25, SV Thorsen and JB Bjorner, unpublished observations).

**Results**

In total, 907 responses from 1316 eligible patients were obtained, yielding an overall response rate of 69%, as detailed elsewhere (16). Of these, 195 were included in the test–retest study, 149 (76%) of whom returned a second response. Eighty-seven of the 195 fulfilled the criteria for inclusion in the analyses (Fig. 1).

Clinical characteristics for the total sample as well as for the test–retest subsample are given in Table 2.

As seen in Fig. 2, the expected high level groups had substantially higher mean scores compared with the expected low level groups on all ThyPRO scales.

All differences were statistically significant using unpaired t-tests (P<0.001 for all scales except hypothyroid symptoms and cosmetic complaints, where P<0.05). The correlations between ThyPRO and HADS scales were 0.72 for depressivity/depression and 0.77 for anxiety (P<0.0001 for both), both with parametric and non-parametric methods.

In the test–retest analyses, all intra-class correlations were above 0.70 (Table 3) and all but two (anxiety (0.77) and hypothyroid symptoms (0.80)) scales were above the ‘almost perfect’ concordance (i.e. >0.81) (26).

**Discussion**

The purpose of the present study was to investigate clinical known-groups validity and test–retest reliability of the Danish version of the thyroid-specific QoL questionnaire, the ThyPRO.
When comparing specified subgroups of patients expected to have high scale scores on each of the 13 QoL scales to subgroups expected to have low scores, we found significant differences on all scales, supporting the clinical validity of the ThyPRO. This is an important and encouraging finding.

The magnitudes of these differences varied: differences on the symptom scales, the cosmetic complaints scale and the impaired social life scale were smaller than the rest. The simple conclusion that these scales are just less sensitive (i.e. are less efficient in detecting differences among groups) can be elaborated by a closer look at the criteria used to define these groups. In fact, the analyses of these scales differed from the rest, in that the criteria defining the high versus low score groups were entirely clinical and biochemical, except for the impaired daily life scale. Criteria for the other scales were based on the consulting clinician’s rating of the overall clinical condition or HADS anxiety and depression scores. To better understand these differences, it is useful to consider a theoretical model for health outcomes. Based on the WHO ICF framework (27) and on the conceptual model for health outcomes proposed by Wilson & Cleary (28), we propose to distinguish four levels of health outcomes: i) biological and physiological variables, ii) symptoms and signs, iii) functioning, i.e. what you can physically do, and iv) participation, i.e. how you function in your social environment. Regarding the symptom scales, the criteria used to define the groups for comparisons relate to biological and physiological variables, whereas the scale scores reflect patients’ experiences of symptoms and signs. The same is true for the cosmetic complaints scale. Thus, since the disease group comparisons do not directly concern the same level of health outcomes as the ThyPRO scales, it is reasonable that the associations are fairly small. On the other hand, the overall clinical rating used for evaluating the other scales, except anxiety and depressivity, is likely to be based on symptoms and signs, and functioning, as well as diagnosis and physiological measures. Thus, strong associations for these scales are likely, which is also what we found, except for the impaired social life scale, which reflects participation rather than functioning. Finally, we are likely to see strong associations between ThyPRO scales for anxiety and depression, and other scales assessing the same domains, i.e. HADS.

For all scales, test–retest reliability was above standard thresholds for adequate reliability, indicating that the measures have appropriate reliability for use in clinical studies (29). Since the test–retest reliabilities found in this study were a little lower than the

<table>
<thead>
<tr>
<th>Scale</th>
<th>Intra-class correlation coefficient (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goitre symptoms</td>
<td>0.87 (0.81–0.91)</td>
</tr>
<tr>
<td>Hyperthyroid symptoms</td>
<td>0.89 (0.82–0.93)</td>
</tr>
<tr>
<td>Hypothyroid symptoms</td>
<td>0.80 (0.71–0.87)</td>
</tr>
<tr>
<td>Eye symptoms</td>
<td>0.86 (0.77–0.92)</td>
</tr>
<tr>
<td>Tiredness</td>
<td>0.85 (0.77–0.91)</td>
</tr>
<tr>
<td>Cognitive complaints</td>
<td>0.88 (0.79–0.93)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.77 (0.64–0.88)</td>
</tr>
<tr>
<td>Depressivity</td>
<td>0.88 (0.82–0.93)</td>
</tr>
<tr>
<td>Emotional susceptibility</td>
<td>0.87 (0.80–0.91)</td>
</tr>
<tr>
<td>Impaired social life</td>
<td>0.84 (0.72–0.91)</td>
</tr>
<tr>
<td>Impaired daily life</td>
<td>0.83 (0.71–0.91)</td>
</tr>
<tr>
<td>Impaired sex life</td>
<td>0.86 (0.74–0.94)</td>
</tr>
<tr>
<td>Cosmetic complaints</td>
<td>0.85 (0.79–0.91)</td>
</tr>
</tbody>
</table>

Figure 2 Mean scale scores for the expected high score groups versus the low score groups for the 13 ThyPRO scales. Vertical lines indicate 95% confidence intervals of the means.
The test–retest reliabilities may even underestimate true reliability slightly, possibly due to less than perfect clinically stable conditions of the patients in the test interval. In fact, the mean score was indeed a little lower at time of re-evaluation, for all scales. No scale achieved a reliability coefficient above the traditional cut-point of 0.90 required for scales used with individual patients.

Since this is the first QoL measure for use in thyroid patients covering the majority of benign thyroid disorders, no previous studies evaluating measurement properties of such measures exist. Regarding individual diagnoses, test–retest reliability and clinical known groups validity have been evaluated (16) for a disease-specific questionnaire for patients with TAO. Test–retest reliability estimates were comparable to the ones found in the present study. However, their clinical validity analyses did not entirely support the validity of the questionnaire, since some of the expected associations with clinical variables were not found (6). The interpretations of these lacking associations offered by the authors are analogous to our interpretations regarding the criteria for evaluations of the symptom scales. They suggested that it was due to the fact that the clinical variables defining the groups related to health outcomes (i.e. biological and physiological variables) differed from the health outcomes relating to QoL (i.e. symptoms, functioning and participation). For another TAO questionnaire, no associations with clinical variables were found at all (8).

To our knowledge, no other study has evaluated clinical validity or test–retest reliability of any of the other existing thyroid-specific (i.e. for individual diagnoses) questionnaires (reviewed in (13)).

Among the strengths of our study is a large sample size, enabling a detailed analysis of each group of benign thyroid disorders with adequate power. Moreover, the patients are clinically well characterized, especially compared to other studies of generic or specific QoL in thyroid patients (13). The fact that each patient was rated by a physician at about the same time that they completed the questionnaire is an important strength of the clinical validity analyses. This is also true for the availability of a ‘stability-measure’ for identification of patients to enter the test–retest substudy. Another strength is the fact that we used only few exclusion criteria, ensuring validity of the questionnaire also for subpopulations with e.g. psychiatric co-morbidities, etc. Of course, future descriptive studies evaluating HRQL in thyroid patients may consider excluding these patient groups in order to limit the description to only impairments related to thyroid diseases.

One limitation of the study relates to its design: due to our cross-sectional design, our analyses of clinical validity are limited to cover differences between groups. Ongoing studies are evaluating the ability of the questionnaire to detect relevant changes in patients over time (responsiveness) and the minimal important differences in scores. However, compared to other disease-specific instruments, including those covering other disease areas than thyroid diseases, the ThyPRO is already very well validated. Another limitation relates to the clinician evaluation. Although it is of value that these are available for each individual patient at time of survey completion, the fact that they are also based on patient history implies that they are not entirely external criteria, as are for example thyroid function tests, etc. Still, they do represent evaluations external to the ThyPRO, and clinician evaluations are often used in studies of criterion validity (30, 31).

Although the patients are clinically well characterized, and a large number of clinical descriptors are available at time of completion, the timing is not perfect; some patients may have had blood samples drawn up to about 3 weeks away from their questionnaire response. However, this would weaken the associations between e.g. thyroid function tests and HRQL data, but we still find positive associations despite such hypothetical bias.

Information from clinical known-groups studies, responsiveness studies and item response theory analyses could be utilized in an attempt to develop shorter versions of the ThyPRO by identifying the best performing items to be included in such abbreviated outcome measures. Ongoing studies are ensuring cross-cultural validity of this measure in seven languages, including English, and versions in further languages are being planned.

We recommend the use of the ThyPRO measure in studies evaluating important clinical questions regarding therapy of thyroid patients, such as whether patients with mild thyroid disease benefit from treatment (32) and whether block replacement therapy with antithyroid drugs and levothyroxine is associated with a better QoL in patients with hyperthyroidism than monotherapy with antithyroid drugs (33).

In conclusion, all scales of the ThyPRO detected clinically relevant differences among thyroid patients, and the ThyPRO was found to have adequate test–retest reliability.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This study has been supported by grants from the Danish Medical Research Council, Agnes and Knut Mørk’s Foundation, Aase and Ejnar Danielsen’s Foundation, Else and Mogens Wedell-Wedellsborg’s Foundation, the Genzyme Corporation, the Novo Nordisk Foundation and the Danish Thyroid Foundation.

Acknowledgements

We wish to express our gratitude to the staffs and colleagues at Department of Endocrinology at Rigshospitalet as well as at Odense University Hospital.
References

The Thyroid-Related Quality of Life Measure ThyPRO Has Good Responsiveness and Ability to Detect Relevant Treatment Effects

Torquil Watt, Per Cramon, Laszlo Hegedüs, Jakob Bue Bjørner, Steen Joop Bønnema, Åse Krogh Rasmussen, Ulla Feldt-Rasmussen, and Mogens Groenvold

Department of Endocrinology (T.W., P.C., Å.K.R., U.F.-R.), Copenhagen University Hospital Rigshospitalet, DK-2100 Copenhagen, Denmark; Department of Endocrinology and Metabolism (L.H., S.J.B.), Odense University Hospital, DK-5000 Odense, Denmark; QualityMetric (J.B.B.), Lincoln, Rhode Island 02865; Department of Public Health (M.G., J.B.B.), University of Copenhagen, DK-1353 Copenhagen, Denmark; National Research Institute for the Working Environment (J.B.B.), DK-2200 Copenhagen, Denmark; and Department of Palliative Medicine (M.G.), Bispebjerg Hospital, DK-2400 Copenhagen, Denmark

Background and Purpose: Patient-reported outcomes have become important endpoints in comparative effectiveness research and in patient-centered health care. Valid patient-reported outcome measures detect and respond to clinically relevant changes. The purpose of this study was to evaluate responsiveness of the thyroid-related quality of life (QoL) instrument ThyPRO in patients undergoing relevant clinical treatments for benign thyroid diseases and to compare it with responsiveness of the generic SF-36 Health Survey.

Methods: A sample of 435 patients undergoing treatment completed the ThyPRO and SF-36 Health Survey (Version 2) at baseline and 6 months after treatment initiation. Responsiveness was evaluated in three thyroid patient groups: patients with hyperthyroidism (n = 66) and hypothyroidism (n = 84) rendered euthyroid after medical therapy, and patients with a clinically detectable nontoxic goiter treated with surgery or radioactive iodine and remaining euthyroid (n = 62). Changes in QoL were evaluated in terms of effect size and compared to the changes predicted by clinical experts. The responsiveness of equivalent scales from ThyPRO and SF-36 Health Survey were compared with the relative validity index.

Results: The ThyPRO demonstrated good responsiveness across the whole range of QoL aspects in patients with hyper- and hypothyroidism. Responsiveness to treatment of nontoxic goiter was also demonstrated for physical and mental symptoms and overall QoL, but not for impact on social life or cosmetic complaints, in contrast to clinicians’ predictions. For all comparable scales except one, the ThyPRO was more responsive to treatment than the SF-36 Health Survey.

Conclusions: The ThyPRO was responsive to treatment across the range of benign thyroid diseases. We suggest implementing this measurement instrument as a patient-reported outcome in clinical studies and in clinical management. (J Clin Endocrinol Metab 99: 3708–3717, 2014)

The impact of a chronic disease on patient’s daily life is very important. Management of disease symptomatology and impact has always been an integral part of clinical care, but the increased focus on outcomes research and on optimizing health has led to a growing interest in quantifying these aspects of diseases (1–3). Health-related quality of life (QoL) has been defined as the subjective assessment of the impact of disease and its treatment...
across the physical, psychological, social, and somatic domains of functioning and well-being (4, 5), and it is evaluated by the patients themselves (6) via standardized questionnaires, ie, via patient-reported outcome (PRO) measures.

Health-related QoL can be assessed by generic or disease-specific instruments. Generic instruments are applicable to patient populations independent of diagnosis. Therefore, comparisons across diseases can be made, the influence of comorbidity may be better captured, and generic instruments often have substantial documentation of their measurement properties (7). The most widely used generic health-related QoL instrument is the SF-36 Health Survey, measuring health-related QoL in eight subdomains (8). Disease-specific instruments are tailored to a specific patient group and are often found to distinguish better between clinically relevant groups (sensitivity) and to detect more accurately clinical change over time (responsiveness) (9–11).

The increased recognition of the importance of measuring PROs has also been evident within the thyroid field (12–23). Thyroid diseases often require long treatment and control, and many different target organ systems are affected. This makes the measurement of such outcomes particularly relevant in descriptive as well as intervention studies of thyroid diseases, and possibly also in daily clinical management (24).

Acknowledging the diffuse symptomatology as well as the frequent overlap and transition between various clinical phenotypes, we have developed a comprehensive thyroid-related PRO that measures the impact of any benign thyroid disease on health-related QoL, the ThyPRO. The development, content, and measurement properties have been described and evaluated in a number of studies (25–28). Reliability (equivalent to the “precision” in medical biochemistry) has been evaluated by internal consistency (27) as well as test-retest reliability (28); validity (equivalent to accuracy) has been evaluated by qualitative studies (24, 25) and cognitive interviewing (26), clinical known-groups comparisons (28), multitrait analyses (27), differential item functioning (29), and structural equation modeling (Watt, T., M. Groenvold, N. Deng, B. Gandek, U. Feldt-Rasmussen, Å. K. Rasmussen, L. Hегедус, S. J. Bonnema, and J. B. Bjorner, unpublished data). Cross-cultural validity of several language versions has recently been evaluated by ordinal logistic regression-based analyses of measurement invariance (Watt, T., G. Barbesino, J. B. Bjorner, S. J. Bonnema, B. Bukvic, R. Drummon, M. Groenvold, L. Hegedus, V. Kantzer, K. E. Lasch, C. Marcocci, A. Mishra, R. Netea-Maier, M. Ekker, I. Paunovic, T. Quinn, Å. K. Rasmussen, A. Russell, M. Sabaretanm, J. Smit, O. Törring, Z. Zivalevic, and U. Feldt-Rasmussen, unpublished data). Other available thyroid-specific instruments focus only on particular thyroid diagnoses (21, 30–34) and of these, only the GO-QoL for Graves’ orbitopathy has been extensively validated (35, 36).

A very important aspect of the validity of a PRO is whether the instrument can detect relevant clinical changes over time, ie, its responsiveness (11, 37). However, the ThyPRO’s responsiveness has yet to be evaluated. Because our instrument covers a broad range of thyroid diseases, responsiveness cannot be expected to be high for all scales in all clinical groups, and it is therefore important that the evaluation of responsiveness is tightly integrated into a robust clinical setting and framework. Particularly, it should be evaluated with a clear identification of which aspects of health-related QoL are expected to change as a consequence of specific clinical treatments of clearly defined patient groups.

The purpose of the present study was 2-fold: first, to evaluate the responsiveness of ThyPRO in patients undergoing relevant clinical treatment for benign thyroid diseases; and second, to compare its responsiveness with that of the widely accepted and applied generic SF-36 Health Survey.

**Patients and Methods**

**Patient population and inclusion procedure**

From 2008 to 2013, patients undergoing treatment for benign thyroid disease at two university hospital outpatient clinics completed ThyPRO and SF-36 Health Surveys (Version 2) prior to and 6 months after treatment. Inclusion criteria were: age above 18 years; ability to complete paper-and-pencil questionnaires in Danish; and referral to and prescription of clinically relevant treatment or change in treatment of the thyroid disease. At Copenhagen University Hospital Rigshospitalet, all referrals were screened consecutively, as were decision reports from all weekly cross-subspecialty board rounds, for eligible patients. The latter included most decisions regarding treatment with radioactive iodine and surgery. Furthermore, patients already followed at the outpatient clinic but prescribed a clinically relevant change in treatment were also included by their treating physicians (eg, patients followed for Graves’ disease experiencing relapse of hyperthyroidism who were referred to a second course of antithyroid drug treatment). At Odense University Hospital, eligible patients were identified through screening all patients referred with a thyroid diagnosis. Exclusion criteria at both university hospitals were: pregnancy; patients undergoing minor adjustments of treatments or referred for second opinion or diagnostic procedures; major comorbidity considered to have substantial influence on QoL; or thyroid malignancy. Eligible patients were sent a booklet containing the two PROs (ThyPRO and SF-36) and sociodemographic questions by mail. One mailed reminder was sent in case of nonresponse. Clinical data including exact diagnosis, previous and current treatment, and biochemical measurements were obtained by medical chart review.
Defining responsive patient groups

In the study protocol, three patient groups undergoing clinically relevant change were defined: 1) patients with hyperthyroidism (ie, suppressed TSH [reference range, 0.35–4.00 mU/L] and elevated nonprotein-bound T4 [reference range, 11.5–22.7 pmol/L] or T3 [reference range, 1.9–2.6 nmol/L]) (Graves’ disease or toxic nodular goiter) rendered euthyroid by treatment 6 months after treatment initiation; 2) patients with autoimmune hypothyroidism (elevated TSH and positive thyroid peroxidase antibodies) rendered euthyroid by treatment 6 months after treatment initiation; and 3) patients with a clinically noticeable goiter treated with ablative therapy (hemithyroidectomy or radioactive iodine for the purpose of volume reduction), also evaluated at follow-up 6 months after treatment.

PRO measures

The ThyPRO measures a range of aspects of QoL relevant to patients with benign thyroid diseases, as identified during patient and expert interviews. It thus covers not only physical symptoms specifically relevant to thyroid diseases, eg, symptoms of hyperthyroidism and goiter, but also nonspecific aspects of high importance to patients with thyroid diseases, eg, fatigue. It consists of 84 items summarized in 13 scales as well as a single item measuring overall impact of thyroid disease on QoL. Each item is rated on a 0–4 Likert scale, from no symptoms/problems = 0 to severe symptoms/problems = 4. The average score of items in a scale is divided by four and multiplied by 100 to yield thirteen 0–100 scales, with higher scores indicating worse health status. The SF-36 Health Survey measures general health, physical and mental well-being, and function as well as impact of health problems on role and social functioning (8). Thirty-five items are summarized into eight 0–100 scales as described above for ThyPRO, but with higher scores indicating better health. One additional single-item measures change in health during the last year.

Clinicians’ ranking of scales

Before analyses, predictions by four thyroid experts (S.J.B., L.H., Å.K.R., and U.F.-R.), each with more than two decades of clinical experience with thyroid patients and familiarity with ThyPRO, were gathered. The experts indicated for each of the 13 ThyPRO scales whether they expected the ThyPRO measures to change as a result of treatment in each of the three predefined patient groups, based on their clinical experience (yes/no/perhaps).

Statistical analyses

Responsiveness of ThyPRO was evaluated as the change in scale scores from baseline to follow-up, 6 months after treatment initiation. Statistical significance of changes in mean scale scores was evaluated by paired Student’s t test. A value of \( P < .05 \) with Hochberg’s adjustment for multiple testing (38) was considered statistically significant. Magnitude of change was evaluated by standardized effect sizes, calculated as mean difference divided by SD of scores at baseline (39, 40). In accordance with Cohen (39), effect sizes between 0.2 and 0.5 were classified as small, 0.5–0.8 as moderate, and >0.8 as large. The magnitude and significance of change were compared with the changes predicted by clinicians.

The responsiveness of ThyPRO compared to SF-36 was evaluated for scales from the two instruments with equivalent content. This comparison was made with the relative validity index, which is a standard measure applied in psychometric comparisons between different measurements of identical concepts (41, 42). It is calculated by dividing the F-statistic for each equivalent scale by the maximum F-statistic among the equivalent scales. Thus, it can attain a value of 0–1, where 1 indicates the most responsive scale. In addition to this standard approach, calculation of 95% confidence intervals (CIs) around the relative validity index was applied using bootstrapping (43). The F-statistic was obtained by squaring the t-statistics from the paired comparisons. To compare higher-order summary measures of the two instruments, responsiveness of a ThyPRO Composite score constructed for two clinical trials (44, 45) was compared with the SF-36 Health Survey Mental Component Summary (MCS). The ThyPRO Composite score was based on items from the Tiredness, Cognition, Anxiety, Depression, Emotional Susceptibility, Impaired Social Life, Impaired Daily Life, and Overall QoL scales. Acceptable scaling properties have been documented for this 26-item Composite score (46).

All analyses were performed with SAS 9.3 software (SAS Institute Inc) (47). According to Danish law, PRO research does not require and thus cannot obtain approval by ethical committees, and a completed questionnaire is regarded as consent. The study was approved by the Danish Data Protection Agency and conducted in accordance with the Declaration of Helsinki.

Results

Of the 544 patients with baseline evaluations undergoing clinically relevant treatment, 435 completed the follow-up survey, yielding a total response rate of 80%. Clinical and sociodemographic characteristics are shown in Table 1. Of the 435 respondents, a total of 212 could be classified into one of the three predefined clinically respondent groups 6 months after treatment initiation: 66 were hyperthyroid rendered euthyroid, 84 were hypothyroid rendered euthyroid, and 62 were patients with a clinically noticeable goiter treated with surgery or radioactive iodine. Thus, 223 patients could not be allocated to any of the predefined response groups, either because they had other diagnoses (eg, Graves’ orbitopathy, subacute thyroiditis de Quervain) or did not comply with the criteria at baseline (eg, had mild hyperthyroidism or did not reach the treatment goals at follow-up (eg, not euthyroid).

Change over time

As shown in Table 2, statistically significant changes were observed in all ThyPRO and SF-36 scales in the total sample. The same was observed in the hypothyroid response sample. In the hyperthyroid response sample, all ThyPRO scales except the Hypothyroid Symptoms scale measuring physical symptoms were significantly better at follow-up, as were five of the eight SF-36 scales.

In the goiter response sample, statistically significant improvements were observed in four of the ThyPRO scales, ie, Goiter Symptoms, Anxiety, Tiredness, and Emo-
tional Susceptibility. In this group, none of the SF-36 scales improved significantly.

Clinician expectations

ThyPRO scales expected by clinicians to change in each of the three response samples are marked in Table 2. As shown, seven of the ThyPRO scales were expected to improve in the hyperthyroid response sample by at least three of the four clinicians, and two additional scales were expected to change by two of the clinicians. In the hypothyroid change sample, five scales were expected to change by at least three of the clinicians, and another three scales by

<p>| Table 1. Clinical and Sociodemographic Characteristics of the Samples |</p>
<table>
<thead>
<tr>
<th>All Patients</th>
<th>Hyperthyroid Patients Treated to Become Euthyroid at Follow-Up</th>
<th>Autoimmune Hypothyroid Patients Treated to Become Euthyroid</th>
<th>Nontoxic Goiter Treated With Surgery or Radioiodine for Volume Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>435</td>
<td>66</td>
<td>84</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>361 (83)</td>
<td>54 (82)</td>
<td>76 (90)</td>
</tr>
<tr>
<td>Men</td>
<td>74 (17)</td>
<td>12 (18)</td>
<td>8 (10)</td>
</tr>
<tr>
<td>Age, y</td>
<td>54 (42–63)</td>
<td>55 (47–65)</td>
<td>45 (32–60)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No vocational training</td>
<td>68 (16)</td>
<td>8 (12)</td>
<td>11 (13)</td>
</tr>
<tr>
<td>Currently studying</td>
<td>12 (3)</td>
<td>0</td>
<td>7 (8)</td>
</tr>
<tr>
<td>Short vocational training</td>
<td>52 (13)</td>
<td>15 (23)</td>
<td>7 (8)</td>
</tr>
<tr>
<td>Apprenticeship</td>
<td>44 (10)</td>
<td>8 (12)</td>
<td>5 (6)</td>
</tr>
<tr>
<td>Short academic education (1–3 y)</td>
<td>116 (27)</td>
<td>19 (29)</td>
<td>12 (14)</td>
</tr>
<tr>
<td>Medium academic education (3–5 y)</td>
<td>90 (21)</td>
<td>13 (20)</td>
<td>23 (27)</td>
</tr>
<tr>
<td>Long academic education (5 y or more)</td>
<td>40 (9)</td>
<td>1 (2)</td>
<td>19 (23)</td>
</tr>
<tr>
<td>Information not available</td>
<td>13</td>
<td>2 (3)</td>
<td>0</td>
</tr>
<tr>
<td>Months from treatment initiation to completion of follow-up survey</td>
<td>6.5 (6.5–6.8)</td>
<td>6.5 (6.4–6.7)</td>
<td>6.5 (6.3–6.8)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nontoxic goiter</td>
<td>135 (31)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Toxic nodular goiter</td>
<td>98 (23)</td>
<td>28 (42)</td>
<td>0</td>
</tr>
<tr>
<td>Graves’ hyperthyroidism</td>
<td>73 (17)</td>
<td>27 (41)</td>
<td>0</td>
</tr>
<tr>
<td>Graves’ orbitopathy</td>
<td>25 (6)</td>
<td>9 (14)</td>
<td>0</td>
</tr>
<tr>
<td>Autoimmune hypothyroidism</td>
<td>86 (20)</td>
<td>0</td>
<td>84 (100)</td>
</tr>
<tr>
<td>Other thyroid diagnosis</td>
<td>18 (4)</td>
<td>2 (3)</td>
<td>0</td>
</tr>
<tr>
<td>Disease duration, mo</td>
<td>0.3 (0–4)</td>
<td>0.2 (0–1)</td>
<td>0 (0–0.2)</td>
</tr>
<tr>
<td>Index treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-T4</td>
<td>111 (26)</td>
<td>0</td>
<td>84 (100)</td>
</tr>
<tr>
<td>Antithyroid medication</td>
<td>86 (20)</td>
<td>36 (55)</td>
<td>0</td>
</tr>
<tr>
<td>Aspiration of thyroid cyst</td>
<td>4 (1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Glucocorticoid pulse therapy</td>
<td>2 (0)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other immunosuppressive treatment</td>
<td>4 (1)</td>
<td>2 (3)</td>
<td>0</td>
</tr>
<tr>
<td>Hemithyroidectomy</td>
<td>64 (15)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total thyroidectomy</td>
<td>37 (9)</td>
<td>5 (8)</td>
<td>0</td>
</tr>
<tr>
<td>Radioactive iodine</td>
<td>127 (29)</td>
<td>23 (35)</td>
<td>0</td>
</tr>
<tr>
<td>Thyroid function at baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSH, mIU/L (undetectably low = 0)</td>
<td>0.42 (0.01–2.64)</td>
<td>0 (0–0.02)</td>
<td>7.4 (4.59–11.4)</td>
</tr>
<tr>
<td>Thyroxine (ie, total T4)</td>
<td>99 (80–133)</td>
<td>141 (112–177)</td>
<td>82 (72–96)</td>
</tr>
<tr>
<td>Thyroid function at follow-up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSH, mIU/L</td>
<td>1.37 (0.46–2.87)</td>
<td>0.76 (0.28–1.3)</td>
<td>2.26 (1.5–4.06)</td>
</tr>
<tr>
<td>Thyroxine (ie, total T4)</td>
<td>93 (77–112)</td>
<td>89 (73–105)</td>
<td>103 (82–121)</td>
</tr>
</tbody>
</table>

Data are expressed as number (percentage) or median (interquartile range [Q1-Q3]).

a These patients had subclinical hyperthyroidism, but the purpose of treatment was volume reduction.
two. In the goiter response sample, only three of the 13 scales were expected to change by three of the clinicians, and one additional scale was expected to change by two of them.

**Relationship between observed changes and clinician expectations**

In the hyperthyroid response sample, a statistically significant change was observed on all scales expected by clinicians to change (Table 2). In the hypothyroid sample, a change was observed on all except one (Hypothyroid symptoms) of the expected scales. In the goiter response sample, significant changes were observed in two of the scales expected to improve, whereas no change was found on the other two expected to change, i.e., Impaired Social Life and Cosmetic Complaints.

**Magnitude of changes (effect sizes)**

The magnitudes of the changes, as estimated by the standardized effect sizes, are shown for each of the ThyPRO and SF-36 scales in Table 3.

In the hyperthyroid response sample, the change was large (i.e., effect size ≥0.80) for two of the ThyPRO scales (Hyperthyroid symptoms and Tiredness), moderate (i.e., 0.50 to 0.80) for five of the scales (Goiter symptoms, Anxiety, Depression, Emotional Susceptibility, Impaired Daily Life) and the Overall QoL impact item, and small (i.e., 0.20–0.50) for the remaining ThyPRO scales. None were below the 0.2-threshold for no change. On the SF-36 scales, no “large” changes were observed in any of the SF-36 scales in this hyperthyroid response sample. “Moderate” changes were observed for the following SF-36 scales: Vitality, Mental Health, Role Physical, and Physical Function. “Small” changes were found in the following SF-36 scales: Social Function, Role Emotional, General Health, and Bodily Pain.

Changes in the hypothyroid response sample were of smaller magnitude than in the hyperthyroid sample. None of the effect sizes were large. Moderate changes were observed for the ThyPRO Tiredness scale and the Overall Health, and Bodily Pain.

<p>| Table 2. ThyPRO and SF-36 Scale Scores at Baseline and at Follow-up 6 Months After Treatment |
|--------------------------------------------|----------------------------------------|------------------------------------------|----------------------------------------|------------------------------------------|</p>
<table>
<thead>
<tr>
<th>All Patients (n = 435)</th>
<th>Hyperthyroidism Treated to Become Euthyroid at Follow-Up (n = 66)</th>
<th>Autoimmune Hypothyroidism Treated to Become Euthyroid (n = 84)</th>
<th>Nontoxic Goiter Treated With Surgery or Radioiodine for Volume Reduction (n = 62)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ThyPRO scale</strong></td>
<td><strong>Baseline</strong></td>
<td><strong>At 6 mo</strong></td>
<td><strong>Change</strong></td>
</tr>
<tr>
<td>Goiter Symptoms</td>
<td>23 (22)</td>
<td>12 (16)</td>
<td>11 (19)</td>
</tr>
<tr>
<td>Hyperthyroid Symptoms</td>
<td>28 (21)</td>
<td>18 (18)</td>
<td>10 (19)</td>
</tr>
<tr>
<td>Hypothyroid Symptoms</td>
<td>24 (22)</td>
<td>20 (21)</td>
<td>4 (18)</td>
</tr>
<tr>
<td>Eye Symptoms</td>
<td>17 (19)</td>
<td>13 (16)</td>
<td>4 (15)</td>
</tr>
<tr>
<td>Tiredness</td>
<td>57 (27)</td>
<td>43 (26)</td>
<td>14 (26)</td>
</tr>
<tr>
<td>Cognition</td>
<td>25 (24)</td>
<td>20 (22)</td>
<td>5 (18)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>29 (24)</td>
<td>17 (20)</td>
<td>12 (21)</td>
</tr>
<tr>
<td>Depression</td>
<td>33 (22)</td>
<td>28 (21)</td>
<td>5 (20)</td>
</tr>
<tr>
<td>Emotional Susceptibility</td>
<td>39 (24)</td>
<td>30 (23)</td>
<td>9 (20)</td>
</tr>
<tr>
<td>Impaired Social Life</td>
<td>14 (20)</td>
<td>10 (17)</td>
<td>4 (17)</td>
</tr>
<tr>
<td>Impaired Daily Life</td>
<td>25 (27)</td>
<td>14 (21)</td>
<td>10 (23)</td>
</tr>
<tr>
<td>Impaired Sex Life</td>
<td>25 (31)</td>
<td>19 (27)</td>
<td>6 (27)</td>
</tr>
<tr>
<td>Cosmetic Complaints</td>
<td>18 (19)</td>
<td>16 (20)</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Overall QoL Impact</td>
<td>40 (34)</td>
<td>22 (28)</td>
<td>18 (32)</td>
</tr>
<tr>
<td>SF-36 scale</td>
<td>Physical Function</td>
<td>77 (24)</td>
<td>82 (23)</td>
</tr>
<tr>
<td>Role Physical</td>
<td>69 (29)</td>
<td>77 (27)</td>
<td>7 (26)</td>
</tr>
<tr>
<td>Bodily Pain</td>
<td>74 (27)</td>
<td>81 (24)</td>
<td>7 (24)</td>
</tr>
<tr>
<td>General Health</td>
<td>66 (21)</td>
<td>68 (21)</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Vitality</td>
<td>46 (25)</td>
<td>56 (25)</td>
<td>10 (23)</td>
</tr>
<tr>
<td>Social Function</td>
<td>78 (27)</td>
<td>85 (23)</td>
<td>7 (23)</td>
</tr>
<tr>
<td>Role Emotional</td>
<td>75 (27)</td>
<td>80 (25)</td>
<td>5 (26)</td>
</tr>
<tr>
<td>Mental Health</td>
<td>66 (20)</td>
<td>73 (19)</td>
<td>7 (18)</td>
</tr>
</tbody>
</table>

Mean (SD) differences in scale scores significantly different from zero at the 0.05 level, after Hochberg’s adjustment for multiple testing, within each sample, are in bold. ThyPRO scales that the four thyroid experts predicted to change are marked with superior black dots, one for each expert who expected the scale to change. Scales are scored 0–100; for ThyPRO, higher scores indicate more problems or symptoms, i.e., worse QoL; whereas higher SF-36 scores indicate better QoL. Differences are in absolute values.
QoL impact item. Effect sizes were small for all the other ThyPRO scales, except Hypothyroid Symptoms, which showed no effect (effect size below 0.20). On SF-36, all changes in the hypothyroid response sample were small, except general health and physical function, which showed no effect (effect size below 0.20).

In the goiter response sample, effect sizes for the ThyPRO scales were large for Goiter Symptoms and moderate for the Anxiety scale. Small changes were observed for five other ThyPRO scales and for the Overall QoL impact item. None of the effect sizes for the change in SF-36 scales were large or moderate. Effect sizes were indicative of a small change in Vitality, Mental Health, Social Function, and Bodily Pain.

### Responsiveness of ThyPRO compared to SF-36 (relative validity)

Relative responsiveness of similar scales could be compared in five instances: ThyPRO Tiredness vs SF-36 Vitality; ThyPRO Anxiety, Depressivity, and Emotional Susceptibility vs SF-36 Mental Health; ThyPRO Impaired social Life vs SF-36 Social Function; ThyPRO impaired Daily Life vs SF-36 role Physical and Role Emotional; and ThyPRO Overall QoL impact vs SF-36 General Health. For all comparisons except for Social Function, the ThyPRO had better responsiveness than SF-36 according to the magnitude of the relative validity index, as shown in Table 4. When examining the CIs, the interval of relative validity included the value 1 for the SF-36 Vitality in all groups, the SF-36 Role Physical in the hypothyroid and the goiter groups, and the SF-36 Role Emotional and General Health in the goiter change group. Similarly, the CI included the value 1 for the ThyPRO Impaired Social Life where the relative validity index was highest for the SF-36 Social Function, in all three groups. The ThyPRO Composite scale had better responsiveness than the SF-36 MCS scale in all three comparisons: relative validity = 1 for the ThyPRO Composite score in all three response samples, vs 0.38 (95% CI, 0.16–0.60), 0.44 (0.20–0.80), and 0.50 (0.13–1.33) for MCS in the hyperthyroid, hypothyroid, and goiter response samples, respectively.

### Discussion

We evaluated the responsiveness of the thyroid-related QoL PRO measure, ThyPRO, and compared it to the responsiveness of the generic SF-36 Health Survey. Because not all scales can be expected to change in all patient groups, our analysis of responsiveness was guided by ex-

### Table 3. Effect Size (mean change/SDbaseline) for ThyPRO and SF-36 Scales

<table>
<thead>
<tr>
<th>ThyPRO scale</th>
<th>Hyperthyroidism Treated to Become Euthyroid (n = 66)</th>
<th>Autoimmune Hypothyroidism Treated to Become Euthyroid (n = 84)</th>
<th>Nontoxic Goiter Treated With Surgery or Radioiodine for Volume Reduction (n = 62)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goiter symptoms</td>
<td>0.61</td>
<td>0.29</td>
<td>1.00</td>
</tr>
<tr>
<td>Hyperthyroid symptoms</td>
<td>0.96</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>Hypothyroid symptoms</td>
<td>0.42</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>Eye symptoms</td>
<td>0.29</td>
<td>0.28</td>
<td>0.30</td>
</tr>
<tr>
<td>Tiredness</td>
<td>0.89</td>
<td>0.63</td>
<td>0.43</td>
</tr>
<tr>
<td>Cognition</td>
<td>0.35</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.78</td>
<td>0.39</td>
<td>0.67</td>
</tr>
<tr>
<td>Depressivity</td>
<td>0.50</td>
<td>0.29</td>
<td>0.07</td>
</tr>
<tr>
<td>Emotional susceptibility</td>
<td>0.63</td>
<td>0.45</td>
<td>0.27</td>
</tr>
<tr>
<td>Impaired social Life</td>
<td>0.30</td>
<td>0.28</td>
<td>0.11</td>
</tr>
<tr>
<td>Impaired daily life</td>
<td>0.64</td>
<td>0.43</td>
<td>0.26</td>
</tr>
<tr>
<td>Impaired sex life</td>
<td>0.41</td>
<td>0.27</td>
<td>0.18</td>
</tr>
<tr>
<td>Cosmetic complaints</td>
<td>0.20</td>
<td>0.30</td>
<td>−0.02</td>
</tr>
<tr>
<td>Overall QoL impact item</td>
<td><strong>0.75</strong></td>
<td><strong>0.59</strong></td>
<td></td>
</tr>
<tr>
<td>SF-36 scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical function</td>
<td>0.55</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>Role physical</td>
<td>0.50</td>
<td>0.34</td>
<td>0.13</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>0.38</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>General health</td>
<td>0.39</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Vitality</td>
<td>0.68</td>
<td>0.49</td>
<td>0.22</td>
</tr>
<tr>
<td>Social function</td>
<td>0.44</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>Role emotional</td>
<td>0.28</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td>Mental health</td>
<td>0.66</td>
<td>0.38</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Small effect sizes (0.2–0.5) are in italic, moderate effect (0.5–0.8) in bold, and large effect sizes (>0.8) in bold italic.
In general, ThyPRO was responsive across the whole range of QoL dimensions, ie, scales covering physical and mental symptoms, well-being, and impact of disease on daily and social life, as well as overall QoL. As expected, the most pronounced changes were observed for the hyperthyroid patients rendered euthyroid when evaluated 6 months after treatment initiation. Changes were also seen across the entire spectrum of QoL for hypothyroid patients rendered euthyroid, but although statistically significant, the observed changes were smaller than for the hyperthyroid group. In patients with nontoxic goiter, changes were observed in physical and mental symptoms and in overall QoL, but not in aspects such as impact on daily or social function.

Significant changes were observed in all instances expected by experts, with three exceptions. First, physical symptoms of hypothyroidism did not improve significantly with treatment in the hypothyroid group. One possible explanation for this could be that currently, the most prominent features of hypothyroidism are not physical symptoms such as dry skin or swollen hands or feet, but rather diffuse symptoms such as fatigue and reduced mental well-being. Alternatively, it may take longer than 6 months for physical consequences of hypothyroidism to resolve after euthyroidism has been reached. Another explanation could be the relatively limited severity of hypothyroidism in this sample. Second, in the goiter group, expected improvement of social life impairment was not observed. According to the low baseline level, this is probably due to the lack of impact on social function of nontoxic goiter. Third, in the same group, Cosmetic Complaints did not improve either. This is more surprising because cosmetic concerns would be expected to be the indication for treatment in a substantial proportion of these patients. However, at baseline, the nontoxic goiter group had in fact lower mean scores on the Cosmetic Complaints scale than the two groups with thyroid dysfunction, indicating that cosmetic issues may be of little concern. This is consistent with a qualitative study comparing patients’ and clinicians’ rating of the importance of QoL issues possibly relevant for patients with nontoxic goiter (25). In that study, the clinicians rated dissatisfaction with appearance as the sixth most important issue, whereas the

Table 4. Relative Validity of ThyPRO and SF-36 Scales With Similar or Closely Related Content

<table>
<thead>
<tr>
<th>ThyPRO Scale</th>
<th>SF-36 Scale</th>
<th>F Value</th>
<th>Relative Validity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiredness</td>
<td>Vitality</td>
<td>47.8</td>
<td>1</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td>31.7</td>
<td>0.66 (0.41–1.05)</td>
</tr>
<tr>
<td>Depressivity</td>
<td></td>
<td>18.3</td>
<td>0.64 (0.27–1.32)</td>
</tr>
<tr>
<td>Emotional susceptibility</td>
<td>Mental health</td>
<td>14.1</td>
<td>0.49 (0.20–0.92)</td>
</tr>
<tr>
<td>Impaired social life</td>
<td>Social function</td>
<td>10.3</td>
<td>1</td>
</tr>
<tr>
<td>Impaired daily life</td>
<td>Role physical</td>
<td>16.4</td>
<td>0.65 (0.31–1.23)</td>
</tr>
<tr>
<td></td>
<td>Role emotional</td>
<td>3.7</td>
<td>0.15 (0.01–0.52)</td>
</tr>
<tr>
<td>Overall QoL impact item</td>
<td>General health</td>
<td>0.03</td>
<td>0 (0.00–0.13)</td>
</tr>
</tbody>
</table>

Relative validity was calculated by dividing the F-statistic for the baseline — 6 months difference of each scale being compared, with the maximal F-statistic among the compared scales. Thus, the scale with maximal change F-statistic has a relative validity of 1. The 95% CIs were calculated with bootstrapping.
patients rated it as the 39th most important. Masking visible signs of goiter was rated as the seventh most important by the clinicians, whereas the patients rated it the 112th most important issue, ie, not important at all.

The comparison of ThyPRO and SF-36 showed that ThyPRO was more responsive than SF-36. For seven of the ThyPRO scales, there was no equivalent SF-36 scale: the four physical symptom scales (Goiter, Hyperthyroid, Hypothyroid, and Eye Symptoms), and for Cognitive problems, Impaired Sex Life, and Cosmetic Complaints. These aspects of QoL were identified through patient interviews and found important by these patients. The two largest effect sizes were, indeed, found among these scales. The SF-36 scales Physical Function and Bodily Pain do not have equivalent ThyPRO scales because they were not rated important in the patient interviews (25). For all other scales except one, Social Function, the ThyPRO had the highest relative validity. In five of the 12 comparisons where ThyPRO had best responsiveness, the relative validity index of the corresponding SF-36 scales included the value of 1, indicating that a larger sample size is warranted for a firm establishment of a larger responsiveness of the ThyPRO on these scales. This was also the case in all three comparisons, where the SF-36 scale (Social Function) had the highest F-value. The high responsiveness demonstrated here is an important verification of the validity of the ThyPRO.

The results of our study are in line with most previous studies in other patient groups, where higher responsiveness for disease-specific compared to generic instruments has also been demonstrated (10, 11, 48), although there are exceptions. In some studies the generic instrument had the highest responsiveness (49–51).

Responsiveness has previously been demonstrated for a PRO developed specifically for patients with Graves’ orbitopathy (the GO-QOL questionnaire) (35), but the present study is the first to evaluate the responsiveness of a QoL instrument in the most common thyroid diseases, ie, patients with hyperthyroidism, hypothyroidism, and goiter.

It is an advantage of our study that responsiveness was evaluated in several clinical phenotypes and that analyses were tied to a clinical framework with specific clinician ratings of expected performance of the instrument. Furthermore, it is a strength that the analyses are compared directly with the performance of a well-known generic instrument, SF-36, with well-documented measurement properties. It is a limitation, however, that the sample size does not allow further subdivision of the effect of various treatments, eg, antithyroid drug treatment, thyroidectomy, and radioactive iodine for hyperthyroidism, which would have profoundly improved our current understanding of evidence-based medicine (52, 53). Such data could have provided information for interpretability, ie, what magnitude of change to expect as a consequence of a specific treatment; could better guide power calculations for future studies; and most importantly, could help prioritize the choice of therapy. A more elaborate and detailed picture of the responsiveness of ThyPRO can be established in future studies if narrower patient and treatment samples are defined and compared. Regarding the hyperthyroid group, symptoms may differ among patients with new-onset and relapsed disease. In this study, they were analyzed jointly because such a difference was not expected to influence the responsiveness as such, but they could be analyzed separately in future, larger studies. Furthermore, for ethical reasons, eligibility for this study did not withhold treatment initiation. For example, treatment with l-T4 was instituted once hypothyroidism was firmly diagnosed. For some patients, a clinical improvement may thus have occurred before the completion of the first questionnaire, which may have led to underestimation of the true responsiveness of ThyPRO. Finally, future studies may apply ThyPRO to evaluate the relationship between QoL and thyroid function within the normal range; this would, however, require additional validation studies.

The advantage of better responsiveness of ThyPRO compared to SF-36 must be weighed against its greater length (85 vs 36 items) and the lack of comparability against other patient groups. On request, norm data from the Danish general population are available for the ThyPRO, and the length can be reduced by selection of only the relevant ThyPRO scales for a particular study. However, such selection should be cautious, until a better understanding of the performance of the ThyPRO and thereby the understanding of the QoL impact of thyroid disease has been gained. Such experience will evolve with increasing application of PRO measures in clinical studies of thyroid diseases. Future studies will focus on the development of abbreviated versions of the ThyPRO, preserving the good measurement properties demonstrated in the present study.

In conclusion, we found that ThyPRO has very good responsiveness, and based on the previously conducted validation studies and the present study, we recommend its use in future clinical studies, trials, as well as in daily clinical practice. Such studies should further evaluate the relationships between clinical variables and QoL, as measured by ThyPRO. We recommend using this instrument if evaluations of changes and comparisons of differences in changes, eg, in randomized clinical trials, are important. If greater emphasis is placed on comparability of QoL with other patient groups, a generic instrument such as SF-36...
can be utilized and supplemented with the most important ThyPRO scales for a particular intervention.

Acknowledgments

Warm thanks to Sofie Larsen Rasmussen and Lene Frydenberg for help with identification and inclusion of patients; to Sofie Larsen Rasmussen and Kira Bang Bové for help with follow-up; to Sofie Larsen Rasmussen, Kim Ebele, Emilie Birch, Thea Christophersen, and Sara Kehlet Watt for help with generation of clinical data; to Selma Watt and Laura Siim Magnussen for meticulous data entry; and to Dan Frendl for comments on an earlier version.

Address all correspondence and requests for reprints to: Torquil Watt, MD, PhD, Department of Endocrinology, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark. E-mail: Torquil.Watt@regionh.dk.

This study has been supported by grants from the Danish Agency for Science, Technology and Innovation; Council for Strategic Research and Council for Independent Research and an unrestricted research grant from the Genzyme Corporation. L.H. is supported by an unrestricted grant from the Novo Nordisk Foundation.

Disclosure Summary: The authors have nothing to disclose.

References

41. Mchonney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993;31:247–263.
Confirmatory factor analysis of the thyroid-related quality of life questionnaire ThyPRO

Torquil Watt1,2*, Mogens Groenvold2,3, Nina Deng4, Barbara Gandek4, Ulla Feldt-Rasmussen1, Åse Krogh Rasmussen1, Laszlo Hegedüs5, Steen Joop Bonnema5 and Jakob Bue Bjorner2,6,7

Abstract

Background and aim: Thyroid diseases are prevalent and chronic. With treatment, quality of life is restored in most, but not all patients. Construct validity of the thyroid-related quality of life questionnaire, ThyPRO, has been established by multi-trait scaling, but not evaluated with more elaborate methods. The purpose of the present study was to evaluate dimensionality of the ThyPRO scales and to attempt to understand possible item misfit through structural equation modeling for categorical data.

Methods: The current 85-item version of ThyPRO consists of 13 scales, covering domains of physical (4 scales) and mental (2 scales) symptoms, function and well-being (3 scales) and participation/social function (4 scales). The data were collected from a cross-sectional sample of 907 thyroid patients. One-factor confirmatory models were fitted to each scale, and evaluated by model fit statistics (comparative fit index >0.95, root mean square error of approximation <0.08), magnitude of factor loadings, model residual correlations and modification indices (MI). Indications of multi-dimensionality were tested in bi-factor models. Possible item misfit was evaluated in a combined, investigational model.

Results: Each ThyPRO scale was adequately represented by a unidimensional model after minor revisions. Eleven items were identified in the unidimensional models as potentially misfitting and were investigated further by multidimensional modeling.

Conclusion: Elaborate psychometric modeling supported the construct validity of the ThyPRO. However, 11 potentially misfitting items and 18 items with local dependence to other items are candidates for removal in future item reduction processes.

Keywords: Patient-reported outcomes, Unidimensionality, Quality of life, Scale validation, Thyroid disease

Introduction

Thyroid diseases are diseases related to the thyroid gland, which is an endocrine, i.e. hormone producing, gland located in the front of the neck. Thyroid diseases are prevalent, affecting approximately 15% of individuals of all ages, with a 4 to 1 women/men ratio [1,2]. The main disease groups comprise non-toxic goiter (enlargement of the gland), hyperthyroidism (either as toxic nodular goiter or Graves’ disease -with or without Graves’ orbitopathy (GO, inflammation and protrusion of the eyes)) - and autoimmune hypothyroidism. The symptomatology is often diffuse, sharing features with many other diseases (fatigue, palpitations, dry skin, depression, uneasiness, etc.) as well as with the non-pathological fluctuations of well-being and function in life. Therefore, thyroid diseases may go un-diagnosed for many years in some patients and at the time of diagnosis, most patients have reduced quality of life [3,4]. The diseases are chronic, but relevant treatment is available. In general though, there is a lag in treatment effect for thyroid diseases of up to several months and population-based studies document excess morbidity and mortality, also when adequately treated [5,6]. Eventually, the quality of life of the majority of patients is restored [4,7]. However, studies indicate that
a substantial minority do not regain their premorbid level of well-being and function [8,9]. Valid and reliable measures of health-related quality of life are necessary in order to describe the patients' experiences of the diseases adequately and for intervention studies attempting to improve treatment efficacy. Therefore, there has been a growing interest within thyroidology in measuring patient-reported outcomes (PRO), leading to the development of a comprehensive PRO measuring thyroid-related quality of life, the ThyPRO. Due to the fact that individual thyroid diseases often co-exist (e.g., goiter and hyperthyroidism) and that treatment of one disease entity may lead to another (e.g., removal of a goiter leading to hypothyroidism), the ThyPRO was developed as a comprehensive thyroid-related measure, aimed at any benign thyroid disease.

The content of the ThyPRO addresses the impact of all benign thyroid diseases [10,11]. The validation of the current version has included evaluation of clinical validity in terms of known-groups comparisons and reliability in terms of internal consistency and test-retest reliability [12,13]. Further, the ThyPRO's dimensionality or construct validity has been established by multi-trait scaling [12]. However, within such a framework, it is not possible to test the overall fit of a model [14], nor can misfit of items be modeled specifically.

The growing interest in applying the ThyPRO in clinical studies [7,15,16] and even in daily clinical practice has motivated efforts to develop shorter versions of the instrument as well as versions applicable to ecological momentary assessments. Development of such versions can be informed by the application of item response theory (IRT) models, which also provide a more detailed description of measurement precision and can provide data for interpretability of the ThyPRO. However, IRT models require additional, more detailed examinations of the dimensionality of the ThyPRO scales.

Structural equation models provide a latent variable modeling framework that is useful in detailed examinations of dimensionality. The measurement part of structural equation models can be used to assess the dimensionality of measured variables such as questionnaire items, using confirmatory factor analysis (CFA) for categorical data. Structural equation modeling can also test relationships among modeled latent variables (i.e., structural part of the models) [17-21]. We will exploit the former in the detailed analyses of the dimensionality of the ThyPRO scales, including overall test of model fit. We will use the structural part of the modeling approach when attempting to understand, through investigative modeling, any possible item misfit identified during the CFA step.

Thus, the purpose of the present study was to evaluate dimensionality of the ThyPRO scales in a sample of patients with a broad spectrum of thyroid diseases and to attempt to understand possible item misfit through investigative structural equation modeling.

**Methods**

**The ThyPRO questionnaire**

The current 85-item version of ThyPRO measures quality of life in 13 scales, covering physical (4 scales) and mental (2 scales) symptoms, function and well-being (3 scales) and participation/social function (4 scales) and one single item about overall quality of life. Content and scale structure were derived from a literature search [8] and from expert and patient interviews [10] and the development was conducted within a classical health-related quality of life theoretical framework [22-25]. Items are rated on a five-point scale from 0 = not at all to 4 = very much, with a reference period of 4 weeks. Thirteen scales are scored by reverting positively worded items and rescaling item scores from 0 (best QoL - absence of symptoms) to 100 (worst QoL – maximum level of symptoms) and taking the average across the items in the scale – i.e., standard summation and linear transformation.

**Patient population**

The patient population comprised a cross-sectional sample of 907 patients attending two university hospital endocrine outpatient clinics during 2007 (Table 1 (For further details, see reference [13])). At one center, all consecutive patients newly referred to the clinic were invited to participate; at the other center, all patients attending the clinic during a specified period of time were invited, regardless of their referral time. Thus, patients from the former were mainly newly diagnosed whereas from the latter most were already receiving treatment. All common benign thyroid diagnoses were represented, as were various stages of disease and treatment. Clinical description of the patients included physical examination, ultrasonographic imaging and biochemical testing. The overall response rate was 69%. The project was approved by the local ethical committee (KF01 2006–1579) and the Danish Data Protection Agency and was registered at ClinicalTrials.gov (NCT00150033).

**Statistical analyses**

Prior to any of the statistical analyses mentioned below, a content analysis of each scale was performed to identify items which might be less associated with the remaining items in the same scale, and item pairs which might be closely related to one another after being accounted for by the scale (local item dependence). This was done to provide a content-based guidance to model fitting.

Then a one-factor confirmatory model for ordinal data was fitted to each individual scale [26,27], using Mplus (version 7.11) [28]. The ordinal items were regressed on the scale-factor by probit regressions estimated by a robust weighted least squares estimator with mean and
variance adjustment (WLSMV) [28,29]. Appropriateness of the initial one-factor model for each scale was assessed by: 1) overall goodness-of-fit statistics including the comparative fit index (CFI) and the root mean square error of approximation (RMSEA), where CFI >0.95 and RMSEA < 0.08 were regarded as appropriate fit [30-34]; 2) magnitude of factor loadings; 3) model residual correlations (RC) and 4) modification indices (MI) [28,35]. For the latter three criteria, their magnitude was evaluated in comparison to other items in the scale and in an integrative manner, taking all three criteria under consideration at once, so no strict thresholds were applied for each criterion. In general though, modification indices >100 and residual correlations > |.10| were taken as indices of lack of fit (local dependence or lack of convergent validity), but smaller values could also give rise to model revision considerations, if several indices pointed in the same direction; e.g., if an item had a modification index of 40 for a specific residual correlation (a “WITH”-statement in Mplus) and also had residual model correlations with several items. Revisions to improve model fit were based on both confirmatory factor modeling and content analysis, including specification of residual correlations among items, omission of poorly associated items from the models, and specification of sub-factors (for example among positively worded items in a scale). For scales where secondary factors seem plausible, a bifactor model was fitted to evaluate the dominance of the primary factor when secondary factors were modeled. A bifactor model specifies that each item is regressed on both a general and a group (secondary) factor, and the general and group factors are uncorrelated with each other [34,36-39]. The magnitude of loadings on the general and group factors were compared. The two-item scale on impaired sex life was not examined in this step, since a separate factor analysis of a two-item scale is not useful.

In an attempt to understand any possible item misfit identified through individual scale analyses, hypotheses which could explain the misfit were sought. These hypotheses were evaluated in a combined, investigational multidimensional model, where the individual scale factors were allowed to correlate freely. Also items were cross-loaded on multiple scale factors when necessary to explore a better understanding of item misfit. For example, if an item in a physical symptoms scale, e.g., “Palpitations”, had low own-factor loadings, it could be hypothesized that this was due to palpitations being influenced by mental health, e.g., as part of anxiety. Then cross-loading of this item on the mental symptoms scales would be specified and evaluated in the combined model.

In order to examine the stability of the model across various estimation techniques, the overall final model was compared with graded response multidimensional IRT models [40], fitted with the Mplus program [28]. For computational reasons, a 13-dimensional IRT model could not be estimated, so the model was broken down to four separate models, each containing scales with cross-loadings across scales. Stability was examined by comparing the estimated factor scores for each patient from the SEM vs. the IRT-model using intra-class correlations.

### Results

#### Fitting unidimensional models to each individual ThyPRO scale

Table 2 shows the results of the content analyses and the confirmatory factor analyses of the ThyPRO scales in their current version. In general, loadings were high in all scales and CFI was also high for the vast majority of scales. In contrast, for most scales, RMSEA was not below the 0.08 threshold for appropriate fit. Model parameters indicative of item misfit are presented to the right in Table 2. The consequential remodeling resulted in the revised scales presented in Figure 1 and the remodeling as well as the overall goodness-of-fit statistics are described separately for each scale in the following text.
Table 2 Content analysis and confirmatory factor analyses of the individual ThyPRO scales

<table>
<thead>
<tr>
<th>Item #</th>
<th>Abbreviated item content</th>
<th>Possible misfit from content analysis</th>
<th>Factor loading</th>
<th>Indication of local dependence</th>
<th>Indication of item misfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goiter Symptoms</td>
<td></td>
<td></td>
<td></td>
<td>CFI=0.95 RMSEA=0.16(0.15-0.16)</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>Sense of fullness in neck</td>
<td>0.87</td>
<td>Mi: LD with 2b</td>
<td>Low loading</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>Visible swelling on neck</td>
<td>0.60</td>
<td>Mi and RC: LD with 2a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2c</td>
<td>Pressure in throat</td>
<td>0.90</td>
<td>RC: LD with 2g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2d</td>
<td>Pain in front of neck</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2e</td>
<td>Throat pain felt in ears</td>
<td>0.60</td>
<td>Low loading and low IC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2f</td>
<td>Lump in throat</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2g</td>
<td>Clear throat often</td>
<td>0.69</td>
<td>Mi: LD with 2l, RC: LD w. 2c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2h</td>
<td>Discomfort swallowing</td>
<td>0.94</td>
<td>Mi: LD with 2i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2i</td>
<td>Difficulty swallowing</td>
<td>0.92</td>
<td>Mi: LD with 2h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2j</td>
<td>Sense of suffocating</td>
<td>0.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2l</td>
<td>Hoarseness</td>
<td>0.56</td>
<td>Mi: LD with 2g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyperthyroid Symptoms</td>
<td></td>
<td></td>
<td></td>
<td>CFI=0.80 RMSEA=0.18(0.17-0.19)</td>
<td></td>
</tr>
<tr>
<td>2m</td>
<td>Trembling hands</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2n</td>
<td>Increased sweating</td>
<td>0.71</td>
<td>Mi: LD with 2q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2o</td>
<td>Palpitations</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2p</td>
<td>Shortness of breath</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2q</td>
<td>Sensitive to heat</td>
<td>0.70</td>
<td>Mi: LD with 2n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2s</td>
<td>Increased appetite</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2t</td>
<td>Loose stools</td>
<td>0.75</td>
<td>Low IC and large neg. RCs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2u</td>
<td>Upset stomach</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothyroid Symptoms</td>
<td></td>
<td></td>
<td></td>
<td>CFI=0.98 RMSEA=0.10(0.06-0.14)</td>
<td></td>
</tr>
<tr>
<td>2r</td>
<td>Sensitive to cold</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2f</td>
<td>Swollen hands or feet</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2gg</td>
<td>Dry skin</td>
<td>0.86</td>
<td>RC: LD with 2hh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2hh</td>
<td>Itching skin</td>
<td>0.63</td>
<td>RC: LD with 2gg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eye Symptoms</td>
<td></td>
<td></td>
<td></td>
<td>CFI=0.94 RMSEA=0.11(0.09-0.11)</td>
<td></td>
</tr>
<tr>
<td>2w</td>
<td>Watery eyes</td>
<td>0.62</td>
<td>Mi and RC: LD with 2x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2x</td>
<td>Bags under the eyes</td>
<td>0.59</td>
<td>Mi and RC: LD with 2w</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2y</td>
<td>Grittiness in eyes</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2z</td>
<td>Reduced sight</td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2aa</td>
<td>Pressure in eyes</td>
<td>0.87</td>
<td>Mi: LD with 2cc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2bb</td>
<td>Double vision</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2cc</td>
<td>Pain in eyes</td>
<td>0.86</td>
<td>Mi: LD with 2aa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2dd</td>
<td>Sensitive to light</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tiredness</td>
<td></td>
<td></td>
<td></td>
<td>CFI=0.99 RMSEA=0.28(0.26-0.28)</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>Been tired</td>
<td>0.90</td>
<td>Mi: LD with 3b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td>Been exhausted</td>
<td>0.93</td>
<td>Mi: LD with 3a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td>Difficult get motivated</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3d</td>
<td>Felt worn out</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4a</td>
<td>Full of life</td>
<td>0.93</td>
<td>Mi and RC: LD with 4b, 4c</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 Content analysis and confirmatory factor analyses of the individual ThyPRO scales (Continued)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Description</th>
<th>Items</th>
<th>CFI</th>
<th>RMSEA</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energetic</td>
<td>With 4a, 4c</td>
<td>0.98</td>
<td>MI and RC: LD with 4a, 4c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Able to cope with life</td>
<td>With 4a, 4b</td>
<td>0.95</td>
<td>MI and RC: LD with 4a, 4b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems remembering</td>
<td>With 5c</td>
<td>0.87</td>
<td>RC: LD with 5d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow or unclear thinking</td>
<td>With 5f</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty finding words</td>
<td>With 5a</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Been confused</td>
<td>*</td>
<td>0.85</td>
<td>RC: LD with 5a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty learning</td>
<td>With 5b</td>
<td>0.92</td>
<td>MI: LD with 5f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty concentrating</td>
<td>With 5b</td>
<td>0.91</td>
<td>MI: LD with 5e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems remembering</td>
<td>With 5c</td>
<td>0.87</td>
<td>RC: LD with 5d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow or unclear thinking</td>
<td>With 5f</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty finding words</td>
<td>With 5a</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Been confused</td>
<td>*</td>
<td>0.85</td>
<td>RC: LD with 5a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty learning</td>
<td>With 5b</td>
<td>0.92</td>
<td>MI: LD with 5f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty concentrating</td>
<td>With 5b</td>
<td>0.91</td>
<td>MI: LD with 5e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nervous</td>
<td></td>
<td>0.90</td>
<td>MI: LD with 6b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afraid or anxious</td>
<td></td>
<td>0.90</td>
<td>MI: LD with 6a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt tension</td>
<td></td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afraid being seriously ill</td>
<td>*</td>
<td>0.70</td>
<td>Low loading, neg. RC's</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unhappy</td>
<td>With 7g</td>
<td>0.92</td>
<td>MI: LD with 7f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>With 7i, 7f</td>
<td>0.76</td>
<td>MI: LD with 7i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-confident</td>
<td>*</td>
<td>0.74</td>
<td>MI: LD with 7i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty coping</td>
<td></td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not like yourself</td>
<td></td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easily stressed</td>
<td></td>
<td>0.81</td>
<td>MI: LD with 8i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood swings</td>
<td></td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irritable</td>
<td>With 8g</td>
<td>0.89</td>
<td>MI: LD with many other items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frustrated</td>
<td></td>
<td>0.91</td>
<td>Large neg. RC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angry</td>
<td>With 8e</td>
<td>0.80</td>
<td>MI: LD with many other items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt in control</td>
<td>With 8i</td>
<td>0.87</td>
<td>MI: LD with many, large neg. RC's</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt in balance</td>
<td>With 8h</td>
<td>0.91</td>
<td>MI: LD with 8, 8c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult with people</td>
<td></td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A burden to people</td>
<td></td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conflicts with people</td>
<td></td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others lack understanding</td>
<td>*</td>
<td>0.71</td>
<td>Low loading, neg. RC's</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult manage life</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limit leisure activities</td>
<td>With 11f</td>
<td>0.95</td>
<td>MI: LD with 11f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult participate in life</td>
<td></td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Goiter Symptoms

Three items were problematic (2b Visible swelling in front of neck, 2e Throat pain felt in ears and 2l Hoarseness), with relatively low loadings and indication of local dependence with other items. Two of these items were identified prior to the modeling as potentially less related to the concept. Two instances of local dependence among other items were identified (2c Pressure in throat vs. 2g Need to clear throat often and 2h Discomfort swallowing vs. 2i Difficulty swallowing, Table 2). When omitting the three items and modeling the local dependencies, an appropriately fitting unidimensional model was reached (Figure 1, CFI = 0.99, RMSEA(90%CI) = 0.08(0.07-0.09)).

Hyperthyroid Symptoms

For one pair of items (2n Increased sweating vs. 2q Sensitive to heat), the modification index suggested local dependence and one item (2t Loose stools) had large negative residual correlations with other items, when the initial model was estimated. When omitting the latter and fitting the local dependence, a unidimensional model obtained an appropriate fit to the data (Figure 1, CFI = 0.97, RMSEA(90%CI) = 0.08(0.07-0.09)).

Hypothyroid Symptoms

When modeling the expected local dependence between the items concerning skin (2gg Dry skin vs. 2hh Itching skin), an appropriate fit between an overall unidimensional model and data was demonstrated for this scale (Figure 1, CFI = 1.0, RMSEA(90%CI) = 0.00(0.00-0.09)).

Eye Symptoms

With the specification of two local dependence-pairs (2w Watery eyes vs. 2x Bags under eyes and 2aa Pressure in eyes vs. 2cc Pain in eyes), an appropriate fit of a unidimensional model was found (Figure 1, CFI = 0.99 RMSEA(90%CI) = 0.06(0.04-0.07)).

Tiredness

Despite quite high factor loadings, overall goodness-of-fit was poor for this scale. To avoid floor problems, three items had been formulated positively for this scale. The positively worded items had high positive residual correlations and modification indices. A bifactor model distinguishing positively from negatively worded items was therefore evaluated (Figure 2, Panel A). Although the positively worded items had high loadings on the positive factor (Vitality), loadings on the general factor were higher. When modeling the local dependence among positively worded items as residual correlations and also allowing for the local dependence between 3a and 3b, the model had good fit (Figure 1, CFI = 1.0, RMSEA(90%CI) = 0.02 (0.00-0.04)).

Cognitive Complaints

All items had high loadings in the initial model (Table 2). When specifying two pairs of local dependence, suggested by modification indices (5a Problems remembering vs. 5d Been confused and 5e Difficulty learning vs. 5f Difficulty concentrating), overall model fit was appropriate (Figure 1, CFI = 1.0 RMSEA(90%CI) = 0.07 (0.05-0.09)).

Anxiety

According to overall goodness-of-fit indices, the initial model did not obtain an appropriate fit to the data (Table 2). When fitting a model by excluding the item identified as less related with the other items (6d Afraid
being seriously ill) and by specifying two item pairs with local dependence (6a Nervous vs. 6b Afraid or anxious and 6e Uneasy and 6f Restless), appropriate fit was obtained (Figure 1, CFI = 1.0, RMSEA(90%CI) = 0.07 (0.04-0.10)).

**Depressivity**

All items had high loadings (Table 2). However, only after specification of two local dependence pairs (7e Crying easily vs. 7f Unhappy and 7g Happy vs. 7i Self-confident), was an appropriate overall fit to data reached (Figure 1, CFI = 1.0 RMSEA(90%CI) = 0.07 (0.05-0.09)).

**Emotional Susceptibility**

In contrast to most other concepts measured by ThyPRO, this scale measures a unique aspect of mental health identified through qualitative analysis of patient views. Thus, it is not classically described as a separate concept. It is, however, an important aspect according to the patients and a prominent feature particularly among patients with thyroid autoimmunity [10]. According to the overall fit indices, these items do not appropriately conform to a unidimensional model, despite high factor loadings (Table 2). Several items had high inter-item residual correlations and were attempted to be modeled as a separate “Anger” sub-factor (Figure 2, Panel B). However, as shown in Figure 2, the sub-factor loadings were rather low. Four items had to be omitted in order to obtain appropriate fit between a unidimensional model and the data (Figure 1, CFI = 1.0 RMSEA(90% CI) = 0.08(0.05-0.11)). A local dependence (8c Easily stressed vs. 8i Felt in balance) was also modeled.
**Impaired Social Life**
Appropriate, albeit not good overall goodness-of-fit indices were found for the initial unidimensional model. Excluding the lowest-loading item (10d People lack understanding), which was also pre-specified as possibly less associated, resulted in a just-identified model, hence with perfect fit (Figure 1, CFI = 1.0 RMSEA(90%CI) = 0.00(0.00-0.00)).

**Impaired Daily Life**
With the specification of one local dependence (11d Difficulty getting around vs. 11e Everything takes longer), a unidimensional model fit the data appropriately (Figure 1, CFI = 1.0, RMSEA (90%CI) = 0.08(0.07-0.10)).

**Cosmetic Complaints**
The initial unidimensional model had almost appropriate goodness-of-fit indices (Table 2). When modeling one local dependence (13a Disease affect appearance vs. 13b Unsatisfied with appearance) and leaving out the very nonspecific item concerning feeling too fat (13g), a good fit between model and data was found (Figure 1, CFI = 1.0 RMSEA(90%CI) = 0.05(0.02-0.08)).

**Investigative modeling of possible item misfit within one combined multidimensional model**
This investigative model is presented in Table 3. The hypotheses concerning the reason for misfit of the omitted items are presented in the second column of the table. In these models, the possible sub-factors tested in bifactor models (Figure 2) were specified as residual correlations among the involved items. In the third column of Table 3, it is specified how these hypotheses were modeled in the combined multidimensional model, where all the factors were evaluated simultaneously and were allowed to correlate freely. The results of this investigative modeling are described in the rightmost column of Table 3. Generally, a closer association was found between items and their own scale for the items in the multidimensional model (e.g. items 2e, 2t and 10e), than in the unidimensional model for each scale. For most items, the hypothesized explanations for the apparent misfit were confirmed. Thus, 2b Visible swelling on neck was indeed associated with Cosmetic Complaints (−0.23). Item 2l Hoarseness did load also on the Hypothyroid Symptoms scale (0.22), 2t Loose stools was negatively associated with particularly Hypothyroid Symptoms (−0.55), and a negative
association between 6d Afraid of being seriously ill and time since diagnosis was found. In contrast, no relationship between item 10e Other people lack understanding and mental health scales was found. Item 13g Feeling too fat was associated with both Hypothyroid Symptoms (−0.16), Anxiety (−0.22) and Depressivity (0.15), and had low loading on its own factor (0.53).

In analyses of concordance of results from SEM and the IRT-model, high intra-class correlation coefficients (0.94-0.99) were found for all 13 scales, when comparing factor scores derived by the SEM with IRT score estimates (Table 4).

**Discussion**

The purpose of the present study was to evaluate the dimensionality of the ThyPRO scales and to detect and understand potential item misfit. Since an established scale structure already exists for the ThyPRO, we used a combination of confirmaory factor analyses of the individual scales and a combined multidimensional model comprising all 13 ThyPRO scales. In case of misfit for each individual scale, we revised the model to achieve the best description of data. In general, items had high loadings on their own factors and the comparative fit indices were high, but for the majority of the scales, the root means square error of approximation indicated that a simple unidimensional model was not fitting the data sufficiently well. Based on prior expectations informed by content analyses, modeling results (model inter-item correlations and model residual correlations) and on model modification indices, the models were adjusted in order to reduce the overall misfit. For all scales, an appropriate fit according to the overall goodness-of-fit indices could be reached. During this process, a total of 11 items were left out of the models and 18 residual correlations indicating local dependence were specified.

In most instances, the magnitude of the residual correlations representing local dependencies was small, and the loading on the relevant general factor was still high. Most of the residual correlations were among very similarly worded items. Such local dependencies are not problematic for the current scoring of the ThyPRO, but may lead researchers to overestimate the precision gained by the instrument, because locally dependent items provide less measurement precision than assumed by standard psychometric analyses [41]. Moreover, one of the items

<table>
<thead>
<tr>
<th>Item</th>
<th>Hypothesized reason for misfit</th>
<th>Investigative modeling of the hypothesized reason for misfit</th>
<th>Results of the investigative modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b Visible swelling on neck from the Goiter Symptoms scale</td>
<td>May relate to cosmetic concerns, rather than being a symptom</td>
<td>Item was allowed to cross-load on the Cosmetic Complaints factor</td>
<td>Loaded −0.23 on the Cosmetic Complaints factor. Loading on own factor: 0.68</td>
</tr>
<tr>
<td>2e Throat pain felt in ears from the Goiter Symptoms scale</td>
<td>May be relevant only for patients with subacute thyroiditis, during the acute inflammatory phase.</td>
<td>No marker of acute inflammation is available in the clinical database describing the patients. Only 9 patients in this sample had subacute thyroiditis</td>
<td>Loading on own factor in the full model: 0.75</td>
</tr>
<tr>
<td>2l Hoarseness from the Goiter Symptoms scale</td>
<td>Hoarseness is also a classical symptom of hypothyroidism. Might relate more to hypothyroidism than to goiter.</td>
<td>Item was allowed to cross-load on the Hypothyroid Symptoms factor</td>
<td>Loaded 0.22 on Hypothyroid Symptoms factor. Loading on own factor: 0.46</td>
</tr>
<tr>
<td>2t Loose stools from the Hyperthyroid Symptoms scale</td>
<td>Might be a non-specific physical symptom</td>
<td>Item was allowed to load on the other physical symptoms factors, except for Eye Symptoms</td>
<td>Loaded 0.15 on Goiter Symptoms factor and −0.55 on Hypothyroid Symptoms. Loading on own factor: 1.20</td>
</tr>
<tr>
<td>6d Afraid of being seriously ill from the Anxiety scale</td>
<td>May be related to not being fully examined yet, and thus an initial fear of e.g. cancer has not yet been ruled out completely</td>
<td>Item was regressed on time since diagnosis. A significant negative association with time since diagnosis was found</td>
<td></td>
</tr>
<tr>
<td>10e Other people lack understanding from the Impaired Social Life scale</td>
<td>May relate more to depressive mood and emotional distress than the other items in the Social Life scale</td>
<td>Item was allowed to cross-load on the Depressivity and the Emotional Susceptibility factor</td>
<td>No significant loading on Depressivity or Emotional Susceptibility was found. Loading on own factor: 1.08</td>
</tr>
<tr>
<td>13g Felt too fat from the Cosmetic Complaints scale</td>
<td>Weight gain is often experienced during hypothyroidism. Feeling too fat may also relate more to a negative self-esteem aspect of depressive mood</td>
<td>Item was allowed to cross-load on the Hypothyroid Symptoms and Depression and Anxiety factors</td>
<td>Loaded −0.16 on Hypothyroid Symptoms factor, −0.22 on Anxiety and 0.15 on Depressivity factor. Loading on own factor: 0.53</td>
</tr>
</tbody>
</table>
involved in such pairs would be potential candidates for omission in future IRT-modeling of the instrument and in the development of abbreviated versions of the ThyPRO.

However, such item reduction should be done with caution and should take clinical analyses and considerations into account.

Although positively worded items did tend to exhibit residual correlations, we found no consistent evidence of a method factor among the positively worded items. Similar studies with other outcome measures have previously found substantial influence of the value of the wording [36,42-44], whereas other studies either did not identify such an effect [45] or the identified effect had only minor influence on the results regarding the substantive factor [46].

We attempted to model potential item misfit identified during the dimensionality analyses of the existing ThyPRO scales. This was done within a model including all scales, which were allowed to correlate, in order to allow for cross-loadings of items to be examined and in order to evaluate if possible misfit identified during individual scale analyses was due to interrelation with other factors. In doing so, the hypothesized reason for misfit was confirmed in five of seven items; Item 2b, about visibility of the goiter, cross-loaded on Cosmetic Complaints. Item 2t, Loose stools, had a large negative loading on Hypothyroid Symptoms, as had 2l, Hoarseness. Both constipation and hoarseness are indeed salient and classical features of hypothyroidism [47]. The rather non-specific item 13g, Feeling too fat, which is a common complaint among hypothyroid patients and among hyperthyroid patients after treatment, had cross-loadings on several other scales and low loading on its own factor, also when modeled multidimensionally. Thus, these four items are very strong candidates for item reduction when developing abbreviated and focused versions of the scales or when fitting models where unidimensionality is a strong assumption, for example as in unidimensional IRT models.

A unique “duration of disease”-effect was observed for one item. Item 6d, Afraid of being seriously ill was negatively associated with time since diagnosis, indicating that the responses to this item reflects a relevant concern early in the disease course, for instance of a goiter being malignant, a concern that wanes as the diagnosis becomes more firmly established and malignancy thus ruled out. It thus measures something different from the other items in the scale, which are more classical indicators of an anxious state.

As an analysis of the robustness and appropriateness of the ordinal confirmatory WLSMV factor analysis, an alternative multidimensional IRT-based analysis was performed. Individual factor scores derived from each of these approaches were very similar, as illustrated by very high intra-class correlation coefficients. This corroborates the current simple scoring approach and the results of the present analyses.

The use of theoretically driven analyses within a clinically well-described and relatively (for thyroid diseases) large sample was a strength of this study. However, the analyses were carried out in one sample and should ideally be confirmed in a new independent sample. Furthermore, although the present sample comprised patients in all stages of disease and treatment, stability of the factor structure across time could not be evaluated, since the data did not contain longitudinal measurements.

In conclusion, each of the ThyPRO scales could be appropriately represented by a unidimensional model after minor revisions. Eleven items were identified in the unidimensional models as potentially misfitting and understood further by multidimensional modeling. Thus, overall the previous initial examinations of the construct validity of the scales [12] were corroborated using a more elaborate technique. Further, advanced psychometric modeling such as IRT, with strong assumptions about dimensionality, can be applied to the reduced scales. Finally, the locally dependent items identified here are strong candidates for removal, in future item reduction processes.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
TW designed the study, conducted the CFA analyses and drafted the manuscript. MG, ND and BG was involved in analysis strategy and provided substantial intellectual input to the manuscript. UFR, ÅKR, LH and SJB was involved in design of the study, inclusion of patients and provided substantial intellectual input to the manuscript. JBB was involved in analysis strategy, conducted the supplemental IRT analyses and provided substantial contributions.
intellectual input to the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This study has been supported by grants from the Danish Medical Research Council, Agnes and Knut Mark’s Foundation, Aase and Ejnar Danielsen’s Foundation, Else and Mogens Wedell-Wedelborg’s Foundation, the Genzyme Corporation, the Novo Nordisk Foundation, Arvid Nilsson’s Fund and the Danish Thyroid Foundation.

*Researchers who want to use the ThyPRO may contact the first author (torquil.watt@regionh.dk).

Author details
1Department of Endocrinology, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen Ø, Denmark. 2Institute of Public Health, University of Copenhagen, Øster Farimagsgade 5, 1014 Copenhagen K, Denmark. 3Department of Palliative Medicine, Bispebjerg Hospital, Bispebjerg Bakke 3, 2400 Copenhagen N, Denmark. 4Department of Quantitative Health Sciences, University of Massachusetts Medical School, 368 Plantation Street, 01605 Worcester, MA, USA. 5Department of Endocrinology and Metabolism, Odense University Hospital, Kløvervænget 6, 5000 Odense M, Denmark. 6National Research Centre for the Working Environment, Lersø Parkallé 105, 2100 Copenhagen, Denmark. 7OptumInsight, 24 Albion Rd, 02865 Lincoln, RI, USA.

Received: 11 March 2014 Accepted: 6 August 2014

References
Few items in the thyroid-related quality of life instrument ThyPRO exhibited differential item functioning

Torquil Watt · Mogens Groenvold · Laszlo Hegeduš · Steen Joop Bonnema · Åse Krogh Rasmussen · Ulla Feldt-Rasmussen · Jakob Bue Bjorner

Accepted: 18 June 2013 / Published online: 30 June 2013 © Springer Science+Business Media Dordrecht 2013

Abstract

Objective To evaluate the extent of differential item functioning (DIF) within the thyroid-specific quality of life patient-reported outcome measure, ThyPRO, according to sex, age, education and thyroid diagnosis.

Study design and setting A total of 838 patients with benign thyroid diseases completed the ThyPRO questionnaire (84 five-point items, 13 scales). Uniform and non-uniform DIF were investigated using ordinal logistic regression, testing for both statistical significance and magnitude ($\Delta R^2 > 0.02$). Scale level was estimated by the sum score, after purification.

Results Twenty instances of DIF in 17 of the 84 items were found. Eight according to diagnosis, where the goiter scale was the one most affected, possibly due to differing perceptions in patients with auto-immune thyroid diseases compared to patients with simple goiter. Eight DIFs according to age were found, of which 5 were in positively worded items, which younger patients were more likely to endorse; one according to gender: women were more likely to report crying, and three according to educational level. The vast majority of DIF had only minor influence on the scale scores (0.1–2.3 points on the 0–100 scales), but two DIF corresponded to a difference of 4.6 and 9.8, respectively.

Conclusion Ordinal logistic regression identified DIF in 17 of 84 items. The potential impact of this on the present scales was low, but items displaying DIF could be avoided when developing abbreviated scales, where the potential impact of DIF (due to fewer items) will be larger.

Keywords Differential item functioning · DIF · Thyroid diseases · Quality of life · Patient-reported outcome

Introduction

Measurement of health-related quality of life (HRQL), in terms of impact of disease and its treatment on physical, psychological, social and somatic domains of functioning and well-being [1] is becoming a natural part of outcome assessments in clinical research. With the development within the field of patient-reported outcomes research leading to measures with improving measurement properties, an interest in implementing such measurements in daily clinical practice is evolving [2, 3]. However, this
requires instruments with good and well-described measurement properties [4].

Clinically as well as scientifically, thyroidology has not had a leading position within this development, despite the fact that thyroid diseases are very common, affect individuals of all ages, and thyroid dysfunction has numerous direct effects on both physical and mental health [5–7]. Until recently, a disease-specific HRQL questionnaire validated in patient populations has only been available for patients with the much rarer eye disease, Graves’ orbitopathy. However, there has been an increasing interest in HRQL measurements for patients with thyroid diseases such as hyperthyroidism (Graves’ hyperthyroidism and toxic nodular goiter) and hypothyroidism (e.g., autoimmune hypothyroidism), which has lead to the development of a comprehensive thyroid-specific patient-reported outcome measure, the ThyPRO [8, 9].

The ThyPRO has been validated and can be used in clinical trials [10, 11]. Further exploration of its measurement properties is required if it is to be included in routine clinical practice. Further, since the current comprehensive version is rather long, development of abbreviated versions is justified, and such abbreviated versions could preferably be based on selection of items with the best measurement properties.

An important aspect of the measurement properties of an instrument and its scales is the degree to which individual items are subject to differential item functioning (DIF) [12–16]. An item in a multi-item scale is subject to DIF if different groups of respondents with the same level of the measured attribute score differently on that item [17]. For example, a mental health item concerning ‘crying’ has been shown to differ among men and women with the same level of depression [18]. When such DIF is present, the response is determined not only by mental health, but also by gender, and the measurement of mental health may thus become biased when comparing men and women.

DIF can be evaluated using either contingency table-based methods, e.g., Mantel-Haenzel analyses [19], logistic regression [14], structural equation modeling [20] or item response theory modeling [13], each having advantages and disadvantages [21]. For the present study, ordinal logistic regression was applied for several reasons [22]: the universality of this methodology and its ability to detect both uniform and nonuniform DIF (where magnitude or even direction of DIF varies across levels of the trait measured), the availability of an effect size measure, its moderate requirements regarding sample size and the possibility for testing continuous variables directly.

The purpose of the present study was to evaluate the extent of DIF within the thyroid-specific quality of life PRO measure, ThyPRO, according to sex, age, education and thyroid diagnosis.

Methods

Patient population

During February to November 2007, 838 nonpregnant patients with benign thyroid diseases (Table 1), attending two university endocrine outpatient clinics in Denmark, completed the ThyPRO questionnaire [10]. The overall response rate was 69 %. Only respondents with no missing item responses were included in the present study. Questionnaires were completed, and blood-samples were drawn a few days prior to the scheduled visit to the clinic. A physical examination, including thyroid ultrasound, was performed and a written, signed informed consent obtained.

Socio-demographic data and information about co-morbidity and nonthyroid medication were self-reported. Data regarding clinical biochemical measurements and thyroid imaging, exact diagnosis, previous and current treatment and time of diagnosis were obtained by chart review. The project was approved by the local ethical committee (KF01 2006-1579) and the Danish data protection agency and registered at ClinicalTrials.gov (NCT00150033).

Patient-reported outcome, ThyPRO

The ThyPRO is an 84-item thyroid-specific patient-reported outcome measuring HRQL by 13 scales [8–11]. Four scales measuring physical symptoms (number of items in parenthesis): Goiter Symptoms (11), Hyperthyroid Symptoms (8), Hypothyroid Symptoms (4) and Eye Symptoms (8); two scales measuring psychological symptoms: Anxiety (6) and Depressivity (7); three scales measuring well-being and function: Emotional Susceptibility (9), Tiredness (7) and Cognitive Impairment (6); and four scales measuring participation: Impaired Social Life (4), Impaired Daily Life (6), Impaired Sexlife (2) and Cosmetic Complaints (6). Each item is rated by the patient on a five-point Likert scale, and the 13 scales are derived by averaging and linearly transforming these item scores into their respective 0–100 scale score. Recent structural equation modeling has confirmed this scale structure (manuscript submitted). However, the analyses flagged some items as possibly misfitting and thus as candidates for reporting as single items or omission as part of an abbreviation process: ‘Visible swelling in front of neck’ and ‘Hoarseness’ from the Goiter Symptoms scale, ‘People lack understanding’ from the Impaired Social Life scale and ‘Felt too fat’ from the Cosmetic Complaints scale.

Statistical analyses

Univariate differences in level of scale scores among the groups were evaluated with ANOVA (SAS PROC GLM).
Table 1  Characteristics of the 838 patients, including mean ThyPRO scale scores (higher scores with worse symptoms/problems)

<table>
<thead>
<tr>
<th>Thyroid diagnosis</th>
<th>n</th>
<th>%</th>
<th>Goiter Symptoms</th>
<th>Hyperthyroid Symptoms</th>
<th>Hypothyroid Symptoms</th>
<th>Eye Symptoms</th>
<th>Tiredness</th>
<th>Cognitive complaints</th>
<th>Anxiety</th>
<th>Depressivity</th>
<th>Emotional Susceptibility</th>
<th>Impaired Social Life</th>
<th>Impaired Daily Life</th>
<th>Impaired Sexlife</th>
<th>Impaired Cosmetic Complaints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nontoxic goiter†</td>
<td>258</td>
<td>(31)</td>
<td>23</td>
<td>19</td>
<td>19</td>
<td>14</td>
<td>46</td>
<td>18</td>
<td>22</td>
<td>28</td>
<td>ns</td>
<td>31</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>Toxic nodular goiter</td>
<td>142</td>
<td>(17)</td>
<td>22</td>
<td>22</td>
<td>19</td>
<td>14</td>
<td>48</td>
<td>19</td>
<td>21</td>
<td>27</td>
<td>ns</td>
<td>29</td>
<td>11</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Graves’ hyperthyroidism</td>
<td>158</td>
<td>(19)</td>
<td>13</td>
<td>27</td>
<td>21</td>
<td>16</td>
<td>49</td>
<td>23</td>
<td>26</td>
<td>32</td>
<td>ns</td>
<td>37</td>
<td>12</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>Graves’ orbitopathy</td>
<td>91</td>
<td>(11)</td>
<td>11</td>
<td>24</td>
<td>22</td>
<td>19</td>
<td>48</td>
<td>27</td>
<td>26</td>
<td>36</td>
<td>ns</td>
<td>36</td>
<td>18</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>Autoimmune hypothyroidism</td>
<td>189</td>
<td>(23)</td>
<td>12</td>
<td>20</td>
<td>28</td>
<td>16</td>
<td>52</td>
<td>26</td>
<td>21</td>
<td>30</td>
<td>ns</td>
<td>35</td>
<td>13</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>728</td>
<td>(87)</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>20</td>
<td>57</td>
<td>20</td>
<td>25</td>
<td>33</td>
<td>ns</td>
<td>33</td>
<td>12</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>Men</td>
<td>110</td>
<td>(13)</td>
<td>14</td>
<td>18</td>
<td>15</td>
<td>15</td>
<td>44</td>
<td>16</td>
<td>17</td>
<td>23</td>
<td>ns</td>
<td>28</td>
<td>8</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Educational level (missing: n = 33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No vocational education†</td>
<td>132</td>
<td>(16)</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>20</td>
<td>57</td>
<td>20</td>
<td>25</td>
<td>33</td>
<td>ns</td>
<td>33</td>
<td>12</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>Studying</td>
<td>12</td>
<td>(1)</td>
<td>21</td>
<td>27</td>
<td>21</td>
<td>13</td>
<td>40</td>
<td>20</td>
<td>20</td>
<td>24</td>
<td>ns</td>
<td>34</td>
<td>10</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Apprenticeship</td>
<td>108</td>
<td>(13)</td>
<td>15</td>
<td>21</td>
<td>17</td>
<td>15</td>
<td>52</td>
<td>20</td>
<td>22</td>
<td>30</td>
<td>ns</td>
<td>33</td>
<td>11</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>Short theoretical (1–3 years)</td>
<td>242</td>
<td>(30)</td>
<td>18</td>
<td>22</td>
<td>22</td>
<td>16</td>
<td>50</td>
<td>23</td>
<td>23</td>
<td>30</td>
<td>ns</td>
<td>35</td>
<td>12</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Medium theoretical (3–5 years)</td>
<td>187</td>
<td>(23)</td>
<td>16</td>
<td>22</td>
<td>24</td>
<td>17</td>
<td>49</td>
<td>25</td>
<td>24</td>
<td>30</td>
<td>ns</td>
<td>35</td>
<td>13</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>Long theoretical (5+ years)</td>
<td>124</td>
<td>(15)</td>
<td>12</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>38</td>
<td>17</td>
<td>18</td>
<td>24</td>
<td>ns</td>
<td>28</td>
<td>9</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18–29 years‡</td>
<td>59</td>
<td>(7)</td>
<td>16</td>
<td>25</td>
<td>22</td>
<td>14</td>
<td>50</td>
<td>23</td>
<td>26</td>
<td>30</td>
<td>ns</td>
<td>38</td>
<td>15</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>30–39 years‡</td>
<td>149</td>
<td>(18)</td>
<td>14</td>
<td>23</td>
<td>24</td>
<td>14</td>
<td>50</td>
<td>22</td>
<td>24</td>
<td>30</td>
<td>ns</td>
<td>39</td>
<td>14</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>40–49 years‡</td>
<td>178</td>
<td>(21)</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td>17</td>
<td>52</td>
<td>24</td>
<td>23</td>
<td>31</td>
<td>ns</td>
<td>37</td>
<td>14</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>50–59 years‡</td>
<td>209</td>
<td>(25)</td>
<td>16</td>
<td>21</td>
<td>22</td>
<td>19</td>
<td>51</td>
<td>24</td>
<td>25</td>
<td>32</td>
<td>ns</td>
<td>35</td>
<td>12</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>60–69 years‡</td>
<td>158</td>
<td>(19)</td>
<td>17</td>
<td>21</td>
<td>21</td>
<td>16</td>
<td>43</td>
<td>19</td>
<td>19</td>
<td>26</td>
<td>ns</td>
<td>26</td>
<td>9</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>70+ years‡</td>
<td>85</td>
<td>(10)</td>
<td>17</td>
<td>19</td>
<td>22</td>
<td>18</td>
<td>49</td>
<td>19</td>
<td>18</td>
<td>26</td>
<td>ns</td>
<td>24</td>
<td>7</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Comorbidity (missing: n = 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No comorbidity†</td>
<td>432</td>
<td>(52)</td>
<td>16</td>
<td>20</td>
<td>20</td>
<td>14</td>
<td>46</td>
<td>20</td>
<td>21</td>
<td>27</td>
<td>ns</td>
<td>33</td>
<td>10</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Somatic comorbidity†</td>
<td>281</td>
<td>(34)</td>
<td>17</td>
<td>21</td>
<td>24</td>
<td>19</td>
<td>50</td>
<td>22</td>
<td>21</td>
<td>28</td>
<td>ns</td>
<td>30</td>
<td>11</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Psychiatric comorbidity‡</td>
<td>18</td>
<td>(2)</td>
<td>21</td>
<td>34</td>
<td>31</td>
<td>24</td>
<td>64</td>
<td>44</td>
<td>47</td>
<td>53</td>
<td>ns</td>
<td>58</td>
<td>22</td>
<td>27</td>
<td>31</td>
</tr>
</tbody>
</table>

† Thyroid diagnosis includes only patients with confirmed diagnosis.

‡ Age is presented as the median age in years.

§ Comorbidity includes only patients with confirmed comorbidity.
DIF was investigated using ordinal logistic regression [22]. For each item, three regression models were compared:

1. $\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_0 + \beta_1 \times \text{Scale}$
2. $\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_0 + \beta_1 \times \text{Scale} + \beta_2 \times \text{Test-variable}$
3. $\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_0 + \beta_1 \times \text{Scale} + \beta_2 \times \text{Test-variable} + \beta_3 \times (\text{Test-variable} \times \text{Scale})$

where $Y$ is the response option and $j = 1, 2, \ldots, c - 1$, where $c$ is the number of response categories in the item.

In Model 1, item responses were regressed on the scale only, i.e. all variation in the item was explained by the scale. In Model 2, the exogenous variable to be tested for DIF was added, and in Model 3, the interaction term between the test variable and the scale (representing nonuniform DIF) was also entered into the regression equation. Significance of DIF was tested by comparing 2log likelihood values for the nested models: An overall test of DIF was obtained by comparing Model 3 with Model 1 (difference test). This yields a $\chi^2$ statistic with degrees of freedom equal to $2 \times (\text{number of groups minus one})$, i.e. $2df$ for the classical two groups comparisons or continuous variables. If a similar comparison of Model 2 and Model 3 (degrees of freedom equal to number of groups minus one, i.e. $1df$ for the classical two-group comparison) is nonsignificant, there is evidence of uniform DIF, i.e. the same ‘degree’ of DIF is irrespective of scale level, whereas if it is significant, there is evidence of nonuniform DIF, i.e. the magnitude of the DIF differs among respondents with varying levels of the attribute. Significance level was set to $p < 0.05$ with a Bonferroni correction for multiple testing [23, 24] within each scale.

In addition to statistical significance, magnitude of DIF (i.e., effect size) was also considered, before flagging an item as evidencing DIF [22, 25]. The magnitude measure was obtained by comparing the three models in terms of $R^2$ (i.e., Nagelkerke ‘Pseudo $R^2$’) [26, 27], i.e. ‘variance explained’ by each model. Magnitude of uniform DIF was obtained by subtracting $R^2$ for Model 1 from $R^2$ for Model 2 and likewise for nonuniform DIF (Model 3 minus Model 1). In order for a possible DIF to be judged substantial, the increase in $R^2$ had to be more than 0.02. In other words, the DIF should explain more than 2 % of the variance in the item score. This is lower than that suggested by other authors (e.g., 3.5 %) [22, 25, 28], but such higher thresholds may lead to high Type II error rates [29, 30], and in our experience, 2 % is a relevant level [27, 31–34].

The scale level was estimated by the sum score, after positively worded items had been reverted. The sum score included also the item being tested for DIF [19, 22, 35], but items with DIF according to the exogenous variable

<table>
<thead>
<tr>
<th>Table 1 continued</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Both somatic and psychiatric complaints</td>
</tr>
<tr>
<td>Comorbidity not specified</td>
</tr>
</tbody>
</table>
| Differences in mean scale scores among the groups were tested with ANOVA (SAS PROC GLM), and the significance level of any identified differences in variance (F-test) is indicated by * (p < 0.05), ** (p < 0.01) and *** (p < 0.001) for groups with mean scale scores statistically different (p < 0.05, parameter estimates in PROC GLM) from the reference group (marked with superscript r), the mean score is highlighted with bold italic  
| a According to a pre-specified list of somatic diseases and other psychiatric disease  
| b Pre-specified list comprising depression, anxiety and other psychiatric disease | 32 4 17 36 26 24 8 16 26 24 | 65 8 16 23 19 10 8 12 18 24 | 32 4 17 36 26 24 8 16 26 24 | 65 8 16 23 19 10 8 12 18 24 | 32 4 17 36 26 24 8 16 26 24 |

Differences in mean scale scores among the groups were tested with ANOVA (SAS PROC GLM), and the significance level of any identified differences in variance (F-test) is indicated by * (p < 0.05), ** (p < 0.01) and *** (p < 0.001) for groups with mean scale scores statistically different (p < 0.05, parameter estimates in PROC GLM) from the reference group (marked with superscript r), the mean score is highlighted with bold italic

a According to a pre-specified list of somatic diseases and other psychiatric disease
b Pre-specified list comprising depression, anxiety and other psychiatric disease

---

123
being tested were excluded, i.e. the sum score was purified [22, 36]. Purification was obtained by iterative DIF tests with subsequent exclusion of items displaying DIF for the exogenous variable in question.

SAS PROC LOGISTIC [37] was used for the ordinal logistic regression analyses, which were performed as complete cases analyses.

Impact of DIF on scale score was evaluated by plotting the expected mean item responses for each DIF grouping, as estimated by the regression model, against the level of the respective purified scales. For categorical groups, the mean item response for each group was obtained from the ordinal regression coefficients, representing the item response threshold values, whereas for age, the expected values were calculated for the median age and the 25th and the 75th percentiles.

**Results**

With univariate comparisons, differences among groups were found, as shown in Table 1. Except for Anxiety and Depressivity, differences were found among diagnostic groups for all scales. Most variation was associated with diagnostic group (e.g., patients with goiter scoring higher on goiter scale, patients with Graves’ disease scoring higher on hyperthyroid symptoms, etc.) and with psychiatric comorbidity.

In total, among the 84 items being tested, 20 instances of DIF were found, involving 17 items within 8 of the 13 ThyPRO scales. Eight instances of DIF according to diagnosis were found, all of which involved symptom scales and all but one of these were physical symptoms scales. The Goiter Symptoms scale was most affected with

---

**Fig. 1** DIF according to diagnosis, in the Goiter scale. For each item with DIF, the mean expected item response, according to the logistic regression model, for each group, for each level of the corresponding scale, is depicted. For example, for DIF1, at the middle level of the scale (dotted vertical line, corresponding to a score of about 50), patients with autoimmune hypothyroidism have a mean response to the relevant item of 2 (on the 0–4 item score), whereas patients with nontoxic goiter have a mean response around 2.7. DIF1–DIF4 refer to the DIF reference designations in Table 2.

**Table 1**

<table>
<thead>
<tr>
<th>Scale</th>
<th>DIF1</th>
<th>DIF2</th>
<th>DIF3</th>
<th>DIF4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense of fullness in neck</td>
<td>2.7</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure in throat</td>
<td></td>
<td></td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Visible swelling in front of neck</td>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td>Hoarseness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend**

- Non-toxic goiter
- Graves’ orbitopathy
- Toxic nodular goiter
- Autoimmune hypothyroidism
- Graves’ hyperthyroidism
4 instances of DIF. As seen in Fig. 1, DIF1–DIF3 were mainly due to patients with goiter scoring higher on these items for a given level of goiter symptoms, compared to patients with the autoimmune thyroid diseases. DIF3 was only just above the 2% cutoff ($\Delta R^2 = 0.022$), whereas DIF1 and DIF4 were 5.3 and 4.0, respectively, and DIF2 was large (12.0%). For the latter, the patients with non-toxic goiter scored more than one category higher than patients with Graves’ orbitopathy (Fig. 1). DIF according to diagnosis was also found for the symptom scales Hyperthyroid Symptoms, Eye Symptoms and Anxiety (Table 2; Fig. 2). For the Hyperthyroid Symptoms scale, patients with autoimmune hyperthyroidism, (i.e., Graves’ hyperthyroidism) had a higher probability of reporting ‘Trembling hands’ than other disease groups. For the Eye Symptoms scale, patients with thyroid eye disease (i.e., Graves’ orbitopathy) were less likely to report ‘Grittiness in eyes,’ for a given level of Eye Symptoms and patients with goiter were less likely to report ‘Reduced sight.’ For the Anxiety scale, patients with a nontoxic goiter, (i.e., an enlarged thyroid, without biochemical thyroid dysfunction) reported more ‘Concern about being seriously ill’ (most often implying fear of cancer), for a given level of Anxiety.

Eight instances of DIF according to age were found (DIF9–DIF16, Table 3). Two of these were nonuniform (DIF10 and DIF14) and are illustrated in Fig. 3. Five of the eight DIFs were in positively worded items, and for all of these, younger patients were more likely to endorse positively on the positively worded items than older patients at the same level of the involved scales (Tiredness and Depression, respectively, cf. regression coefficients in Appendix 1 of ESM for details).

One instance of DIF according to gender (DIF17, uniform; women more likely to respond positively to the item about crying easily, provided the same level of depressivity, compared to men) and three instances according to educational level were found (Table 4). Two of the latter were nonuniform and are illustrated in Fig. 3. However, the number of students (i.e., patients currently studying at, e.g., a university) in this sample was very low (cf. Table 1). Analyses regarding DIF according to educational level were therefore re-run without students: The DIF in the Cosmetic Complaints scale was no longer present (DIF20), but DIF was still present within the Depression and the Social Life scale (DIF18 and DIF19). This provided some clearer results: Patients without an education scored lower (i.e., had less difficulty being with other people, for a given scale score) on the item ‘Difficulty being with other people,’ compared to patients with an education. This item is the most emotionally neutrally formulated item within the scale.

Table 2 Differential item functioning (DIF) according to thyroid diagnosis

<table>
<thead>
<tr>
<th>Abbreviated item wording</th>
<th>Significance of DIF</th>
<th>Variance explained by DIF (%)</th>
<th>Direction of DIF</th>
<th>DIF reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goiter Symptoms scale (threshold for significance level with Bonferroni correction: 0.0045)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of fullness in neck</td>
<td>$&lt;0.0001$</td>
<td>5.3</td>
<td>Patients with nontoxic goiter and toxic nodular goiter scored higher</td>
<td>DIF1</td>
</tr>
<tr>
<td>Visible swelling in front of neck</td>
<td>$&lt;0.0001$</td>
<td>12.0</td>
<td>Patients with nontoxic goiter and toxic nodular goiter scored higher</td>
<td>DIF2</td>
</tr>
<tr>
<td>Pressure in throat</td>
<td>0.0001</td>
<td>2.2</td>
<td>Patients with nontoxic goiter and toxic nodular goiter scored higher</td>
<td>DIF3</td>
</tr>
<tr>
<td>Hoarseness</td>
<td>0.0001</td>
<td>4.0</td>
<td>Patients with Graves’ orbitopathy scored higher</td>
<td>DIF4</td>
</tr>
<tr>
<td>Hyperthyroid Symptoms scale (threshold for significance level with Bonferroni correction: 0.0063)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trembling hands</td>
<td>0.0001</td>
<td>3.0</td>
<td>Patients with Graves’ hyperthyroidism and Graves’ orbitopathy scored higher</td>
<td>DIF5</td>
</tr>
<tr>
<td>Grittiness in eyes</td>
<td>0.0039</td>
<td>2.0</td>
<td>Patients with Graves’ orbitopathy scored lower</td>
<td>DIF6</td>
</tr>
<tr>
<td>Reduced sight</td>
<td>0.0009</td>
<td>3.3</td>
<td>Patients with Graves’ hyperthyroidism and Graves’ orbitopathy scored lower</td>
<td>DIF7</td>
</tr>
<tr>
<td>Anxiety scale (threshold for significance level with Bonferroni correction: 0.0083)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concerned being seriously ill</td>
<td>$&lt;0.0001$</td>
<td>2.7</td>
<td>Patients with nontoxic goiter scored markedly higher</td>
<td>DIF8</td>
</tr>
</tbody>
</table>

Patients were classified according to their state at time of diagnosis: nontoxic goiter, toxic nodular goiter, Graves’ hyperthyroidism, Graves’ orbitopathy and autoimmune hypothyroidism. Abbreviated wording of item and its associated scale is presented, along with the $p$ value from the $\chi^2$ test for DIF and the increase in explained variance (i.e., increment in $R^2$) when adding the DIF term to the regression equation. Finally, the direction of the DIF is indicated, and a reference for use in the text and figures is assigned (DIF reference). All DIFs according to diagnosis were uniform.

© Springer
with the other items indicating some negative aspects of social life impairment: ‘A burden to other people,’ ‘Conflicts with other people’ and ‘Surroundings lack understanding.’

In terms of the influence of the identified DIF on the scale score, the item with largest effect size (DIF2) only had minor influence on its scale because the scale consists of 11 items. From Fig. 1, bottom, left, the difference between the two extreme groups is about one item category, corresponding to a score difference of 2.3 on the 0–100 Goiter Symptoms scale (11 five-point items yield a raw summary score of 44; therefore, a difference of one measurement point corresponds to 100/44 = 2.3 on the 0–100 metric), which is below the typical size of a minimal important difference (MID, around 8 for a 0–100 scale [38]). In fact, all differences between the DIF groups according to the graphs were small (0.1–2.3, which can be extracted from the graphs and the number of items in each scale), except for DIF19 and DIF20, which at the scale level where they are largest (around scale level 4/5 of scale max for DIF 19 [see Fig. 3, top, right], and 2/3 of scale max for DIF 20 (Fig. 3, bottom, right)) corresponds to a difference of 9.8 and 4.6 points (DIF 19: a raw difference of 1.56 points (Fig. 3) out of a raw scale range of 0–16 points yields a 9.8 points difference on a 0–100 scale; DIF 20: a raw difference of 1.1 out of a raw scale range of 0–20 points yields a 4.6 point difference on a 0–100 scale). As mentioned above, DIF20 disappeared, when excluding students, but DIF19 remained and it had a magnitude above the typical MID.

Discussion

The purpose of the present study was to test for differential item functioning according to thyroid diagnosis, gender, age and education of the ThyPRO questionnaire, using ordinal logistic regression with both statistical significance and effect size criteria after iterative scale purification.
Table 3 Differential item functioning (DIF) according to age

<table>
<thead>
<tr>
<th>Abbreviated item wording</th>
<th>Significance of DIF</th>
<th>Variance explained by DIF (%)</th>
<th>Direction of DIF</th>
<th>DIF reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiredness scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full of life</td>
<td>&lt;0.0001</td>
<td>2.7</td>
<td>Younger patients more full of life</td>
<td>DIF9</td>
</tr>
<tr>
<td>Energetic*</td>
<td>&lt;0.0001</td>
<td>3.5</td>
<td>Younger patients more energetic</td>
<td>DIF10</td>
</tr>
<tr>
<td>Able to cope with life</td>
<td>&lt;0.0001</td>
<td>3.8</td>
<td>Younger patients more able to cope with life</td>
<td>DIF11</td>
</tr>
<tr>
<td>Depressivity scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crying easily</td>
<td>&lt;0.0001</td>
<td>2.6</td>
<td>Younger patients cry more easily</td>
<td>DIF12</td>
</tr>
<tr>
<td>Happy</td>
<td>&lt;0.0001</td>
<td>2.1</td>
<td>Younger patients more happy</td>
<td>DIF13</td>
</tr>
<tr>
<td>Self-confident*</td>
<td>&lt;0.0001</td>
<td>2.1</td>
<td>Younger patients more self-confident</td>
<td>DIF14</td>
</tr>
<tr>
<td>Impaired Social Life scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult being with other people</td>
<td>0.0033</td>
<td>3.4</td>
<td>Younger patients less difficulty being with other people</td>
<td>DIF15</td>
</tr>
<tr>
<td>A burden to other people</td>
<td>0.0020</td>
<td>2.3</td>
<td>Younger patients less a burden to other people</td>
<td>DIF16</td>
</tr>
</tbody>
</table>

Abbreviated wording of item and its associated scale is presented, along with the p value from the 2 test for DIF (uniform or uniform plus nonuniform, in case of nonuniform DIF) and the increase in explained variance (i.e., increment in $R^2$), when adding the DIF term to the regression equation. Finally, the direction of the DIF is indicated, and a reference for use in the text and figures is assigned (DIF reference). Nonuniform DIF marked with *

A total of 17 of the 84 items showed DIF. Eight DIFs according to diagnosis were found, eight according to age, one to gender and three according to education.

Most DIFs were small, judged by the effect size measures ($\Delta R^2 = 0.02–0.04$), but two items had larger DIF: 12% of additional variance in the item ‘Visible swelling in front of neck’ and 5.3% for ‘Sensation of fullness in the neck’ (both from the Goiter Symptoms scale) were explained when adding diagnosis to the regression equation. As illustrated above, the DIF regarding ‘visible swelling in front of neck’ only had minor impact on the 11 item Goiter Symptoms scale (it corresponds to a difference of 2.3 on the 0–100 scale). However, if it was part of an abbreviated scale of only three five-point items, such DIF would lead to a biased difference of 1 * 100/12 = 8.3 points on a 0–100 scale (assuming the other two items in such a scale is homogeneous).

Two general patterns of DIF were observed: (1) Only symptom scales were affected by DIF according to diagnosis, and (2) positively worded items were endorsed more readily by younger patients compared to older. Gender DIF was observed only for the item about crying. There was no obvious pattern in the DIFs according to education.

The fact that some items in the symptom scales relate slightly different to the overall scale in some patient groups compared to others seems intuitively reasonable. For the majority of items within the Goiter Symptoms scale displaying DIF, i.e. ‘Visible swelling in front of neck’ ($\Delta R^2 = 0.120$), ‘Sensation of fullness in the neck’ ($\Delta R^2 = 0.053$) and ‘Pressure in throat’ ($\Delta R^2 = 0.022$), this was in terms of patients with nodular goiter having greater likelihood of endorsing the items for a given level of goiter symptoms than patients with the autoimmune diseases (Graves’ hyperthyroidism, Graves’ orbitopathy and autoimmune hypothyroidism). An explanation for this could be that the sensation of the goiter by patients with nodular goiter (i.e., a swelling of the thyroid without accompanying autoimmunity) is different from the sensation of neck symptoms experienced by patients with thyroid autoimmunity. The symptoms in patients with a nodular goiter may be more directly related to the increased size of the goiter (visible swelling, sensation of fullness and pressure), whereas patients with autoimmune diseases experience more diffuse, irritative symptoms (bourseness, pain, globulus, swallowing problems and the need to clear the throat more often).

Also the DIF according to diagnosis regarding the Anxiety scale is clinically interpretable. Patients with nontoxic goiter have a higher probability of a positive response to the item concerning fear of being seriously ill, conditioned on their level of anxiety, than the other patient groups. It is indeed often such a concern that leads these patients to their physician: they discover a swelling of their neck, and fear that it is cancer. So, it is a relevant concern, independent of their level of anxiety in general, in contrast to, e.g. patients with hyperthyroidism, where anxiety as an affective state is part of the clinical picture.

The findings regarding DIF according to age in the positively worded items raise the question as to whether this is due to a method factor, to which younger patients are more...
sensitive, or whether positive well-being and energy, rather than representing one end of a continuum, as conceptualized within the ThyPRO, is in fact a separate dimension, i.e. 'vitality' in contrast to tiredness. As mentioned, previous factor analyses supported a one-factor solution for this scale, yielding the former explanation the most likely: younger patients are more influenced by a 'positive wording'-method factor, making them more likely to report positively on positively worded items than older patients are, provided the same level of tiredness/depressivity.

Although the concept of DIF concerns group comparisons, and thus is cross-sectional in nature, it is a limitation that the clinical significance of the DIFs found here cannot be tested in a longitudinal setting, since we only have cross-sectional datasets of a magnitude allowing for DIF analyses [39]. If longitudinal data from patients undergoing relevant changes were available, the influence of DIF on the responsiveness to these changes could be tested. Further, it is possible that some items are prone to DIF across time, i.e. that an item functions differentially at baseline compared to follow-up, during a particular disease/treatment course, which would also require longitudinal data [40].

The overall sample size is fairly large for DIF analyses using ordinal logistic regression. All subgroups had acceptable sample sizes (well above 100 and most around 150, for all groups), except for patients with Graves' orbitopathy (the rarest of the examined diagnoses, $n = 91$) and students ($n = 12$) [22, 41], although one simulation study has suggested a minimum of $n = 200$ in each group [42]. However, those analyses did not include any purification, which has been shown to increase detection rates [43].

### Fig. 3 Items with nonuniform DIF

For each item with nonuniform DIF, the mean expected item response, according to the logistic regression, for each group, for each level of the corresponding scale, is depicted. DIF10–DIF20 refer to the designations in Tables 3 and 4.

**Item score (0-4)**

<table>
<thead>
<tr>
<th>DIF according to age</th>
<th>DIF according to education</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIF10: Energetic</td>
<td>DIF19: Difficult being with other people</td>
</tr>
<tr>
<td>DIF14: Self-confident</td>
<td>DIF20: Disease affect appearance</td>
</tr>
</tbody>
</table>

**Scale**

<table>
<thead>
<tr>
<th>Tiredness scale</th>
<th>Impaired Social Life scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressivity scale</td>
<td>Cosmetic Complaints scale</td>
</tr>
</tbody>
</table>

**Min**

- Age 40 years
- Age 52 years
- Age 62 years

**Max**

- No education
- Studying
- Apprenticeship
- Short theoretical (1-3 years)
- Medium theoretical (3-5 years)
- Long theoretical (5+ years)
Previous studies have used different levels of cutoff for the effect size (i.e., $\Delta R^2$), typically higher than the one used here. Applying a higher cutoff, e.g. 3.5 %, as previously suggested [25], would reduce the number of instances of DIFs to only six. However, simulation studies have indicated that a level of 3.5 % leads to relatively high levels of Type II errors, and substantial instances of DIF may thus be missed [29, 30]. Moreover, the fact that a well-known DIF, as the gender/crying DIF [44–47] is identified only if the 2 % level is maintained, corroborates the appropriateness of this level.

Another limitation is the fact that analyses were only performed in one sample; ideally, they should be re-tested in an independent sample. However, this is somewhat compensated for by the fact that the sample spans a broad range of thyroid diagnoses, allowing for comparison across all the relevant diagnoses.

For future directions, the following recommendations can be based on the present analyses: All items with DIF should be flagged as possible candidates for item reduction in any forthcoming abbreviated versions of the ThyPRO. Of the items flagged, both in previous latent variable models and in the present analyses, two are strong candidates for reporting as single items or for item reduction: ‘Visible swelling in front of neck’ and ‘Hoarseness’ from the Goiter Symptoms scale. For the other DIFs according to diagnosis, a plausible and clinically meaningful interpretation could be found, and potential impact is only relevant when direct comparisons are important across various diagnostic groups. We cannot think of instances where this will be the main focus, so they are only weak candidates for reduction. Particularly, they may very well be items that differentiate well between patients. This should, as a minimum, be evaluated in longitudinal clinical studies before firm conclusions regarding their fate in a reduced scale structure can be ascertained. Regarding DIF according to gender, the item regarding tendency to cry is a strong candidate for removal from the Depressivity scale; this item also displayed DIF according to education.

In conclusion, few instances of DIF were identified. Most of these were due to clinically interpretable differences among various diagnostic subgroups and thus only potentially influential when comparisons across sub-diagnosis are important as such. The potential impact on scale score was low, but may be larger if abbreviated scales are constructed.

Acknowledgments This study has been supported by grants from the Danish Agency for Science, Technology and Innovation: Council for Strategic Research and Council for Independent Research. LH is supported by an unrestricted research grant from the Novo Nordisk Foundation.
Conflict of interest None of the authors have any financial conflicts of interest to declare. The ThyPRO was developed by the research team authoring this paper.

References


33. Martin, M., Blaisdell, B., Kwong, J. W., & Bjorner, J. B. (2004). The Short-Form Headache Impact Test (HIT) was...
psychometrically equivalent in nine languages. Journal of Clinical Epidemiology, 57(12), 1271–1278.


45. Yang, F. M., & Jones, R. N. (2007). Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. Journal of Clinical Epidemiology, 60(11), 1195–1200.


Paper #5
Cross-cultural validity of the thyroid-specific quality-of-life patient-reported outcome measure, ThyPRO

Torquil Watt · Giuseppe Barbesino · Jakob Bue Bjorner · Steen Joop Bonnema · Branka Bukvic · Russell Drummond · Mogens Groenvold · Laszlo Hegedüs · Valeska Kantzer · Kathryn E. Lasch · Claudio Marcocci · Anjali Mishra · Romana Netea-Maier · Merel Ekker · Ivan Paunovic · Terence J. Quinn · Åse Krogh Rasmussen · Audrey Russell · Mayilvaganan Sabaretnam · Johannes Smit · Ove To¨ rring · Vladan Zivaljevic · Ulla Feldt-Rasmussen

Accepted: 15 August 2014 / Published online: 7 September 2014 © Springer International Publishing Switzerland 2014

Abstract

Background and purpose Thyroid diseases are common and often affect quality of life (QoL). No cross-culturally validated patient-reported outcome measuring thyroid-related QoL is available. The purpose of the present study was to test the cross-cultural validity of the newly developed thyroid-related patient-reported outcome ThyPRO, using tests for differential item functioning (DIF) according to language version.

Methods The ThyPRO consists of 85 items summarized in 13 multi-item scales and one single item. Scales cover physical and mental symptoms, well-being and function as well as social and daily function and cosmetic concerns. Translation applied standard forward–backward methodology with subsequent cognitive interviews and reviews. Responses (N = 1,810) to the ThyPRO were collected in seven countries: UK (n = 166), The Netherlands (n = 147), Serbia (n = 150), Italy (n = 110), India (n = 148), Denmark (n = 902) and Sweden (n = 187). Translated versions were compared pairwise to the English version by examining uniform and nonuniform DIF, i.e., whether patients from different countries respond differently to a...
particular item, although they have identical level of the concept measured by the item. Analyses were controlled for thyroid diagnosis. DIF was investigated by ordinal logistic regression, testing for both statistical significance and magnitude \((\Delta R^2 > 0.02)\). Scale level was estimated by the sum score, after purification.

**Results** For twelve of the 84 tested items, DIF was identified in more than one language. Eight of these were small, but four were indicative of possible low translatability. Twenty-one instances of DIF in single languages were identified, indicating potential problems with the particular translation. However, only seven were of a magnitude which could affect scale scores, most of which could be explained by sample differences not controlled for.

**Conclusion** The ThyPRO has good cross-cultural validity with only minor cross-cultural invariance and is recommended for use in international multicenter studies.

**Keywords** Thyroid diseases · Quality of life · Patient-reported outcome measure · Cross-cultural validity · Differential item functioning

**Introduction**

Patients with thyroid diseases may experience impact of their disease on quality of life (QoL) [1–4], i.e., on physical, psychological and social functioning and well-being [5]. Measurement of QoL [here synonymous with patient-reported outcome measures (PRO)] is becoming routine in clinical research outcome assessments, and it is generally recommended to apply both a generic and a disease-specific PRO measure. Despite the fact that thyroid diseases are common, thyroid-specific PROs have only recently been developed. Some are only applicable to specific thyroid diseases such as hypothyroidism [6] or Graves’ orbitopathy [7]. We have developed a thyroid disease-specific PRO questionnaire (ThyPRO), which measures QoL across the entire range of benign thyroid diseases [8–11]. To date, the ThyPRO has only been quantitatively validated in Danish. As there is an urgent and unmet need for measuring disease-specific QoL in thyroidology internationally, the ThyPRO has become available in 13 linguistically and qualitatively validated language versions. This cross-cultural development procedure has followed scientifically accepted standards [12, 13].

However, in order to compare results across languages and cultures, it is important that PRO measures have also been cross-culturally validated quantitatively [14–16], since the qualitative translation and equilibrating process may not identify item wordings conveying different meaning in various cultures.

An efficient method for testing cross-cultural validity is analysis of differential item functioning (DIF), using ordinal logistic regression [17–21]. An item in a multi-item scale is subject to DIF if different groups of respondents with the same level of the measured attribute score differently on that item [22–27].

The purpose of the present study was to evaluate cross-cultural validity of the thyroid-specific QoL PRO measure, ThyPRO, by testing for differential item functioning according to language.

**Methods**

**Patient populations**

This study compiles data from 7 samples in 7 languages: English (Scotland and Ireland), Dutch, Serbian, Hindi (India), Italian, Danish [10] and Swedish. The clinical and sociodemographic characteristics of the samples are summarized in Table 1.

In Denmark, the Netherlands, UK and Italy, patients referred to the participating university hospital endocrinological outpatient clinics for treatment of benign thyroid diseases, as well as patient attending the clinics for monitoring of their treatment, completed the ThyPRO questionnaire and had blood samples drawn a few days prior to or as part of their scheduled visit to the clinics.

In India and Serbia, patients referred to the endocrine surgery department for surgical treatment of their benign thyroid diseases were included. ThyPRO was completed, and blood samples were drawn a few days prior to thyroidectomy.

In Sweden, responses to the ThyPRO were collected as part of a follow-up study among patients treated 7–9 years earlier for hyperthyroidism.

In all samples, data regarding diagnoses and treatments were obtained from medical records.

**Multicultural development, translation procedure and qualitative validation**

The ThyPRO is an 85-item thyroid-specific patient-reported outcome measuring QoL with 13 scales [8–11] covering physical (4 scales) and mental (2 scales) symptoms, function and well-being (3 scales) and participation/social function (4 scales) (cf. Appendix A and B) and a single item about overall QoL. Items are rated on a five-point scale from 0 = not at all to 4 = very much, with a reference period of 4 weeks. The average score of items in a
scale is divided by four (maximum item response) and multiplied by 100 to yield thirteen 0–100 scales. Questionnaire content (i.e., issues of relevance for thyroid patients) was generated from a literature review [1] and interviews with Danish experienced clinicians and 80 patients [8] within a classical health-related QoL theoretical framework [28–31]. Items were constructed in Danish and immediately thereafter translated using standard forward–backward translation (by translators native in target language and fluent in source language) and subsequent cognitive debriefing to English [12]. Subsequently, in-depth interviews were initiated in Danish (by TW) and shortly thereafter also in English (by KEL). The interviews were conducted in iterative rounds (in a total of 48 patients

<table>
<thead>
<tr>
<th>Table 1 Clinical and sociodemographic characteristics of the seven samples</th>
<th>English</th>
<th>Dutch</th>
<th>Serbian</th>
<th>Italian</th>
<th>Indian</th>
<th>Danish</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>166</td>
<td>147</td>
<td>150</td>
<td>110</td>
<td>148</td>
<td>902</td>
<td>187</td>
</tr>
<tr>
<td>Gender [n (%)]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>142 (86)</td>
<td>122 (83)</td>
<td>132 (88)</td>
<td>97 (88)</td>
<td>116 (78)</td>
<td>784 (87)</td>
<td>158 (84)</td>
</tr>
<tr>
<td>Men</td>
<td>24 (14)</td>
<td>25 (17)</td>
<td>18 (12)</td>
<td>13 (12)</td>
<td>32 (22)</td>
<td>118 (13)</td>
<td>29 (16)</td>
</tr>
<tr>
<td>Age in years [median (Q1–Q3)]</td>
<td>49 (40–60)</td>
<td>48 (39–59)</td>
<td>52 (41–61)</td>
<td>50 (38–58)</td>
<td>40 (33–50)</td>
<td>51 (40–61)</td>
<td>50 (39–61)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only primary school (0–10 years)</td>
<td>6 (4)</td>
<td>15 (10)</td>
<td>29 (19)</td>
<td>8 (7)</td>
<td>30 (20)</td>
<td>151 (17)</td>
<td></td>
</tr>
<tr>
<td>High school/vocational (11–13 years)</td>
<td>18 (11)</td>
<td>53 (36)</td>
<td>75 (50)</td>
<td>17 (15)</td>
<td>29 (20)</td>
<td>110 (12)</td>
<td></td>
</tr>
<tr>
<td>Upper secondary/higher educ. (14–16 years)</td>
<td>48 (29)</td>
<td>44 (30)</td>
<td>19 (13)</td>
<td>36 (33)</td>
<td>35 (24)</td>
<td>261 (29)</td>
<td></td>
</tr>
<tr>
<td>High educ., e.g., University (&gt;16 years)</td>
<td>30 (18)</td>
<td>32 (22)</td>
<td>27 (18)</td>
<td>45 (41)</td>
<td>20 (14)</td>
<td>346 (38)</td>
<td></td>
</tr>
<tr>
<td>Information not available</td>
<td>64 (39)</td>
<td>3 (2)</td>
<td>0</td>
<td>4 (4)</td>
<td>34 (23)</td>
<td>34 (4)</td>
<td>187 (100)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nontoxic goiter</td>
<td>14 (8)</td>
<td>10 (7)</td>
<td>53 (46)</td>
<td>46 (42)</td>
<td>81 (55)</td>
<td>259 (29)</td>
<td>0</td>
</tr>
<tr>
<td>Toxic nodular goiter</td>
<td>17 (10)</td>
<td>5 (3)</td>
<td>17 (13)</td>
<td>13 (12)</td>
<td>14 (9)</td>
<td>145 (16)</td>
<td>40 (21)</td>
</tr>
<tr>
<td>Graves hyperthyroidism</td>
<td>78 (47)</td>
<td>39 (27)</td>
<td>15 (15)</td>
<td>15 (14)</td>
<td>39 (26)</td>
<td>166 (18)</td>
<td>112 (60)</td>
</tr>
<tr>
<td>Graves orbitopathy</td>
<td>11 (7)</td>
<td>3 (2)</td>
<td>3 (9)</td>
<td>9 (8)</td>
<td>14 (9)</td>
<td>94 (10)</td>
<td>32 (17)</td>
</tr>
<tr>
<td>Autoimmune hypothyroidism</td>
<td>31 (19)</td>
<td>77 (52)</td>
<td>10 (19)</td>
<td>19 (17)</td>
<td>0</td>
<td>198 (22)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Other thyroid diagnosis</td>
<td>15 (9)</td>
<td>12 (8)</td>
<td>1 (8)</td>
<td>8 (7)</td>
<td>0</td>
<td>40 (4)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Disease duration in months [median (Q1–Q3)]</td>
<td>10 (0–42) [3]</td>
<td>90 (42–174) [8]</td>
<td>60 (18–156)</td>
<td>84</td>
<td>24 (10–70)</td>
<td>27 (4–79)</td>
<td>91</td>
</tr>
<tr>
<td>[info not avail.]]</td>
<td></td>
<td></td>
<td></td>
<td>[24–156]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbidity [n (%)]</td>
<td>70 (42) [3]</td>
<td>72 (49) [0]</td>
<td>150</td>
<td>12 (11) [7]</td>
<td>148</td>
<td>505 (56) [0]</td>
<td>61 (33)</td>
</tr>
<tr>
<td>[info not available]]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current treatment: [n (%)]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-thyroxine</td>
<td>57 (34)</td>
<td>125 (85)</td>
<td>27 (18)</td>
<td>47 (43)</td>
<td>0</td>
<td>288 (32)</td>
<td>116 (62)</td>
</tr>
<tr>
<td>Anti-thyroid drug</td>
<td>60 (36)</td>
<td>12 (8)</td>
<td>37 (25)</td>
<td>9 (8)</td>
<td>67 (45)</td>
<td>162 (18)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>6 (4)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (0)</td>
<td>19 (10)</td>
</tr>
<tr>
<td>Previous treatment: [n (%)]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radioiodine</td>
<td>10 (6)</td>
<td>27 (18)</td>
<td>0</td>
<td>13 (12)</td>
<td>0</td>
<td>114 (13)</td>
<td>87 (47)</td>
</tr>
<tr>
<td>Thyroid surgery [(hemi-)</td>
<td>2 (1)</td>
<td>16 (11)</td>
<td>24 (16)</td>
<td>16 (15)</td>
<td>0</td>
<td>129 (14)</td>
<td>35 (19)</td>
</tr>
<tr>
<td>thyroidectomy]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thyroid function tests: [median (Q1–Q3)] [n info not avail.]]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSH mIU/L (value 0 means below detection level)</td>
<td>0.95 [0.03–2.55] [2]</td>
<td>0.89 [0.26–1.97] [8]</td>
<td>0.97 [0.39–1.99] [0]</td>
<td>1 [0.32–2.5] [7]</td>
<td>0.68 [0.1–2.1] [9]</td>
<td>1.12 [0.36–2.7] [16]</td>
<td>0 (0–0) a [8]</td>
</tr>
</tbody>
</table>

* Measured at time of diagnosis
with thyroid diseases), and results from previous interviews in both languages were subsequently incorporated into both language versions. The final English version was translated (forward–backward) into Dutch, Italian, German, Portuguese and French. In each language, at least five patients were interviewed in depth using cognitive interviewing. Issues arising during this procedure were addressed at a harmonization meeting, and resulting changes in a language version were subsequently tested in new cognitive interviews. Issues addressed were mainly concerning strengths of terms at various languages (e.g., the degree of tiredness associated with a specific term describing tiredness) or conceptual breadth of terms in different languages.

After the initial parallel development process described above, the ThyPRO has been translated into Serbian, Swedish, Hebrew, Hindi and Tamil. For these, the method described above was applied, except for the fact that for the Danish and English master versions, there were no further revisions; all issues noted during the translation process should be addressed in the new language version.

In a recent study, the ThyPRO scales were found to be uni-dimensional, according to structural equation modeling analyses [32].

Statistical analyses

The English version was used as the reference for all other versions, in the DIF analyses. DIF was investigated using ordinal logistic regression [33]. For each item, three regression models were compared:

\[
\ln \left( \frac{P(Y \leq j)}{P(Y > j)} \right) = \beta_{0j} + \beta_1 \times \text{Scale} \quad (1)
\]

\[
\ln \left( \frac{P(Y \leq j)}{P(Y > j)} \right) = \beta_{0j} + \beta_1 \times \text{Scale} + \beta_2 \times \text{Language} \quad (2)
\]

\[
\ln \left( \frac{P(Y \leq j)}{P(Y > j)} \right) = \beta_{0j} + \beta_1 \times \text{Scale} + \beta_2 \times \text{Language} + \beta_3 \times (\text{Language} \times \text{Scale}) \quad (3)
\]

where \( Y \) is the response option and \( j = 0, 1, 2, \ldots, c-1 \), where \( c \) is the number of response categories in the item. All models were controlled for diagnosis by adding it to the regression equation (but are left out of the presented equations above, for simplicity).

In Model (1), item responses were regressed on the scale only, i.e., all variation in the item was explained by the scale. In Model (2), language was added as a covariate, and in Model (3), the interaction term between language and the scale (representing nonuniform DIF) was also entered into the regression equation. Significance of DIF was tested by comparing 2log likelihood values for the nested models: An overall test of DIF was obtained by comparing Model 3 with Model 1 (difference test). This yields a \( \chi^2 \) statistic with degrees of freedom equal to 2 * (number of groups minus one), i.e., \( 2df \) for the classical two groups comparisons or continuous variables. If a similar comparison of Model 2 and Model 3 (degrees of freedom equal to number of groups minus one, i.e., 1df for the classical two-group comparison) is nonsignificant, there is evidence of uniform DIF, i.e., the same “degree” of DIF irrespective of scale level, whereas if it is significant, there is evidence of nonuniform DIF, i.e., the magnitude of the DIF differs among the language samples. Significance level was set to \( p < 0.05 \) with a Bonferroni correction for multiple testing [34, 35] within each scale.

In addition to statistical significance, magnitude of DIF (i.e., effect size) was also considered, before flagging an item as evidencing DIF [33, 36]. The magnitude measure was obtained by comparing the three models in terms of \( R^2 \) (i.e., Nagelkerke “Pseudo \( R^2 \)” [37, 38]), i.e., “variance explained” by each model. Magnitude of uniform DIF was obtained by subtracting \( R^2 \) for Model 1 from \( R^2 \) for Model 2 and likewise for nonuniform DIF (Model 3 minus Model 1). In order for possible DIF to be judged substantial, the increase in \( R^2 \) had to be more than 0.02. In other words, the DIF should explain more than 2 % of the variance in the item score. This is lower than that suggested by other authors (e.g., 3.5 %) [18, 33, 36], but such higher thresholds may lead to high type II error rates [39, 40], and in our experience, 2 % is a relevant level [19, 38, 41–43].

The scale level was estimated by the sum score, after positively worded items had been reverted. The sum score included also the item being tested for DIF [33, 44, 45], but items with DIF were excluded, i.e., the sum score was purified [33, 46]. Purification was obtained by iterative DIF tests with subsequent exclusion of items displaying DIF.

SAS PROC LOGISTIC [47] was used for the ordinal logistic regression analyses, which were performed as complete cases analyses.

Impact of DIF on scale score was evaluated by plotting the expected item responses for each level of scale score and language, as estimated by the regression model, against the level of the respective purified scales [48]. The expected item response was calculated as

\[
\sum_{j=1}^{c} P(j) \times j, \text{ where}
\]

\[
P(j) = \frac{e^{\text{linpred}_j}}{1 + e^{\text{linpred}_j}} - \frac{e^{\text{linpred}_{j-1}}}{1 + e^{\text{linpred}_{j-1}}}, \text{ and}
\]

\[
\text{linpred}_j = \beta_{0j} + \beta_1 \times \text{Scale} + \beta_2 \times \text{Language} + \beta_3 \times (\text{Language} \times \text{Scale})
\]

In case of uniform DIF, \( \beta_3 \) was set to zero.
All instances of DIF were reviewed by language and clinical experts in order to identify possible linguistic, cultural or clinical explanations for the DIF. In general, DIF for an item in only one country was considered a cross-cultural or translational issue for that particular translation, whereas DIF in more than one language was considered an indication of a translatability or cross-cultural issue for the involved item.

Results

Clinical and sociodemographic characteristics are presented in Table 1. The samples had similar age, gender and education distributions. The English-speaking sample had shorter disease duration than the other samples. Thyroid disease subcategories differed among countries. Only few patients with nontoxic goiter were sampled in Scotland/Ireland, The Netherlands and Sweden. A high proportion of patients with Graves’ disease were included in UK and Sweden, in comparison with low proportions in Serbia, Italy and Denmark. In the Dutch sample, the majority of patients had autoimmune hypothyroidism, particularly compared to India, wherefrom no patients with hypothyroidism were sampled. Despite these dissimilarities in sample compositions, the thyroid function tests were comparable, except for the Swedish sample, but this was due to the fact that thyroid function tests from the Swedish sample were reported at time of diagnosis, not at time of questionnaire completion.

Mean level of the 13 ThyPRO scales according to language and diagnosis is presented in Table 2. The English-speaking sample had higher scores in comparison with other languages in most scales. Least differences were observed with Hindi followed by the Serbian sample. When comparing score levels according to diagnosis, a number of differences were also found. As shown in Table 2, scores in the four physical symptoms scales were comparable, except for the Swedish sample, but this was due to the fact that thyroid function tests from the Swedish sample were reported at time of diagnosis, not at time of questionnaire completion.

Mean level of the 13 ThyPRO scales according to language and diagnosis is presented in Table 2. The English-speaking sample had higher scores in comparison with other languages in most scales. Least differences were observed with Hindi followed by the Serbian sample. When comparing score levels according to diagnosis, a number of differences were also found. As shown in Table 2, scores in the four physical symptoms scales were comparable, except for the Swedish sample, but this was due to the fact that thyroid function tests from the Swedish sample were reported at time of diagnosis, not at time of questionnaire completion.

Mean level of the 13 ThyPRO scales according to language and diagnosis is presented in Table 2. The English-speaking sample had higher scores in comparison with other languages in most scales. Least differences were observed with Hindi followed by the Serbian sample. When comparing score levels according to diagnosis, a number of differences were also found. As shown in Table 2, scores in the four physical symptoms scales were comparable, except for the Swedish sample, but this was due to the fact that thyroid function tests from the Swedish sample were reported at time of diagnosis, not at time of questionnaire completion.

Mean level of the 13 ThyPRO scales according to language and diagnosis is presented in Table 2. The English-speaking sample had higher scores in comparison with other languages in most scales. Least differences were observed with Hindi followed by the Serbian sample. When comparing score levels according to diagnosis, a number of differences were also found. As shown in Table 2, scores in the four physical symptoms scales were comparable, except for the Swedish sample, but this was due to the fact that thyroid function tests from the Swedish sample were reported at time of diagnosis, not at time of questionnaire completion.

Discussion

The purpose of the present study was to evaluate the cross-cultural validity, in terms of differential item functioning across seven different languages, of the ThyPRO. Standard ordinal logistic regression analyses of DIF were applied. This is the first study to evaluate cross-cultural validity of a PRO measure among patient with benign thyroid diseases. The availability of a cross-culturally valid
thymoid-specific PRO measuring QoL across a broad span of thyroid diseases is a major advantage of this study.

Very high cross-cultural validity was found, as might also be expected due to the rigorous translation methodology applied. Hardly any difference was identified between the English, Danish, Dutch and Swedish versions of the questionnaire. Small differences were found in the Italian, Serbian and Indian versions, which were most likely caused by differences in thyroid subdiagnoses. Other differences may reflect cultural differences in conceptualizations of specific item content in relation to the concept measured by the scale. However, none of these instances of DIF led to substantial change in the summary scale score, as estimated by the expected mean scores modeling the DIFs.

Twelve items (i.e., those in Figs. 1, 2 and 3) had differential item functioning in more than one country, indicating possible low translatability or cross-cultural equality of these items. However, differential item functioning is only potentially harmful to the validity of a measurement, if it has an impact on the scale score. Although the DIFs identified in this study are above the explained variance cut-off level for substantiality, they have little impact on the ThyPRO scale scores. For example, for the four items with substantially diverging expected item mean score lines in Figs. 1, 2, 3, the estimated differences on the 0–100 point scale scores caused by the DIFs are as follows: “Visible swelling in front of neck”: 3.9 point, “Been confused”: 5.0 points, “Feeling not like yourself”: 2.6 points and “Felt too fat”: 6.1 points. All of these are below the typical level for minimally important differences in these types of scales of around 7 points, although the latter is of a magnitude that may impact conclusions in clinical studies. The reason for the limited effect on scale score is that the items are part of multi-item scales of at least four items. So for “Visible swelling in neck,” for example, the difference in mean item scores between the two extremes (Hindi and Swedish) is 3.2 minus 1.5, as illustrated in Fig. 1, i.e., 1.7 point scores between the two extremes (Hindi and Swedish).

The cross-cultural equivalence was higher than expected, since the study encompassed patients from as different cultures as, e.g., the Danish (the majority in this study) and the Indian. A possible reason for this may be the universality of the concepts measured, embedded in a unifying clinical context (thyroid disease). This was also reflected in the early, qualitative phases of the project, where the rephrasing of item content by patients was very much in line with the prespecified and rephrased content of the originals.
Table 3  Differential item functioning (DIF) in single languages

<table>
<thead>
<tr>
<th>Item</th>
<th>Serbian wording</th>
<th>Scale</th>
<th>Direction of DIF</th>
<th>Item</th>
<th>Italian wording</th>
<th>Scale</th>
<th>Direction of DIF</th>
<th>Item</th>
<th>Hindi (India) wording</th>
<th>Scale</th>
<th>Direction of DIF</th>
<th>Item</th>
<th>Swedish wording</th>
<th>Scale</th>
<th>Direction of DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>had difficulty swallowing?</td>
<td>Goiter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td>ošećao gutala?</td>
<td>Serbian more difficulty swallowing</td>
<td>$\Delta R^2$ 2.8</td>
<td>Non-linear relationship</td>
<td>3a</td>
<td>felt full of life?</td>
<td>Tired</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c</td>
<td>- ha avuto tremore alle mani?</td>
<td>Hyper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d</td>
<td>- ha avuto il fiato corto?</td>
<td>Non-linear relationship</td>
<td>$\Delta R^2$ 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1e</td>
<td>- has your thyroid disease affected your appearance?</td>
<td>Cosmetic comp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1f</td>
<td>had the sensation of a lump in your throat?</td>
<td>Goiter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1g</td>
<td>- haft klinian sv svt ha en klump i halsen?</td>
<td>$\Delta R^2$ 2.8</td>
<td>Swedish more sensation of lump</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11a</td>
<td>- haft orolig mage?</td>
<td>$\Delta R^2$ 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11b</td>
<td>- had an upset stomach?</td>
<td>Hyper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11c</td>
<td>- felt uneasy?</td>
<td>Anxiety</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11d</td>
<td>- kän dig orolig?</td>
<td>$\Delta R^2$ 5.0</td>
<td>Swedish more uneasy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11e</td>
<td>- felt less restful</td>
<td>Anxiety</td>
<td>$\Delta R^2$ 2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11f</td>
<td>- felt less happy</td>
<td>Depressivity</td>
<td>$\Delta R^2$ 2.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11g</td>
<td>- feel you were a burden to other people?</td>
<td>Social Impact</td>
<td>$\Delta R^2$ 2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11h</td>
<td>- have difficulty managing daily life?</td>
<td>Daily Impact</td>
<td>$\Delta R^2$ 7.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each instance of DIF which was only present in one of the six tested languages, the English reference wording, the tested language wording, direction of DIF, scale to which the item belongs and the magnitude of DIF, in terms of the variance explained ($\Delta R^2$ in percent), is presented. In case of nonlinearity, the direction of DIF is ambiguous and cannot be presented briefly. Item numbers refer to ThyPRO questionnaire in Appendix A.
A limitation of this study is the clinical sample dissimilarity, due to different sampling strategies, which is a consequence of practical issues within each participating site. Since different diagnoses have indeed previously been shown to cause DIF in the ThyPRO, albeit only to a limited extent, one might question whether the identified DIFs are consequences of the differences in sample distributions, rather than cross-cultural differences [49]. However, against such a concern speaks the fact that some of the samples which were clinically most dissimilar to the English-speaking sample (i.e., Dutch, Danish and Swedish) had virtually no DIF. Thus, the DIF identified in this study is likely to reflect cross-cultural and linguistic differences, rather than clinical differences.

Another concern could be that the determination of DIF based on its presence in more than one country may be overly simplistic and potentially misleading. However, if the present results are used as guidance to select items in a future shorter version, avoiding these items could optimize the translatable of such a shorter version.

These results could have implications at various levels. It could have implications for the level of comparability of studies conducted in different countries, in terms of analysis strategies for multi-language clinical studies (e.g., international multi-center clinical trials) and in terms of revised or abbreviated versions of the ThyPRO. Further, the results could document the extent to which it is justified to extrapolate validation studies conducted in one language to other languages. In terms of the former, the degree of DIF among these language versions is of such small magnitude that it does not hamper comparisons of studies conducted in different countries. The ThyPRO is being used in international multicenter studies, e.g., in evaluations of adjunctive treatment of nontoxic goiter [50], and also an ongoing EU funded RCT of L-thyroxine in sub-clinical hypothyroidism, as co-primary outcome. In such
studies, one should consider taking the four major instances of DIF into account in the analysis strategy, by conducting sensitivity analyses of the results when the scales are scored without the items displaying DIF. Alternatively, one could adjust the scale scores by the magnitude and direction of the DIF [48]. However, in clinical trials, this will seldom be a significant problem, since participants are usually stratified by center in the randomization procedure.

Perhaps the most important implication of these results is their potential to guide short-form versions of the ThyPRO applicable in clinical trials and possibly clinical practice. The current version of the ThyPRO is rather long, which is also part of the explanation for the limited impact of DIF on the scale scores. However, when short-forms are developed, since the impact of a particular DIF on a short scale will be larger, we would recommend that the items with DIF are flagged for avoidance in future cross-culturally applicable short-forms of the ThyPRO.

In conclusion, good cross-cultural validity of the ThyPRO was found. Twelve items had differential item functioning in more than one country, but the DIF had little impact on the ThyPRO scale scores. Small single-language differences were found in the Italian, Serbian and Indian versions, but they were most likely caused by differences in thyroid subdiagnoses. We recommend the ThyPRO implemented as an outcome in clinical studies, trials and also as monitoring and communication instrument in clinical practice across the various translations. We recommend application of standard rigorous translation procedures applying state-of-the-art mixed methods research principles, when translating instruments to new languages. Further, if future short-forms of ThyPRO are developed, we recommend such abbreviation approaches to use the present results as guidance for the item selection procedure, to increase translatability of a short-form into further languages.
Fig. 3 Differential item functioning (DIF) in the anxiety, depressivity and emotional susceptibility scales in two languages or more. For each item with DIF, the mean expected item score, according to the logistic regression model, for each group, along the continuum of the corresponding scale, is depicted (see Fig. 1 for an example). In addition, the variance explained by the DIF, $\Delta R^2$, is provided for each country, wherein the DIF was detected.

Acknowledgments This study has been supported by grants from the Danish Agency for Science, Technology and Innovation: Council for Strategic Research and Council for Independent Research and by Genzyme Corporation by an unrestricted research grant. Special thanks to Sabrina Sereni, Supriya Mathur, Kajsa Bergman, Gitte Karina Geil, Kim Øbelø, Selma Flora Watt and Laura Siim Magnussen for assistance with data entry and logistics.

Conflict of interest None of the authors have any financial conflict of interest to declare. The ThyPRO was developed by TW, UFR, AKR, JBB, MG, SB and LH.

References


Is Thyroid Autoimmunity per se a Determinant of Quality of Life in Patients with Autoimmune Hypothyroidism?

Torquil Watt\textsuperscript{a, b} Laszlo Heged\text{s}u\textsuperscript{e} Jakob Bue Bjorner\textsuperscript{b, d} Mogens Groenvold\textsuperscript{b, c} Steen Joop Bonnema\textsuperscript{a} Åse Krogh Rasmussen\textsuperscript{a} Ulla Feldt-Rasmussen\textsuperscript{a}

\textsuperscript{a}Department of Endocrinology, Copenhagen University Hospital Righospitalet, \textsuperscript{b}Institute of Public Health, University of Copenhagen, \textsuperscript{c}Department of Palliative Medicine, Bispebjerg Hospital, and \textsuperscript{d}National Research Centre for the Working Environment, Copenhagen, and \textsuperscript{e}Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark

---

**Key Words**

Autoimmune hypothyroidism \cdot Quality of life \cdot Clinical variables \cdot Autoimmunity

---

**Abstract**

**Purpose:** To evaluate the relationship between thyroid variables and health-related quality of life (QoL) in patients with autoimmune hypothyroidism, using the thyroid-specific QoL questionnaire ThyPRO. **Methods:** In a cross-sectional study, responses to the ThyPRO from 199 outpatients with autoimmune hypothyroidism were analyzed in relation to thyroid volume, thyroid function and markers of thyroid autoimmunity. Based on a classical QoL framework, we hypothesized that physiological dysfunction caused specific physical and psychological symptoms, which affected functioning and well-being, and consequently participation in life and QoL. These hypotheses were tested through multiple regression and multivariate path analysis models. **Results:** None of the thyroid function tests were associated with QoL scores. However, in the pairwise regression, the thyroid peroxidase antibody (TPOAb) level was associated with several QoL outcomes: Goitre Symptoms (p = 0.024), Depressivity (p = 0.004), Anxiety (p = 0.004), Emotional Susceptibility (p = 0.005) and Impaired Social Life (p = 0.047). In the multivariate model, the TPOAb level was related to Goitre Symptoms (r = 0.17, p = 0.019), Depressivity (r = 0.24, p = 0.001), and Anxiety (r = 0.23, p = 0.002), but no longer to Emotional Susceptibility or Impaired Social Life, indicating that the effect on these were mediated through an effect on the symptom scales (i.e. Goitre Symptoms, Depressivity and Anxiety). **Conclusion:** Health-related QoL, evaluated with state-of-the-art QoL methodology, was related to TPOAb level but not to thyroid function. This raises the hypothesis that autoimmunity, independent of thyroid function, impacts on QoL in patients with autoimmune hypothyroidism, especially in terms of psychological symptoms. Longitudinal studies, in initially untreated patients, are needed to test this hypothesis.

---

**Introduction**

Autoimmune hypothyroidism is characterized by lymphocytic infiltration of the thyroid gland and most commonly various degrees of hypofunctioning of the gland, in the presence of thyroid autoantibodies against thyroid peroxidase (TPOAb) and/or thyroglobulin...
Table 1. Clinical and sociodemographic characteristics of the 199 patients with autoimmune hypothyroidism

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women/men, n</td>
<td>184 (92%)/15</td>
</tr>
<tr>
<td>Age, years</td>
<td>44 (19–88)</td>
</tr>
<tr>
<td>Months since diagnosis</td>
<td>21 (0.7 to 307)</td>
</tr>
<tr>
<td>Diagnosed with mild hypothyroidism</td>
<td>69 (36)</td>
</tr>
<tr>
<td>Currently treated with L-thyroxine</td>
<td>152 (77%)</td>
</tr>
<tr>
<td>Thyroid volume, ml</td>
<td>9 (0–55)</td>
</tr>
<tr>
<td>TSH, mIU/l</td>
<td>3.1 (&lt;0.01–54.8)</td>
</tr>
<tr>
<td>Free T₄, pmol/l</td>
<td>16 (0.2–163)</td>
</tr>
<tr>
<td>Free T₃, pmol/l</td>
<td>4.8 (1.6–7.6)</td>
</tr>
<tr>
<td>TPOAb, U/l</td>
<td>4,266 (0–21,952)</td>
</tr>
<tr>
<td>TSHRAb, U/l</td>
<td>0 (0–7)</td>
</tr>
</tbody>
</table>

Data presented as numbers and percentages and medians and ranges. TSHRAb = TSH receptor antibody levels.

Negative duration due to a patient completing the questionnaire before a final diagnosis was reached.

Serum TSH above reference range and peripheral thyroid hormones within reference range (also termed subclinical hypothyroidism).

Data for determination of mild or overt hypothyroidism at time of diagnosis (i.e., up to several years prior to questionnaire completion) were unavailable for 8 patients.

The condition is associated with a wide range of unspecific symptoms.

In recent years, a change of paradigm has occurred in the treatment of many chronic diseases. Restoring physiological imbalances is no longer seen as an aim in itself, but rather as a means towards prolonging life, relieving symptoms and improving function. Thus, evaluation of such outcomes is central in chronic-disease treatment evaluation [1]. While the paradigm shift has come late to thyroid diseases, this area represents obvious beneficial applications. Thyroid diseases such as autoimmune hypothyroidism are common, chronic, and affect the patients’ function and quality of life (QoL) [2, 3]. Despite this high prevalence, few systematic studies have evaluated the impact of autoimmune hypothyroidism on QoL and until recently, no validated measure of QoL for thyroid patients was available [4].

QoL is a complex and multidimensional construct. Studies of QoL in patient populations should acknowledge this, by incorporating analyses of QoL into a theoretical and statistical model reflecting this complexity [5]. In order for such a model to reflect the disease in focus, the theoretical model should incorporate clinical variables and the hypothesized relationship between these and the QoL framework [5, 6]. The advent of a comprehensive, thoroughly validated thyroid-specific patient-reported outcome measuring health-related QoL for patients with thyroid diseases, i.e., ThyPRO [7–10], calls for studies evaluating these relationships between clinical variables and QoL.

The purpose of the present study was therefore to evaluate the relationship between clinical measures of disease activity and health-related QoL in patients with autoimmune hypothyroidism using the thyroid-specific QoL questionnaire ThyPRO and subjecting data to multivariate analysis within a theoretical QoL framework.

Methods

Patients and Procedures

During February to November 2007, 199 patients with autoimmune hypothyroidism (defined as any degree of serum thyroid-stimulating hormone (TSH) above the reference range at two consecutive measurements, with or without associated thyroid hormone levels below the reference range, and TPOAb level >60 IU/l), attending the endocrinological outpatient clinics at Copenhagen University Hospital Rigshospitalet and Odense University Hospital were recruited (Table 1). None of the patients had eye involvement. Recruitment has previously been described in detail [9] and the overall response rate was 69%. Questionnaires were completed and blood samples were drawn a few days prior to the scheduled visit to the clinic. A physical examination, including thyroid ultrasound, was performed and written, signed informed consent obtained.

Sociodemographic data and information about comorbidity and nonthyroid medication were self-reported. Data regarding clinical biochemical measurements and thyroid imaging, exact diagnosis, previous and current treatment and time of diagnosis were obtained by chart review. Biochemical thyroid tests, using in-house routine assays, were TSH, total thyroxine (T₄), total triiodothyronine (T₃), non-protein-bound thyroxine (FT₄), non-protein-bound triiodothyronine (FT₃) and TPOAb. Thyroid volume was determined by ultrasound using the ellipsoid method [11]. The project was approved by the local ethical committee (KF01 2006-1579) and the Danish data protection agency and registered at ClinicalTrials.gov (NCT00150033).

Patient-Reported Outcome, ThyPRO

The ThyPRO is an 84-item thyroid-specific patient-reported outcome measuring QoL with 13 scales covering physical and mental symptoms, functioning and well-being as well as impaired participation in important life activities. The content was derived by literature review and in-depth interviews with experts and patients [7]; the instrument has been validated for thyroid patients [8–10]. Each item is rated by the patient on a five-point Likert scale, and the 13 scales are derived by averaging and linearly transforming these item scores into their respective 0–100 scale score. Figure 1 is a representation of these scales, organized in a conceptual model linking health-related QoL and clinical variables [6, 12] for patients with any benign thyroid disease (see below).
**Theoretical QoL Model**

The theoretical QoL model, incorporating the hypothesized relationships, is presented in figure 1: thyroid volume is related to goitre symptoms; low serum TSH, high serum fT4, and high serum fT3 cause hyperthyroid symptoms; high serum TSH, low serum fT4 and low serum fT3 cause hypothyroid symptoms; Clinical Activity Score (CAS) [13] and NOSPECS, an orbitopathy grading system [14], are associated with eye symptoms (not relevant in this sample), and TPOAb and TSHRAb could be associated with all the symptom scales. Further, physical and psychological symptoms are associated with reduced function and well-being and these reductions again affect social life, daily life, sex life and body image.

The segregation of the thyroid function into higher versus lower function is due to the fact that a linear, or just incremental, relationship between thyroid function and QoL cannot be expected. This relationship must be U-shaped with e.g. decreasing serum TSH below the reference range leading to decreasing QoL due to hyperthyroidism whereas increasing TSH above the reference range also leads to decreasing, and not increasing, QoL. Therefore, when analyzing thyroid hyperfunction within the model, serum TSH values above the general population median, and serum T4 and serum T3 values below the median were set to missing. Similarly, when analyzing hypofunction, serum TSH values below the median in the general population and serum T4 and serum T3 above the median were set to missing.

---

**Statistical Analyses**

Initially, pairwise relationships between all variables in the theoretical model were analyzed univariately, controlling for age, gender and educational level, using SAS 9.1.3 PROC GLM [15].

Subsequently, statistically significant associations (p < 0.05) were joined in one overall path model, using Mplus [16] with maximum likelihood estimation. Model results were checked for colinearity between hormone measurements and nonnormal distribution of clinical variables, particularly serum TSH, which was also entered as squared and cubed values.

To test the stability of the results, several post hoc analyses were performed. In the univariate pairwise analyses of the relationship between QoL scales and thyroid function, the latter was also operationalized as (a) current thyroid function, i.e. euthyroidism (n = 92, 46%), mild hypothyroidism (n = 77, 39%), overt hypothyroidism (n = 3, 2%), mild hyperthyroidism (i.e. overtreated) (n = 14, 7%) and overt hyperthyroidism (n = 8, 4%); (b) current euthyroidism (euthyroidism vs. any dysfunction) and (c) degree of dysfunction at the time of diagnosis (mild vs. overt hypothyroidism). The multivariate analysis concerning relationships with thyroid autoimmunity was repeated, controlling for (a) age, (b) duration of disease, (c) whether or not on 1-thyroxine replacement therapy, (d) dosage and duration of 1-thyroxine therapy and (e) current thyroid function (euthyroidism, mild/overt hypothyroidism and hyperthyroidism) by adding these variables as covariates to the regression equations.

---

*Watt/Hedégüs/Bjönr/Grøenved/ Bonnema/Rasmussen/Feldt-Rasmussen*
Results

Univariate Associations
In the univariate, controlled comparisons, clinical variables measuring goitre (i.e. thyroid volume), thyroid dysfunction (i.e. TSH, T4 and T3) and eye involvement (i.e. CAS and DOSPECS) were not related to QoL scores. TPOAb levels were statistically significantly associated with several QoL dimensions: Goitre Symptoms (p = 0.024), Depressivity (p = 0.004), Anxiety (p = 0.004), Emotional Susceptibility (p = 0.005) and Impaired Social Life (p = 0.047).

In the controlling post hoc analyses, none of the alternative operationalizations of thyroid function was associated with QoL scores.

Multivariate Path Model
In the path analysis multivariate model, TPOAb was related to the symptom scales: Goitre (p = 0.019), Depressivity (p = 0.001) and Anxiety (p = 0.002) but no longer to Emotional Susceptibility or Impaired Social Life (fig. 2). Post hoc analyses controlled for age, duration of disease, whether the patient was being treated with L-thyroxine and the dosage and duration of this treatment, current thyroid function status as well as transforming relevant clinical variables (see above) did not significantly change the results (data not shown).

Discussion
The purpose of the present study was to evaluate the relationship between clinical measures of disease activity and health-related QoL in patients with autoimmune hypothyroidism.

In univariate comparisons between the relevant clinical variables and the ThyPRO QoL scales, no relationship between measures of thyroid dysfunction was found. In contrast, TPOAb levels were related to several ThyPRO scales, both symptom scales and scales measuring well-being and function as well as participation. The findings were repeated when the data were subjected to multivariate analysis within a theoretical QoL framework, with the exception that now TPOAb levels were only related to the symptom scales. The relationship between TPOAb and the function and well-being scale, Emotional Susceptibility, and the participation scale, Impaired Social Life, thus seem to be through an association with Depressivity.
It is surprising that we did not find any relationship between thyroid function and QoL. This finding was consistent, regardless of the way thyroid (dys-)function was operationalized, as exemplified in our post hoc analyses. We therefore also conducted post hoc repetitions of the multivariate analyses, where various operationalizations of thyroid (dys-)function were entered as covariates to all paths involving TPOAb. We also controlled the analyses for age and various other clinically relevant variables. Still, only TPOAb and not thyroid function was related to QoL. However, this is probably due to the fact that the vast majority of patients were only marginally dysthyroid at the time of participation, as a consequence of the sampling procedure spanning all disease stages, including well-treated (and even over-treated) patients. Almost half of the patients were euthyroid at the time of the study and only 2% were overtly hypothyroid. This may have been different if more patients with current overt hypothyroidism had been included. Indeed, Canaris et al. [17] and others [18, 19] have demonstrated a relationship between hypothyroid symptoms and thyroid function, when investigating patients with overt hypothyroidism and euthyroid controls. Another recent study among 597 patients receiving L-thyroxine (73% of whom had autoimmune hypothyroidism) found an association between psychological well-being and fT4 and serum TSH, but not fT3 or TPOAb positivity [20]. However, TPOAb were analyzed as a dichotomized, and not as a continuous variable as in our study, which could explain this difference. In accordance with our findings, Ott et al. [21] demonstrated a correlation between the level of TPOAb, but not thyroid function tests, and scores on an unvalidated symptom questionnaire, in a study of patients undergoing thyroidectomy for benign goitre. Another field of research has focused on the role of thyroid autoimmunity in musculoskeletal complaints, particularly fibromyalgia [22, 23]. For example, Bazzichi et al. [22] found comorbidity with and higher levels of symptoms of fibromyalgia in patients with Hashimoto’s thyroiditis and subclinical hypothyroidism, in contrast to patients with nonautoimmune subclinical hypothyroidism. Additionally, in a population-based study, a higher prevalence of TPOAb positivity was found in respondents with musculoskeletal complaints, compared to those without [24]. An interesting study supposed a possible mechanism of action: Marquez et al. [25] found microvascular alterations similar to those observed in other autoimmune diseases in skeletal muscle from patients with autoimmune thyroiditis, independent of thyroid function. These capillary alterations included changes in basement membranes and endothelial thickening, but also signs of capillary degeneration. It may be hypothesized that such microvascular alterations could also be present in other biological structures, including the central nervous system, and thus explain the associations between autoimmunity and neck-related symptoms as well as mental health observed in the present study.

Several studies have indicated a link between the thyroid gland and mental diseases [26]. As for autoimmune thyroid diseases, Pop et al. [27] found a three times higher risk of current depression (according to the Edinburg Depression Scale) in individuals with positive TPOAb in a large population-based study and the presence of TPOAb during pregnancy has been identified as a risk factor for postpartum depression [28].

A direct relation between thyroid autoimmunity and affected QoL has also been hypothesized to be at play in the rare and controversial diagnosis of Hashimoto’s encephalopathy, where a variety of neurological symptoms have been linked to the presence of TPOAb [29].

As its main strength, our study is the first to apply a thoroughly validated, thyroid-specific QoL instrument. Investigation of the associations between clinical and QoL variables in an explicit theoretical framework, in clinically well-characterized individuals, adds further validity and novelty to our findings.

The main limitation of our study is its cross-sectional design, the relatively limited range in thyroid dysfunction, and the few severely hypothyroid individuals. Future hypothesis-testing studies should adopt a longitudinal design, compare changes in all available thyroid autoantibodies, or even better various epitopes of these antibodies [30–34], and in thyroid function, with changes in QoL. Also, the relationship between such indicators of thyroid autoimmunity and QoL could be evaluated in a sample of patients with autoimmune thyroiditis and normal thyroid function. Applying diagnostic imaging, such as e.g. MR spectroscopy [35], may offer further insight into the possible CNS effect and thereby the link with affected QoL measures.

Future interventional research may target the autoimmune component of the disease more directly. For example, by investigating the effect on QoL of selenium supplementation [36], which appears to lower TPOAb levels, glucocorticoid therapy or novel biological therapies, such as rituximab [37]. Future studies focusing on QoL methodology could further develop and investigate the theoretical framework and its relationship with clinical variables. Accepting that patients with chronic autoimmune thyroiditis, with or without thyroid dysfunc-
tion, have increased morbidity and decreased QoL, but seemingly no increase in mortality [38], remains a conundrum and needs further study.

In conclusion, TPOAb level, but not thyroid function, was related to thyroid-specific QoL. This raises the hypothesis that thyroid autoimmunity may play a role, independent of thyroid dysfunction, in the QoL impairment associated with autoimmune hypothyroidism. Future studies should focus on individuals with a broader range of TPOAb as well as thyroid dysfunction in longitudinal and interventional studies, employing a validated questionnaire such as the ThyPRO for QoL determination.

Acknowledgement

This study has been supported by grants from the Danish Medical Research Council, Agnes and Knut Mørk's Foundation, Aase and Ejnar Danielsen's Foundation, Else and Mogens Wedell-Wedellsborg's Foundation, the Genzyme Corporation, Novo Nordisk Foundation, Danish Thyroid Foundation and Arvid Nilsson's Foundation.

Disclosure Statement

The authors have no conflicts of interest to declare and have nothing to disclose.

References


Paper #7
Selenium supplementation for patients with Graves’ hyperthyroidism (the GRASS trial): study protocol for a randomized controlled trial

Torquil Watt1,*, Per Cramon1, Jakob Bue Bjorner2,3, Steen Joop Bonnema4, Ulla Feldt-Rasmussen1, Christian Gluud5, Jeppe Gram6, Jane Lindschou Hansen7, Laszlo Hegedüs4, Nils Knudsen7, Pernille Bach-Mortensen8, Runa Nolsøe9, Birte Nygaard10, Flemming Pociot11, Maria Skoog5, Per Winkel5 and Åse Krogh Rasmussen1

Abstract

Background: Graves’ hyperthyroidism is an autoimmune disease causing hyperfunction of the thyroid gland. The concentration of selenium is high in the thyroid gland and two important groups of enzymes within the thyroid are selenoproteins, that is, they depend on selenium. Selenium may have beneficial effects on autoimmune hypothyroidism and on Graves’ orbitopathy, but the effects of selenium on Graves’ hyperthyroidism is unknown. We hypothesize that adjuvant selenium may be beneficial in the treatment of Graves’ hyperthyroidism. The objective is to investigate if selenium supplementation plus standard treatment with anti-thyroid drugs versus standard treatment with anti-thyroid drugs will lead to a decrease in anti-thyroid drug treatment failure (that is, failure to remain euthyroid, without further treatment, one year after cessation of anti-thyroid drug treatment), faster and longer lasting remission (that is, anti-thyroid drug treatment success), and improved quality of life in patients with Graves’ hyperthyroidism.

Methods and design: The trial is an investigator-initiated, randomised, blinded, multicentre clinical trial. Inclusion criteria are: age 18 years or older; diagnosis of active Graves’ hyperthyroidism within the last two months; and informed consent. Exclusion criteria are major co-morbidity; previous radioactive iodine treatment; ongoing anti-thyroid drug treatment for more than two months; treatment with immunomodulatory drugs; known allergy towards the components in the selenium and placebo pills; pregnancy or breast-feeding; and intake of selenium supplementation above 70 μg per day. We plan to include 492 participants, randomised (1:1) to two tablets of 100 μg selenium once daily for the 24 to 30 months intervention period versus two identical placebo tablets once daily. The primary outcome is the proportion of participants with anti-thyroid drug treatment failure (see above) at the end of the intervention period (24 to 30 months). Secondary outcomes are: thyroid-specific quality of life during the first year after randomisation; level of thyroid stimulating hormone-receptor antibodies at 18 months after randomisation and at the end of the intervention period (24 to 30 months); hyperthyroid symptoms during the first year after randomisation; eye symptoms during the first year after randomisation, and at the end of the intervention period (24 to 30 months); adverse reactions during the intervention period; and serious adverse events during the intervention period.

(Continued on next page)
Background

Graves’ hyperthyroidism is an autoimmune disease causing hyperfunction of the thyroid gland by a mechanism, where auto-antibodies bind to and stimulate the thyroid stimulating hormone (TSH) receptor. It affects individuals of all ages and is five to ten times more common in women than in men, with an overall annual incidence in Denmark of about 1,700 patients in 5.5 million people [1]. Treatment of Graves’ hyperthyroidism comprises: 1) anti-thyroid drugs (ATD) (for example, methimazole or propylthiouracil); 2) thyroidectomy, that is, surgical removal of the thyroid gland, and 3) radioactive iodine treatment (I\(^{131}\)), which permanently reduces thyroid function. In Europe, the primary treatment is usually ATD, which leads to resolution of hyperthyroidism (that is, euthyroidism) in 85% to 90% of patients within 6 weeks [2]. This treatment is usually continued for a period of 12 to 18 months and then tapered off (ATD treatment withdrawal). However, 30% to 60% of patients will experience relapse of hyperthyroidism during the following years [2]. Patients with Graves’ hyperthyroidism suffer from a wide range of symptoms and have major impairments in most areas of quality of life (QoL) [3].

Selenium is an essential trace element important for human health. The main dietary sources of selenium in Denmark are meat, poultry, dairy products, bread, cereals and fish [4,5]. The recommended daily intake of selenium is 40 \(\mu\)g for women and 50 \(\mu\)g for men. The estimated actual daily intake in Denmark is considered sufficient, but it cannot be ruled out that about 10% of the Danish population could benefit from selenium supplementation [4]. The upper tolerable level of selenium intake is set to 400 \(\mu\)g per day in the USA [6] and 300 \(\mu\)g per day in the EU [7].

The thyroid gland has the highest selenium concentration per unit weight among all tissues. Selenium is incorporated into key enzymes involved in several metabolic pathways. The main selenoprotein families are the glutathione peroxidases, the thiorexin reductases and the iodothyronine deiodinases [5]. It is hypothesized that the glutathione peroxidases and the thiorexin reductases participate in a complex defence system maintaining normal thyroid function by protecting the gland from both hydrogen peroxide, which is produced by the thyrocytes, and reactive oxygen intermediates [8,9]. It has thus been hypothesized, that selenium may have a beneficial role in autoimmune thyroid diseases, by blunting the autoimmune process.

We have not identified any published randomised trials of selenium supplementation in patients with Graves’ hyperthyroidism. On ClinicalTrials.gov, one ongoing randomised trial is registered on selenium supplementation in Graves’ thyrotoxicosis (NCT01247077). This trial assesses neuropsychological well-being (not otherwise specified), after nine months intake of 200 \(\mu\)g selenium, and includes 44 participants. In a multicentre trial among 159 patients with mild Graves’ orbitopathy, without hyperthyroidism, 200 \(\mu\)g selenium selenite improved disease-specific QoL (\(P <0.001\)) and reduced eye disease severity (\(P = 0.01\)) [10]. Another trial evaluated the effect of adding a mixture of antioxidants, including 60 \(\mu\)g of selenium (not otherwise specified) to standard ATD treatment in 29 patients with Graves’ disease. During the 60-day follow-up period euthyroidism was reached more rapidly in patients receiving antioxidants [11-13]. In contrast, several randomised trials have evaluated the effect of selenium on the other major autoimmune thyroid disease: autoimmune hypothyroidism. In six [14-19] of seven [14-20] placebo-controlled clinical trials, selenium treatment reduced thyroid peroxidase antibody (TPOAb) levels, indicating a beneficial effect on the autoimmune activity.

We hypothesize that the addition of selenium supplementation to the standard treatment with ATD in patients with active Graves’ hyperthyroidism will lead to a decrease in ATD treatment failure (that is, fewer patients with relapse), faster remission, and improved quality of life.
Methods and design

Objectives
The primary objective is to investigate the effect of selenium supplementation on the proportion of participants with ATD treatment failure, that is, failure to remain euthyroid without further treatment one year after cessation of ATD treatment.

The secondary objectives are to investigate the effect on thyroid-specific QoL, level of TSH-receptor antibody (TRAb), hyperthyroid symptoms, eye symptoms, adverse reactions, and serious adverse events. Further, we wish to explore the effect of selenium on ATD treatment duration, incidence of Graves’ orbitopathy, and hypothyroid symptoms.

Design
The GRASS (GRAves’ disease Selenium Supplementation trial) trial is an investigator-initiated, randomised, blinded, multicentre clinical trial of selenium supplementation versus placebo in patients with Graves’ hyperthyroidism. The trial has a parallel-arm design with 1:1 allocation to the experimental intervention group and the control intervention group, and involving seven clinical trial sites in Denmark (see Figure 1). The trial also includes a questionnaire and register-based follow up period up to ten years after completion of the intervention period.

Trial participants
All patients with current hyperthyroidism, who are referred to or being followed at the participating clinical trial sites, are considered for participation. Patients are eligible for the GRASS trial, if they comply with the following inclusion and exclusion criteria.

Inclusion criteria
The inclusion criteria are: Graves’ hyperthyroidism (first-time diagnosis, defined as patients not yet receiving ATD treatment, or having received ATD treatment continuously for less than two months, or relapse of Graves’ hyperthyroidism defined as patients previously having received and discontinued treatment with ATD); active Graves’ hyperthyroidism (TSH <0.1 mU/L and positive TRAb according to local laboratory results) measured within the last two months prior to the inclusion date; age 18 years or older; provision of written informed consent.

Exclusion criteria
The exclusion criteria are: major co-morbidity, rendering the participants unlikely to continuously receive the trial intervention; previous treatment with radioactive iodine; ongoing ATD treatment for more than two months; treatment with immunomodulatory drugs, such as cyclosporine A, methotrexate, cyclophosphamide; allergy to the components in the selenium and placebo pills; pregnancy or breast-feeding; intake of selenium supplementation above 70 μg per day; inability to read and understand Danish; lack of provision of informed consent.

Trial intervention
Selenium
The compound used in this trial is organic selenium, in the form of selenium yeast, which mainly consists of selenomethionine. The specific product is Organisk selen, 100 μg tablets and is produced by Jemo-Pharm A/S, Stege, Denmark (http://www.jemo-pharm.dk/frame.cfm/cms/id=977/sprog=2 grp=6/menu=1/). The daily dose is set at 200 μg (two tablets taken in the morning). The trial dosage is based on the available clinical data, is not considered to cause adverse reactions, and is lower than the upper tolerable intake level of 300 μg per day [6,7].

Placebo
Placebo tablets, identical in size, appearance, taste, smell, and solubility to the experimental intervention tablet are produced by Jemo-Pharm A/S. They have the same content as Organisk selen but are without selenium, as the selenium yeast has been exchanged with yeast grown in selenium-deplete media. The placebo regimen is identical to the selenium regimen.

Randomisation
Randomisation will be performed centrally. The allocation sequence is computer-generated with a varying block size kept unknown to the investigators. Randomisation is stratified by clinical trial site and disease status (incident or relapse), and the allocation ratio is 1:1.

Informed consent procedure
Potential participants are identified at referral or visits to the outpatient clinics and include all patients with current hyperthyroidism (that is, elevated thyroid hormone levels). Where Graves’ hyperthyroidism is confirmed (that is, TRAb is positive), the potential participant is invited to an information visit by letter. The visit consists of obtaining history, blood sampling, information about trial contact, and randomisation.

Duration
The intervention period for each participant will be 24 to 30 months, as selenium will be given until 12 months after cessation of ATD treatment, which usually lasts 12 to 18 months (Figure 2). The total trial duration is expected to be about 4 years (inclusion period about 17 months and up to 30 months intervention period).

Concomitant medication or treatment
The trial participants receive conventional ATD treatment. Treatment of hyperthyroidism, will take place within the
participating clinical trial sites, according to the clinical standards set by the departments. ATD treatment withdrawal must be considered or attempted 18 months after randomisation at the latest.

Participants are advised not to take extra selenium supplementation during the trial. A dose of <70 μg per day is allowed (the content in multivitamin tablets is 55 μg). The participants’ consumption of additional selenium supplements will be monitored during the trial through self-report forms.

**Monitoring for intervention compliance**

Participant compliance with the intervention will be monitored by self-reported tablet intake at 6 and 12 weeks, and 6, 12 and 24 months (Table 1), and by tablet-counting at trial visits at 18 months and at the end of the intervention (24 to 30 months). An investigator will contact the patient if the trial data management system flags a participant as non-compliant (or overdosing), according to self-report of the number of unopened containers.

**Discontinuation**

A participant who no longer wishes to participate in the trial can withdraw his/her informed consent at any time without need of further explanation, and this will not have any consequences for the participant’s further treatment. In order to conduct intention-to-treat analyses with as few missing data as possible, the investigator may ask the participants which aspects of the trial, they wish to withdraw from. These can include the following:
receipt of the trial intervention; participation in the remaining follow-up assessments, and analysis of data already collected. The investigators will discontinue a participant’s taking of the trial intervention at any time, if the participant: experiences intolerable adverse reactions; is diagnosed with any of the exclusion criteria during the intervention period; is referred for ablative therapy (radioactive iodine or thyroid surgery) during the intervention period. In all three cases, the investigator and/or the treating physician will, if possible, encourage the participant to continue with follow-up assessment and to allow the use of collected data in the analyses.

**Blinding**

Blinding will be maintained for all parties in the trial, throughout all aspects of the trial. The trial interventions will be identical, and will be packed in identical packages by the Capital Region Pharmacy, and therefore, knowledge

---

**Table 1 Trial schedule for assessments**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Inclusion (baseline)</th>
<th>6 weeks</th>
<th>12 weeks</th>
<th>6 months</th>
<th>12 months</th>
<th>18 months</th>
<th>24 months</th>
<th>12 (± 1) months after ATD treatment withdrawal†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATD treatment</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thyroid function</td>
<td>x,†</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSH-Receptor Antibodies</td>
<td>x,†</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum selenium</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creatinine/iodine ratio in spot urine</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage samples (blood and urine)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ThyPRO</td>
<td>x,†</td>
<td>x,†</td>
<td>x,†</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Tablet count</td>
<td>x,†</td>
<td>x,†</td>
<td>x,†</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Consumption of additional selenium</td>
<td></td>
<td>x,†</td>
<td>x,†</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Adverse events*</td>
<td>x,†</td>
<td>x,†</td>
<td>x,†</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Serious adverse events</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Referral to ablative therapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

All assessments must be made at the time points specified above. If the assessment is not possible at the specified time, the assessment shall still be conducted, and the time shall be noted in the electronic case report form (eCRF). Following this, deviations from the protocol can be assessed. †As part of usual clinical practice (f = free (non-protein bound)), that is, all measured analyses are collected continuously from medical systems; x, analysed on stored plasma/serum samples after trial completion; x, self-reported data by participant; †participants for whom ATD treatment withdrawal has not been attempted, or has been unsuccessful, will be followed up at 36 months after randomisation (± 1 month). ATD, anti-thyroid drugs; TSH, thyroid stimulating hormone; ThyPRO, thyroid patient-reported outcome.
of allocated intervention group will be unknown to participants and investigators. All outcome assessments will be performed blinded and statistical analyses will be performed with the blinding intact.

**Safety**

Selenium is tolerated in short-term doses up to 10,000 μg (that is, 100 experimental GRASS tablets). Acute intoxication is very rare and has only been related to accidental or suicidal intake [4,21,22]. Chronic intoxication requires long-term intake of at least 800 μg per day [23], according to epidemiological studies in areas with very high selenium content in the soil. Thus, no signs of chronic intoxication have been observed in areas with daily intakes of up to about 800 μg daily. In geographically highly exposed groups without signs of intoxication the serum concentrations have been reported to be 148 to 363 μg per L [24] and 284 to 472 μg per L [23], respectively. In a study of 200 μg selenium supplementation per day, the serum concentration reached a plateau at around 190 μg per L [25].

According to the Danish National Food Institute, the 99th percentile for daily selenium intake through diet in Denmark is 93 μg for men and 72 μg for women. Therefore, a dosage of 200 μg per day should not bring trial participants above the established upper tolerable intake of 300 μg per day. A review of the safety of selenium supplementation with selenium yeast, which will be used in the GRASS trial [26], concludes that in about a dozen supplementation studies, none has shown evidence of toxicity even up to an intake of 800 μg selenium per day over a period of years. In conclusion, the experimental intervention with 200 μg selenium per day is not expected to cause adverse reactions. Regardless, participants will be monitored for adverse events.

Overdosing can lead to gastrointestinal discomfort, hair and nail malformations and loss, peripheral neurological symptoms, fatigue and dizziness, and, in the case of very large selenium loads (1,000 times the daily trial dose), cardiovascular collapse and respiratory distress [21,22].

Assessment and reporting of adverse reactions (ARs): participants are prompted to self-report ARs at 6 and 12 weeks, and 6, 12 and 24 months, and are questioned about ARs at the study visits at 18 months of treatment and 12 months after stopping ATD treatment, respectively. ARs will be reported as a trial outcome. In addition, participants are instructed to contact their trial contact person if they experience symptoms suggestive of ARs.

Assessment and reporting of serious ARs (SARs), serious suspected SARs (SUSARs) and serious adverse events (SAEs): data on hospital admissions and mortality will be obtained through national registries at the end of the trial. Also, participants are informed and instructed to contact their trial contact person if they are admitted to a hospital for selenium intoxication, experience a clinical picture indicative of selenium intoxication, or experience a clinical picture that is unexpected but suspected to be related to selenium intoxication. When a possible serious event (SAE, SAR, or SUSAR) is identified, details will be sought from the patient’s medical record and through direct contact with the patient. Any SAE, SAR or SUSAR will be reported as an outcome measure.

**Outcomes**

Outcomes will be assessed seven times during the trial (Table 1).

**Primary outcome**

The primary outcome is the proportion of participants with the composite outcome of ATD treatment failure in participants receiving ATD treatment during the last 12 months (± 1 month) of the intervention period, who have had thyroid hyperfunction (TSH <0.1) during the last 12 months (± 1 month) of the intervention period, or have been referred for ablative therapy (radioactive iodine or thyroid surgery) at some point during the entire intervention period.

**Secondary outcomes**

The secondary outcomes are each component of the primary outcome as follows: proportion of participants who receive ATD treatment (at any level) during the last 12 months (± 1 month) of the intervention period (separate component of the primary outcome); proportion of participants who have thyroid hyperfunction (TSH <0.1) during the last 12 months (± 1 month) of the intervention period (separate component of the primary outcome); proportion of participants who have been referred to ablative therapy (radioactive iodine or thyroid surgery) at some point during the entire intervention period (separate component of the primary outcome); thyroid-specific QoL during the first year after randomisation, and at the end of the intervention period (24 to 30 months), as measured by the global score in the ThyPRO questionnaire (Appendix 1); level of TRAb at 18 months, and at the end of the intervention period (24 to 30 months); hyper-thyroid symptoms (ThyPRO subscale) during the first year after randomisation; eye symptoms (ThyPRO subscale) during the first year after randomisation, and at the end of the intervention period (24 to 30 months); number of patients with ARs during the intervention period, and number of patients with serious adverse events during the intervention period.

**Exploratory outcomes**

The following outcomes are of an exploratory nature: time to ATD treatment withdrawal (unsuccessful participants will be censored at 18 months); cost-effectiveness
of the experimental intervention; incidence of Graves’ orbitopathy during the intervention period, assessed as clinical activity score (CAS) >1 among patients with CAS scoring in the medical chart, and hypothyroid symptoms (ThyPRO subscale) during the intervention period.

**The trial data management system**
As a result of the pragmatic design with minimal participant-trial interaction, a large part of the data collection, trial conduct, and trial surveillance and timing is handled by the trial data management system. This trial data management system consists of a patient-survey-interface, a trial-personnel-interface, a system-integration interface and a programme *motor*. The system will be used for collection of outcomes, adverse events, and other trial-relevant information; for timing of trial events, that is, time for patient-reported outcomes and trial visits; for identification of need for actions (for example, contact to a participant); and for delivery of output to personnel or participants (for example, email notifications).

**Research biobank**
A research biobank for serum/plasma and urine samples will be established. The samples will initially be kept at each clinical trial site, and will later be analysed at central laboratories. Participants are informed verbally and in writing, and will consent to the withdrawal and storing of biological material in the GRASS trial. Any remaining samples will be transferred to a Biobank designated for future use. These samples will be used for genome-wide association studies (of predictors for remission and experimental intervention effects, or other indicators of autoimmunity), as may be specified in forthcoming protocols.

**Monitoring**
The trial will be monitored according to the International Committee of Harmonization (ICH) guidelines for good clinical research practice [27] by internal monitoring.

**Statistical analysis**

**Primary outcome: ATD treatment failure - sample size estimation**
Prior data indicate that the proportion of patients in the placebo group with ATD treatment failure is 50% [28]. If the true proportion with ATD treatment failure is 37.5% among selenium-treated participants (that is, a relative risk reduction of 25%), we will need to include 492 participants (246 experimental and 246 control participants) to be able to reject the null hypothesis with a power of 80% and a risk of type I error of 5%.

The patient catchment areas of the participating centres include 1,640,000 persons. Assuming an annual incidence of Graves’ hyperthyroidism of 40 per 100,000 after iodine fortification [1,29] and 50% recurrence rate, the incident population (including recurrences) will be 788 patients annually. Assuming an inclusion of 45% of the potential patients, this will lead to about 29 participants per month and an inclusion period of about 17 months.

**Secondary outcomes - power estimation**

For the ThyPRO Global score, hyperthyroid symptoms and eye symptoms, if the true difference between experimental and control participants is 5 (on a scale of 0 to 100) with SD of 20, we will be able to reject the null hypothesis that the population means of the experimental and control groups are equal with a probability (power) of 79%. The associated type I-risk is 5%.

For the level of TRAb, if the true difference in TRAb between experimental and control participants is 0.15 IU per L with SD of 0.5 IU per L, we will be able to reject the null hypothesis with a probability (power) of 91% and a type I error-risk of 5%.

For the number of patients with adverse reactions, prior data indicate that the proportion of participants who experience adverse reactions in the control group is about 5% [26]. If the true proportion of participants in the experimental group who experience adverse reactions is 10%, we will be able to reject the null hypothesis that the failure rates for experimental and control participants are equal with probability (power) 50%. The type I error probability associated with this test is 5%.

**Data analysis**
All analyses will be intention-to-treat analyses with the intervention group concealed until two conclusions are drawn. The significance test will be at the 5% level and two-sided. Table 2 shows the priority for each outcome, when it will be measured, the mathematical type of measure, and the analytical procedure to be used when analyzing the outcome values.

**Analytical procedures**
Depending on the specific type of outcome measure, one of five types of regression analysis will be applied (Table 2). Indicator variable I (1 if X and 0 if Y) is included as a covariate and the outcome measure (y) as the dependent variable. All analyses will be conducted both as unadjusted analyses and adjusted for stratification variables (clinical trial site and disease status (incident or relapse)). Discrepancies between the results of the two analyses will be discussed. For each covariate, an exploratory analysis, including the interaction between the covariate and the intervention indicator, will be conducted.

If the distribution of the primary outcome measure differs significantly between the two intervention groups, and the percentage of missing values is larger than 5%, multiple imputations (MI) will be used. If so, the result obtained by imputation will be the primary result.
Table 2 Outcome measures, their priorities, times of measurement, mathematical types and analytical categories (defining the statistical analysis to which they will be subjected)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Times of measurements</th>
<th>Type of quantity</th>
<th>Regression analysis to be applied</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATD treatment failure</td>
<td>End of trial</td>
<td>Binary</td>
<td>Logistic regression</td>
</tr>
<tr>
<td><strong>Secondary outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. ATD treatment within last 12 months</td>
<td>End of trial</td>
<td>Binary</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>2. Thyroid hyperfunction after ATD treatment withdrawal</td>
<td>End of trial</td>
<td>Binary</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>3. Ablative therapy</td>
<td>End of trial</td>
<td>Binary</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>4. Global QoL ThyPRO score</td>
<td>a) Time sequence of five measurements(†)</td>
<td>Numerical</td>
<td>a) Mixed-model with repeated measures (MMRM)</td>
</tr>
<tr>
<td></td>
<td>b) 12 months following ATD treatment withdrawal(†)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Level of TRAb</td>
<td>After 18 months and at the end of intervention period</td>
<td>Numerical</td>
<td>General linear univariate model. As sensitivity analysis: Mann-Whitney test</td>
</tr>
<tr>
<td>6. ThyPRO - hyperthyroid symptoms</td>
<td>a) Time sequence of five measurements(†)</td>
<td>Numerical</td>
<td>a) Mixed-model with repeated measures (MMRM)</td>
</tr>
<tr>
<td></td>
<td>b) 12 months following ATD treatment withdrawal(†)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. ThyPRO - eye symptoms</td>
<td>a) Time sequence of five measurements(†)</td>
<td>Numerical</td>
<td>a) Mixed-model with repeated measures (MMRM)</td>
</tr>
<tr>
<td></td>
<td>b) 12 months following ATD treatment withdrawal(†)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Adverse reactions</td>
<td>End of trial</td>
<td>Rate = count/period of intervention/day</td>
<td>Generalised linear model, Poisson distribution, link = log. As sensitivity analysis: Mann-Whitney test</td>
</tr>
<tr>
<td>9. Serious adverse reactions</td>
<td>End of trial</td>
<td>Rate = count/period of intervention/day</td>
<td>Generalised linear model, Poisson distribution, link = log. As sensitivity analysis: Mann-Whitney test</td>
</tr>
<tr>
<td><strong>Exploratory outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Time to ATD withdrawal §</td>
<td>End of trial</td>
<td>Numerical (time until ATD withdrawal (or censoring))</td>
<td>Cox proportional hazard rate model. As sensitivity analysis: Kaplan-Meier estimates of survival function</td>
</tr>
<tr>
<td>2. Incidence of Graves’ orbitopathy - CAS score</td>
<td>End of trial</td>
<td>Binary (CAS &gt;1)</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>3. ThyPRO - Hypothyroid symptoms</td>
<td>a) Time sequence of five measurements(†)</td>
<td>Numerical</td>
<td>a) Mixed-model with repeated measures (MMRM)</td>
</tr>
<tr>
<td></td>
<td>b) 12 months following ATD treatment withdrawal(†)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(†\)This analysis (2) includes four measurements relative to the reference time (6 weeks, 12 weeks, 6 months and 12 months). This analysis (3) includes one measurement (12 months following ATD treatment withdrawal or failure to withdraw (alternative treatment instituted or deadline expired)) relative to a different reference time. \(§\)End of ATD treatment or censoring at 18 months after randomisation or ablative surgery, provided the latter takes place during ATD treatment and prior to time of censoring. ATD, anti-thyroid drugs; QoL, quality of life; ThyPRO, thyroid patient-reported outcome; TRAb, TSH-receptor antibodies; CAS, clinical activity score.

events, the potential for bias caused by non-random missing values will be assessed using a worst- and best-case scenario. The gate keeping method of Dmitrienko et al. [30] will be used to adjust the observed P-values.

**Ethical considerations**

The GRASS trial has been approved by the Regional Ethics Committee (H-4-2012-026). The trial will be conducted in compliance with the guidelines of the latest Declaration of Helsinki and the International Conference on Harmonization (ICH) Good Clinical Practice (GCP) Guidelines [27].

**Publication plan**

The aim is to publish all results, positive, neutral, and negative, in peer-reviewed international journals. Authorship will be determined according to the guidelines from the International Committee of Medical Journal Editors [31].

**Discussion**

The design of the GRASS trial has faced three major challenges: the intent to conduct it as a pragmatic trial with as little interference with daily clinical management as possible, while still measuring QoL meticulously; the intent to comply with the ICH guidelines for GCP, even...
though the intervention is widely considered to be non-toxic and rather, a supplemental nutrient than a drug; and the need for a very long intervention period.

It was of great importance to the initiators of this trial, that the results would be directly applicable to daily clinical practice. Therefore, it was decided to conduct a pragmatic trial [32], with as little interference with daily clinical practice as possible. This could be accomplished by letting the patients follow their usual treatment at their usual hospitals by whichever physician was involved in their treatment. At the same time, we found it important to collect high quality data on the clinical course and QoL. We have, therefore, put great effort into the design of a trial management system, which could solve this schism of distance from participants compared to close monitoring. The system initiates and keeps track of patient input (QoL-measurements and other patient-reported outcomes), identifies need for trial personnel input and action, and collects data on thyroid function from medical chart systems. The system also identifies the site of the information-provider, as well as the disease status of new participants, and delivers randomisation codes stratified by site and disease status. Meticulous follow up on missing responses to the QoL measurements is incorporated into the system, to minimise the usual major problem with missing QoL data in clinical trials.

Somewhat similarly, the need to monitor adverse reactions and events in accordance with the ICH guidelines was in conflict with the intent to interfere as little as possible and with the fact that no previous selenium trial has identified any adverse reactions, which would also be quite surprising, given the wide therapeutic range of this dietary supplement. This is solved by a combination of thorough instruction of the participants to contact their trial person in case of symptoms indicative of adverse reactions or events, surveillance of patient-responses to prompts through the trial management system, and integration with national databases regarding hospitalisations.

The third issue was the decision to continue trial intervention until one year after stopping anti-thyroid drugs (ATD treatment withdrawal). Patients are not considered in remission unless they are still euthyroid one year after stopping medication. Since selenium is considered to have an attenuating effect on thyroid autoimmunity, a long duration of treatment was considered necessary to test our hypothesis, that is, that selenium can indeed lead to more patients staying in remission (and not just reaching euthyroidism faster and obtaining better QoL).

**Trial status**
The first patient was enrolled in December 2012.

**Abbreviations**
ATD: Anti-thyroid drug; AR: Adverse reaction; CAS: Clinical activity score; CRF: Case report form; GCP: Good Clinical Practice; ICH: International Conference on Harmonization; QoL: Quality of life; SAE: Serious adverse event; SAR: Serious adverse reaction; s-Se: Serum selenium; SmPC: Summary of products characteristics; SUSAR: Suspected serious adverse reaction; T3: Triiodothyronine; T4: Thyroxine; ThyPRO: Thyroid patient related outcome (thyroid-specific quality of life questionnaire); TPOAb: Thyroid peroxidase antibody; TSH: Thyroid stimulating hormone; TRAB: Thyroid stimulating hormone receptor antibody (TSH receptor antibody).

**Competing interest**
None of the investigators have any financial or non-financial competing interest.

**Author contributions**
All authors contributed to the design of the trial, preparation and review of the manuscript and all authors read and approved the final manuscript.

**Funding**
The GRASS trial is part of the ThyQoL project, a strategic project headed by Ulla Feldt-Rasmussen and funded by the The Danish Agency for Science, Technology and Innovation (grant 271-09-0143). It is also part of a grant given to Torquil Watt by the Danish Council for Independent Research (grant 09–066886).

**Author details**
1. Department of Medical Endocrinology, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.
2. National Research Centre for the Working Environment, Copenhagen, Denmark.
3. Institute of Public Health Science, University of Copenhagen, Copenhagen, Denmark.
4. Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark.
5. Copenhagen Trial Unit, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark.
6. Department of Endocrinology, Hospital of Southwest Denmark Esbjerg, Esbjerg, Denmark.
7. Department of Endocrinology and Gastroenterology, Copenhagen University Hospital Bispebjerg, Copenhagen, Denmark.
8. Department of Endocrinology, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark.
9. Department of Cardiology and Endocrinology, Endocrine Unit, Copenhagen University Hospital Hillerød, Hillerød, Denmark.
10. Department of Internal Medicine, Endocrine Unit, Copenhagen University Hospital Herlev, Copenhagen, Denmark.
11. Glostrup Research Institute, Copenhagen University Hospital Glostrup, Copenhagen, Denmark.

**Received: 4 September 2012 Accepted: 10 April 2013**

**Published: 30 April 2013**

**References**


Development of a Short Version of the Thyroid-Related Patient-Reported Outcome ThyPRO

Torquil Watt,1,2 Jakob Bue Bjorner,3,4 Mogens Groenvold,4,5 Per Cramon,1 Kristian Hillert Winther,6 Laszlo Hegedu˝s,6 Steen Joop Bonnema,6 Åse Krogh Rasmussen,1 John E. Ware, Jr., 7 and Ulla Feldt-Rasmussen1

Background: Thyroid diseases affect quality of life (QoL). The Thyroid-Related Patient-Reported Outcome (ThyPRO) is an international comprehensive well-validated patient-reported outcome, measuring thyroid-related QoL. The current version is rather long—85 items. The purpose of the present study was to develop an abbreviated version of the ThyPRO, with conserved good measurement properties.

Methods: A cross-sectional (N = 907) and a longitudinal sample (N = 435) of thyroid patients were analyzed. A graded item response theory (IRT) model was fitted to the cross-sectional data. Short-form scales with three items were aimed for, by selecting items with best fit according to the IRT model, avoiding cross-culturally noninvariant items. Seven scales measuring mental and social well-being and function as well as one overall QoL impact item were analyzed in a bifactor model, to develop a supplementary composite score. Short-form scales were linked to original scales with IRT-based summed-score-linking. Agreement between the short and long form was estimated by agreement plots, intraclass correlations, and mean score levels. Responsiveness was compared by relative validity indices, clinical validity by ability to detect clinically relevant differences, and test–retest reliability by intra–class correlation.

Results: One four-item scale was not abbreviated and one two-item scale was omitted from the short-form. For the 11 scales undergoing abbreviation, 10 with three and one with four items were developed. A bifactor model with good overall fit was fitted to the composite score, including the single QoL item. Responsiveness and clinical validity of the short-form scales were preserved, as were test–retest reliability (0.75–0.89). Short- versus long-form intraclass correlations were high (0.89–0.98), and the mean scale levels were similar.

Conclusions: A 39-item version of the ThyPRO, with good measurement properties, was developed and is recommended for clinical use.

Introduction

Diseases related to the thyroid gland are common, affecting around 10–15% of the adult population in most countries (1–3). Irrefutably, thyroid diseases affect quality of life (QoL) (4–9), work role function (10,11), as well as morbidity and mortality (12,13). Nonetheless, research focusing on QoL among these patients has been scarce, and until recently, no validated thyroid-specific patient-reported outcome (PRO) measuring thyroid-related QoL across different diseases has been available.

The ThyQoL project was launched to address this shortage (14). The Thyroid-Related Patient-Reported Outcome (ThyPRO) instrument was developed as a comprehensive thyroid-related standalone PRO for patients with any benign thyroid disease (15–17). It was crucial to the developers that it covered all benign thyroid diseases in order for it to maintain content validity when patients “converted” from one diagnosis to the other because of treatment (e.g., patients with nontoxic goiter becoming hypothyroid and requiring thyroid hormonal substitution after thyroidectomy). The ThyPRO is now in use in many studies worldwide (e.g., Mishra et al. (7), Watt et al. (18), Watt et al. (19), Graf et al. (20), Fast et al. (21), Bukvic et al. (22), Bukvic et al. (23), and Winther et al. (24)). However, the current version is rather long (85 items) and is reported in numerous (i.e., 13) scales (Supplementary...
Appendix S1: Supplementary Data are available online at www.liebertpub.com/thy). A shorter version for use as, for example, secondary outcome in clinical trials and in daily clinical practice would further advance its applicability (25).

The purpose of the present study was to develop an abbreviated version of the ThyPRO with good cross-cultural validity and with maximum preservation of favorable measurement properties in terms of construct and clinical validity, test–retest reliability, and responsiveness to relevant clinical treatments.

Methods

Study population

Data from two previously described patient populations were used (16,17,26). The cross-sectional sample comprised thyroid patients followed at or referred to two university hospital outpatient clinics (Copenhagen University Hospital Rigshospitalet and Odense University Hospital), in 2007–2008 (16) (Table 1). Thus, this sample comprised patients with newly diagnosed thyroid disease, as well as patients controlled for ongoing treatment. A subset of these, the retest subsample in Table 1, was evaluated twice, at two-week intervals (17). The longitudinal sample comprised patients undergoing treatment for thyroid diseases at the above-mentioned centers during 2008–2012, evaluated before and six months after treatment (26).

Patient-reported outcome measure

The ThyPRO measures a range of aspects of QoL relevant to patients with benign thyroid diseases, as identified during patient and expert interviews (14). It thus covers both physical symptoms specifically relevant to thyroid diseases, for example, symptoms of hyperthyroidism and goiter, and nonspecific aspects of high importance to patients with thyroid diseases, for example fatigue. The full-length ThyPRO consists of 85 items summarized in 13 scales, as well as a single item measuring overall impact of thyroid disease on QoL (Supplementary Appendix S1). Each item is rated on a 0–4 Likert scale from 0 = “no symptoms/problems” to 4 = “severe symptoms/problems.” The average score of items in a scale is divided by four and multiplied by 100 to yield thirteen 0–100 scales, with higher scores indicating worse health status.

Abbreviation strategy

The analyses were conducted in three separate steps: (a) selection of items for the short form, including selection of

<table>
<thead>
<tr>
<th>Table 1. Clinical and Basic Characteristics of the Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional sample (N = 907)</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Sex, n (%):</strong></td>
</tr>
<tr>
<td>Women</td>
</tr>
<tr>
<td>Men</td>
</tr>
<tr>
<td><strong>Age in years, median (Q1–Q3)</strong></td>
</tr>
<tr>
<td>Months from treatment initiation to completion of follow-up survey, median (Q1–Q3)</td>
</tr>
<tr>
<td><strong>Diagnosis:</strong></td>
</tr>
<tr>
<td>Nontoxic goiter</td>
</tr>
<tr>
<td>Toxic nodular goiter</td>
</tr>
<tr>
<td>Graves’ hyperthyroidism</td>
</tr>
<tr>
<td>Graves’ orbitopathy</td>
</tr>
<tr>
<td>Autoimmune hypothyroidism</td>
</tr>
<tr>
<td>Other thyroid diagnoses</td>
</tr>
<tr>
<td><strong>Disease duration in months, median (Q1–Q3)</strong></td>
</tr>
<tr>
<td><strong>Treatment, n (%):</strong></td>
</tr>
<tr>
<td>L-thyroxine</td>
</tr>
<tr>
<td>Anti-thyroid medication</td>
</tr>
<tr>
<td>Aspiration of thyroid cyst</td>
</tr>
<tr>
<td>Glucocorticoid pulse therapy</td>
</tr>
<tr>
<td>Other immunosuppressive treatment</td>
</tr>
<tr>
<td>Hemithyroidectomy</td>
</tr>
<tr>
<td>Total thyroidectomy</td>
</tr>
<tr>
<td>Radioactive iodine</td>
</tr>
<tr>
<td>No treatment (ever)</td>
</tr>
<tr>
<td><strong>Thyroid function at baseline, median (Q1–Q3):</strong></td>
</tr>
<tr>
<td>TSH mIU/L (undetectably low = 0)</td>
</tr>
<tr>
<td>Thyroxine (i.e., total T4) nmol/L</td>
</tr>
<tr>
<td><em><em>Thyroid function at follow-up</em>:</em>*</td>
</tr>
<tr>
<td>TSH mIU/L</td>
</tr>
<tr>
<td>Thyroxine (i.e., total T4) nmol/L</td>
</tr>
</tbody>
</table>

Data are in number (percent) or median (interquartile range [Q1–Q3]).

*Only in longitudinal sample.
scales for a composite score; (b) scoring of short scales; and (c) validation of the short form.

**Item selection.** The Hypothyroid scale already consists of only four items and has relatively low reliability, and it was therefore decided to retain it in full length in an abbreviated instrument. The Impaired Sex Life scale had higher occurrence of missing responses in previous studies, as an indication of lower acceptability than the remaining scales. It was therefore decided to exclude it in the abbreviated version. For each of the remaining 11 scales, items previously shown to fit a unidimensional factor model, (27) were analyzed using Samejima’s graded item response theory (IRT) model (28,29). In case of significant item misfit at the $p < 0.01$ level, according to Orlando and Thissen’s S-X² item fit index (30–33), a reduced model without the least fitting item was respecified, until a model without misfit was identified. Based on knowledge about the ThyPRO instrument, and on the IRT-derived item information functions (29), the best items were selected. Knowledge about ThyPRO stemmed from the initial qualitative content validation studies (14,15) and from other validation studies evaluating cross-cultural invalidity (34), and differential item functioning (DIF) according to diagnosis or sociodemographic characteristics (35). Since the latter DIFs were found to be minor, items with DIF were not excluded if content or measurement considerations advocated their preservation. The aim was to reduce each scale to three items, which was considered the optimal minimum number enabling meaningful evaluation of scale properties (e.g., dimensionality and DIF). In case of similar item characteristics and information curves, items were selected to cover core content of the construct, as well as important subsaspects of the construct measured (e.g., including both positive and negative aspects, such as positive energy vs. fatigue).

**Scoring composite and short-form scales.** For simplicity in future reporting, it was decided to develop a supplemental composite summary score based on factor analysis. Scales measuring mental and social well-being and function have previously been shown to be highly correlated (16,27) (i.e., the scales concerning Tiredness, Cognitive Complaints, Anxiety, Depressivity, Emotional Susceptibility, Impaired Social Life, Impaired Daily Life, as well as the single overall QoL-impact item) and were modeled as one general factor (see Fig. 1). The individual scales were modeled as subfactors, that is, a bifactor model was fitted: Each item was regressed on both the general factor and the subfactor (representing the individual, abbreviated scales). The subfactors were specified as uncorrelated with each other and with the general factor (36–40).

**FIG. 1.** Bifactor modeling evaluating the composite scale, summarizing the seven short-form mental and social well-being and function scales (left-hand side). Factor loadings are presented at the relevant arrows. All items except one had higher loading on the general factor representing the composite score.
Table 2. Results of IRT Analyses (Slopes $a$ and Item Thresholds $b_{1-4}$) of the Original Long-Form Scales, the Scales That Were Modified Based on IRT Analyses of Long-Form Scales and the Short-Form Scales

<table>
<thead>
<tr>
<th>Scale name</th>
<th>Original model</th>
<th>Modified model</th>
<th>Short form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thresholds</td>
<td>Thresholds</td>
<td>Thresholds</td>
</tr>
<tr>
<td></td>
<td>Slope 1 2 3 4</td>
<td>Slope 1 2 3 4</td>
<td>Slope 1 2 3 4</td>
</tr>
<tr>
<td><strong>Goiter symptoms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of fullness in neck</td>
<td>2.9 0.1 0.8 1.4 2.3</td>
<td>3.3 0.1 0.8 1.3 2.3</td>
<td></td>
</tr>
<tr>
<td>Visible swelling in front of neck</td>
<td>2.0 1.1 1.7 2.4 3.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure in throat</td>
<td>3.7 0.1 0.8 1.3 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain in front of neck</td>
<td>5.7 0.1 0.7 1.2 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throat pain felt in ears</td>
<td>1.3 1.2 2.2 3.1 4.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lump in throat</td>
<td>3.3 0.0 0.7 1.2 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear throat often</td>
<td>1.7 –0.2 0.7 1.4 2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty swallowing</td>
<td>2.1 0.5 1.2 1.8 2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of suffocating</td>
<td>2.3 1.2 1.8 2.2 2.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoarseness+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hyperthyroid symptoms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trembling hands</td>
<td>1.4 0.6 1.7 2.6 3.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased sweating</td>
<td>1.8 –0.4 0.5 1.1 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpitations</td>
<td>1.7 –0.2 0.8 1.6 2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shortness of breath</td>
<td>2.1 –0.2 0.7 1.5 2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity to heat</td>
<td>1.7 0.1 0.9 1.7 2.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased appetite</td>
<td>1.2 0.5 1.4 2.2 3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loose stools+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upset stomach</td>
<td>1.4 0.0 1.0 1.9 3.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Eye symptoms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watery eyes</td>
<td>1.2 0.2 1.3 2.1 3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bags under the eyes</td>
<td>1.9 –0.2 0.7 1.4 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grittiness in eyes</td>
<td>1.9 0.5 1.3 2.1 3.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced sight</td>
<td>2.0 1.2 1.9 2.3 2.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure in eyes</td>
<td>*2.2 0.9 1.6 2.2 2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double vision</td>
<td>1.8 0.3 1.2 1.7 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain in eyes</td>
<td>1.2 0.2 1.3 2.1 3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitive to light</td>
<td>1.9 –0.2 0.7 1.4 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tiredness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Been tired</td>
<td>*3.3 –1.2 –0.2 0.4 1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Been exhausted</td>
<td>4.1 –0.5 0.1 0.6 1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty getting motivated</td>
<td>*3.9 –0.5 0.3 0.7 1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt worn out</td>
<td>*4.5 –0.5 0.2 0.6 1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full of life</td>
<td>*2.6 –2.3 –0.9 0.1 1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energetic</td>
<td>*2.7 –2.4 –0.9 0.1 1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Able to cope with life</td>
<td>*2.7 –2.4 –1.1 –0.1 0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cognitive complaints</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems remembering</td>
<td>3.2 –0.4 0.6 1.3 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow or unclear thinking</td>
<td>5.0 0.0 0.8 1.4 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty finding words</td>
<td>2.8 –0.1 0.8 1.5 2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Been confused</td>
<td>2.7 0.2 1.1 1.8 2.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty learning</td>
<td>3.7 0.2 0.9 1.6 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty concentrating</td>
<td>3.8 –0.3 0.7 1.3 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Anxiety</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nervous</td>
<td>3.6 0.0 0.9 1.4 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afraid or anxious</td>
<td>3.4 0.3 1.1 1.6 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt tension</td>
<td>3.5 –0.3 0.7 1.3 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concerned being seriously ill+</td>
<td>*3.2 0.3 1.1 1.6 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uneasy</td>
<td>4.3 –0.2 0.7 1.3 1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restless</td>
<td>2.3 –0.1 0.8 1.5 2.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
The individual short-form scales were scored, using the Orlando and Thissen IRT-based summed-score linking (41), where the scales are linked based on derived summed-score-to-IRT-score translation tables, to make scale levels on the abbreviated scales comparable to the original scales. For the purpose of this linking, all items in each of the original 11 scales were calibrated using the graded IRT model. Agreement between the short- and long-form scales was estimated by agreement plots, mean score levels, and intraclass correlation with empirical bootstrap confidence intervals (42,43).

Scale validation. Responsiveness of abbreviated scales was compared to the long-form scales with calculation of effect sizes (44,45) and relative validity indices (46) in patients undergoing relevant clinical treatment, as evaluated in a previous study (26). Responsiveness in specific diagnostic groups undergoing treatment (hyper- and hypothyroidism, respectively, treated to euthyroidism, and volume reduction of goiter) was also compared to the responsiveness previously reported for the long form (26). Clinical validity was tested by evaluating whether the short-form scales were able to differentiate among clinically relevant patient groups (sensitivity), similar to what was found previously for the long-form scales (17). DIF according to age was evaluated for the short form and compared to findings from the long form (35). Finally, test–retest reliability was evaluated.

<table>
<thead>
<tr>
<th>Scale name</th>
<th>Abbreviated item wording</th>
<th>Long form</th>
<th>Initial model</th>
<th>Modified model</th>
<th>Short form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Thresholds</td>
<td>Slope 1 2 3 4</td>
<td>Slope 1 2 3 4</td>
<td>Slope 1 2 3 4</td>
</tr>
<tr>
<td><strong>Depressivity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sad</td>
<td>Sad</td>
<td>5.8</td>
<td>–0.3 0.7 1.2 1.9</td>
<td>7.5 –0.3 0.6 1.2 1.9</td>
<td>6.3 –0.3 0.7 1.3 1.9</td>
</tr>
<tr>
<td>Depressed</td>
<td></td>
<td>4.6</td>
<td>0.2 0.9 1.4 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discouraged</td>
<td></td>
<td>5.2</td>
<td>0.0 0.8 1.4 2.1</td>
<td>4.7 0.0 0.8 1.4 2.2</td>
<td></td>
</tr>
<tr>
<td>Crying easily</td>
<td></td>
<td>2.1</td>
<td>0.0 0.8 1.5 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unhappy</td>
<td></td>
<td>4.0 –0.2 0.7 1.3 1.9</td>
<td>3.8 –0.3 0.7 1.3 1.9</td>
<td>4.2 –0.2 0.7 1.3 1.9</td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td></td>
<td>*1.8 –2.1 –0.2 0.9 2.4</td>
<td></td>
<td>1.3 –2.3 –0.4 0.8 2.1</td>
<td></td>
</tr>
<tr>
<td>Self-confident</td>
<td></td>
<td>*1.6 –2.1 –0.4 0.8 2.0</td>
<td>1.4 –2.3 –0.4 0.8 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Emotional susceptibility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty coping</td>
<td></td>
<td>2.3 –0.5 0.5 1.3 2.5</td>
<td>3.3 –0.5 0.5 1.2 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not like yourself</td>
<td></td>
<td>2.5 –0.1 0.7 1.3 2.2</td>
<td>3.2 –0.1 0.7 1.2 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easily stressed</td>
<td></td>
<td>2.6 –0.6 0.4 1.0 1.8</td>
<td>3.1 –0.6 0.3 0.9 1.7</td>
<td>2.7 –0.6 0.4 1.0 1.8</td>
<td></td>
</tr>
<tr>
<td>Mood swings</td>
<td></td>
<td>*3.7 –0.5 0.4 1.0 1.7</td>
<td>2.7 –0.6 0.4 1.1 1.9</td>
<td>3.6 –0.5 0.4 1.0 1.8</td>
<td></td>
</tr>
<tr>
<td>Irritable</td>
<td></td>
<td>*3.6 –0.6 0.4 1.1 1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frustrated</td>
<td></td>
<td>4.3 –0.2 0.6 1.2 1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angry</td>
<td></td>
<td>2.4 –0.1 0.9 1.5 2.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Felt in control</td>
<td></td>
<td>*1.7 –2.0 –0.3 0.8 2.0</td>
<td>1.6 –2.1 –0.3 0.8 2.1</td>
<td>1.5 –2.1 –0.3 0.9 2.1</td>
<td></td>
</tr>
<tr>
<td>Felt in balance</td>
<td></td>
<td>*2.1 –1.9 –0.6 0.6 1.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Impaired social life</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult being with other people</td>
<td>*3.7 0.6 1.2 1.8 2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A burden to other people</td>
<td></td>
<td>4.1</td>
<td>0.7 1.3 1.9 2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conflicts with other people</td>
<td></td>
<td>*2.3 0.9 1.8 2.6 3.0</td>
<td></td>
<td>*2.3 0.9 1.8 2.6 3.0</td>
<td></td>
</tr>
<tr>
<td>People lack understanding+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Impaired daily life</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty managing daily life</td>
<td>4.9 0.3 1.0 1.5 2.2</td>
<td>4.6 0.3 1.0 1.5 2.3</td>
<td>4.9 0.3 1.0 1.5 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limit leisure activities</td>
<td></td>
<td>*5.8 0.3 0.9 1.2 1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty participating in life</td>
<td>6.5 0.5 1.0 1.4 2.0</td>
<td>5.2 0.5 1.0 1.4 2.1</td>
<td>5.3 0.5 1.0 1.4 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty getting around</td>
<td></td>
<td>2.7 0.7 1.3 1.7 2.3</td>
<td>2.8 0.7 1.3 1.7 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Everything takes longer</td>
<td></td>
<td>*2.8 –0.1 0.8 1.2 1.9</td>
<td>3.1 –0.1 0.7 1.2 1.9</td>
<td>2.8 –0.1 0.7 1.2 1.9</td>
<td></td>
</tr>
<tr>
<td>Difficulty managing job</td>
<td></td>
<td>3.3 0.5 1.1 1.4 1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Appearance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease affect appearance</td>
<td></td>
<td>2.9 –0.3 0.5 1.1 1.9</td>
<td></td>
<td>2.4 –0.3 0.5 1.2 2.0</td>
<td></td>
</tr>
<tr>
<td>Unsatisfied with appearance</td>
<td></td>
<td>7.1 0.2 0.7 1.1 1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mask visible signs</td>
<td></td>
<td>2.2 1.2 1.6 2.0 2.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bothered by people looking</td>
<td></td>
<td>2.6 1.3 1.8 2.2 2.5</td>
<td>2.6 1.3 1.8 2.2 2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Influence on clothes worn</td>
<td></td>
<td>2.3 1.0 1.4 1.8 2.6</td>
<td>2.5 0.9 1.4 1.8 2.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Items flagged for nonunidimensionality in previous confirmatory factor analyses were omitted (+).

+Items with item level S-X2 misfit p-value<0.01.

IRT, item response theory.
among stable patients responding twice to the ThyPRO (17), using intraclass correlation (42,47–49). Confidence intervals (CI) were estimated by empirical bootstrap (43,50).

**Statistical analysis**

Descriptive analyses, summed-score linking, sensitivity tests, responsiveness comparisons, DIF, and test–retest intraclass correlations were performed with SAS v9.3 (51). Bifactor models were estimated with Mplus v7.1 (52). IRT modeling was performed with IRTPRO (53,54). DIF was evaluated using the ordinal logistic regression approach (55).

**Study approval and conduct**

According to Danish law, PRO research does not require and thus cannot obtain approval by ethical committees, and a completed questionnaire is regarded as consent. The study was approved by the Danish Data Protection Agency (#2007-58-0015) and conducted in accordance with the Declarations of Helsinki.

**Results**

Item parameters from the IRT modeling are shown in Table 2. The Impaired Social Life scale only included three unidimensional items, and was thus not abbreviated further. In the original long-form scales—Goiter Symptoms, Hyperthyroid Symptoms, Cognitive Complaints, Anxiety, Social Life, and Appearance—all items had good fit to the IRT model. For the Eye Symptoms, Tiredness, Depressivity, Emotional Susceptibility, and the Daily Life scales, item-level misfit was eliminated through reduction of the scales (Table 2, middle column). Abbreviated scales with three items each were obtained for 10/11 scales, whereas the Hyperthyroid Symptoms scale had four items. The short form is presented in Supplementary Appendix S2.

With specification of a method factor comprising the positively worded items, a bifactor model with acceptable overall fit (comparative fit index [CFI] = 0.97, root mean square error of approximation [RMSEA] = 0.085) was fitted to the composite score (Fig. 1). One item concerning memory problems had higher loading on its subfactor; all other items had higher loading on the general factor.

The scale transformations are presented for each scale in Table 3. Thus, a raw sum score of 0 on, for example, the Goiter Symptoms scale (answer “not at all” to all three items) should be rescaled to a value of 2, and a raw sum score of 12 (answer “very much” to all three items) should be rescaled to 84, which is the maximum score on the Goiter short-form scale.

Agreement plots of the new, rescaled short-form scales versus the original long-form scales are presented in Figure 2. Good and uniform agreement was shown across the entire range of scores. In each plot, the mean CI long-form score and the mean short-form score for the cross-sectional sample are provided too. Only the Tiredness short-form mean score was outside the CI for the long-form scale. However, the difference between the two mean levels was only three points on the 0–100 scale.

Effect sizes and responsiveness in groups of patients undergoing treatment were preserved in the short-form scales, as shown in Table 4. Only the short-form Anxiety scale had smaller effect size and slightly lower responsiveness than the original long-form. Conversely, the short-form Appearance scale had slightly higher effect size and responsiveness compared to the long form.

Test–retest reliability was similarly preserved in the short-form scales, with only the Appearance scale having significantly but marginally lower reliability (Table 4).

Very high short- versus long-form intraclass correlations were found (0.89–0.98; Table 4). Further, the previously shown discriminant validity was reproduced. Thus, for all

<table>
<thead>
<tr>
<th>Raw sum score</th>
<th>Goiter</th>
<th>Hyper</th>
<th>Eye</th>
<th>Tired</th>
<th>Cognition</th>
<th>Anxiety</th>
<th>Depressivity</th>
<th>Susceptibility</th>
<th>Social Life</th>
<th>Daily Life</th>
<th>Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>13</td>
<td>14</td>
<td>17</td>
<td>14</td>
<td>18</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>25</td>
<td>21</td>
<td>26</td>
<td>22</td>
<td>21</td>
<td>25</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>23</td>
<td>25</td>
<td>33</td>
<td>29</td>
<td>34</td>
<td>29</td>
<td>28</td>
<td>33</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>28</td>
<td>32</td>
<td>42</td>
<td>37</td>
<td>41</td>
<td>37</td>
<td>36</td>
<td>42</td>
<td>38</td>
<td>43</td>
</tr>
<tr>
<td>6</td>
<td>37</td>
<td>33</td>
<td>38</td>
<td>50</td>
<td>44</td>
<td>49</td>
<td>45</td>
<td>44</td>
<td>50</td>
<td>46</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>38</td>
<td>45</td>
<td>58</td>
<td>52</td>
<td>56</td>
<td>54</td>
<td>52</td>
<td>58</td>
<td>54</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td>49</td>
<td>44</td>
<td>52</td>
<td>67</td>
<td>60</td>
<td>63</td>
<td>63</td>
<td>60</td>
<td>67</td>
<td>62</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>57</td>
<td>49</td>
<td>60</td>
<td>75</td>
<td>68</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>68</td>
<td>75</td>
<td>71</td>
</tr>
<tr>
<td>10</td>
<td>64</td>
<td>55</td>
<td>68</td>
<td>83</td>
<td>76</td>
<td>79</td>
<td>80</td>
<td>77</td>
<td>83</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>11</td>
<td>73</td>
<td>60</td>
<td>78</td>
<td>92</td>
<td>85</td>
<td>87</td>
<td>89</td>
<td>86</td>
<td>92</td>
<td>89</td>
<td>87</td>
</tr>
<tr>
<td>12</td>
<td>84</td>
<td>66</td>
<td>89</td>
<td>100</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>95</td>
<td>100</td>
<td>98</td>
<td>96</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The raw score (left column) is derived by simple summation of item values (0–4) separately for each scale. The corresponding final, IRT-linked rescaled short-form score is tabulated for each scale. Thus, when scoring short-form scales, items are summed (0–12 for 3-item scales and 0–16 for 4-item scale), and the rescaled score is derived from the table.
short-form scales, the group expected to score higher, as specified in Watt et al. (17), did have significantly higher mean scores than the group expected to score low (data not shown). Responsiveness in the three diagnostic groups evaluated was identical to that demonstrated for the long form (26). DIF according to age in the short form were also identical to the small effects previously identified in the long form (35). For a given level of Tiredness and Depressivity, respectively, younger patients had a tendency to endorse positively worded items ("felt energetic" and "self-confident," respectively). For

FIG. 2. Agreement plots of short-form (horizontal axis) versus original long-form (vertical axis) scale levels with regression lines. For each scale, the mean score (95% CI) for the long form is presented in the upper-left corner, and the mean score for the short form in the lower right. Short-form mean outside long form is marked by an asterisk.
Table 4: Responsiveness and Test–Retest Reliability of the Short-Form Versus the Long-Form Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Effect size short scale</th>
<th>F short</th>
<th>Test–retest reliability short scale</th>
<th>Effect size long scale</th>
<th>F long</th>
<th>Test–retest reliability long scale</th>
<th>Relative validity short scale</th>
<th>Relative validity long scale</th>
<th>Intraclass correlation short scale</th>
<th>Intraclass correlation long scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperthyroid Symptoms</td>
<td>0.52</td>
<td>0.49</td>
<td>0.94 [0.84–1.06]</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Tiredness</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Cognitive Complaints</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.52</td>
<td>0.46</td>
<td>0.90 [0.79–1.02]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Depressivity</td>
<td>0.25</td>
<td>0.19</td>
<td>0.94 [0.84–1.06]</td>
<td>0.25</td>
<td>0.19</td>
<td>0.94 [0.84–1.06]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Emotional Susceptibility</td>
<td>0.32</td>
<td>0.27</td>
<td>0.94 [0.84–1.06]</td>
<td>0.32</td>
<td>0.27</td>
<td>0.94 [0.84–1.06]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Impaired Social Life</td>
<td>0.22</td>
<td>0.17</td>
<td>0.94 [0.84–1.06]</td>
<td>0.22</td>
<td>0.17</td>
<td>0.94 [0.84–1.06]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Impaired Daily Life</td>
<td>0.37</td>
<td>0.30</td>
<td>0.94 [0.84–1.06]</td>
<td>0.37</td>
<td>0.30</td>
<td>0.94 [0.84–1.06]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
<tr>
<td>Appearance</td>
<td>0.19</td>
<td>0.15</td>
<td>0.94 [0.84–1.06]</td>
<td>0.19</td>
<td>0.15</td>
<td>0.94 [0.84–1.06]</td>
<td>0.98 [0.96–1.04]</td>
<td>0.90 [0.84–0.94]</td>
<td>0.84 [0.77–0.93]</td>
<td>0.88 [0.79–0.91]</td>
</tr>
</tbody>
</table>

For each scale, the effect size for the new short form is calculated and the relative validity index calculated as $F_{short}/F_{long}$, i.e., the version with the best responsiveness has a relative validity index of 1. In addition, the test–retest reliability for the short form is presented to the right.

Discussion

The purpose of the present study was to develop an abbreviated version of the ThyPRO. This goal was achieved successfully. Based on previous validation studies and IRT modeling, an abbreviated version of the ThyPRO was developed containing (a) four physical symptom scales, two with three items (Goiter and Eye Symptoms) and two with four items (Hypo- and Hyperthyroid physical symptoms); (b) seven three-item scales about physical, mental, and social well-being and function; (c) one three-item scale concerning appearance; and (d) one single item about impact on overall QoL. Thus, the abbreviated version consists of 39 items, if all physical symptom scales are administered. Each of the 12 short-form scales and the single QoL item can be reported separately, but the seven well-being and function scales can also be summarized in one single composite score.

The validation analyses showed that the abbreviated scales had very high agreement with the original long-form ones, including roughly similar mean levels, and comparable measurement quality. Thus, good test–retest reliability, responsiveness to clinical change, and sensitivity to relevant clinical differences were demonstrated. This preservation of good measurement properties in scales with much fewer items is interpreted as being a result of selection of items with best measurement properties, under consideration of the conceptual model and content validity, thereby reducing random and systematic measurement error.

The primary strength of this study is the integration of several studies and methodologies in the item reduction process. Thus, several modalities within modern psychometrics (DIF, structural equation modeling for ordinal data, item response modeling) were applied within a firm clinical framework among patients from several clinical studies, including cross-cultural samples. Further, analyses were conducted within the original ThyQoL conceptual model (56) with focus on content validity. However, the final short-form has not been tested as a stand-alone form in an independent, novel clinical sample. This should be considered in future studies. Further, although the aim was to develop and test the instrument in a broad, heterogeneous sample, as specified in the introduction, and although the cross-sectional sample size was fairly large, it was not large enough to permit multigroup analyses (57) according to diagnosis. However, previous studies using an ordinal regression approach (55), less dependent on sample size, have shown only minimal DIF of the ThyPRO scales, according to diagnosis (35). Application of a short form may lead to loss of content validity. The extent to which this has occurred can only be evaluated in qualitative studies (58). However, since previous validation studies have confirmed that the individual scales are unidimensional, the potential loss should theoretically be minimal. As evident from the agreement plots, the short-form scales have fewer measurement points along the entire spectrum, and application of the short forms may also lead to poorer discrimination at the extremes. Another potential weakness was the fact that five of the scales were slightly modified (some items were omitted) to avoid item-level misfit in the IRT model. Since
the rescaling was based on these IRT analyses, this may lead to weaker linking between the two versions. On the other hand, as mentioned, the correlation, agreement and mean levels between the two versions of each scale all supported the appropriateness of the present linking.

The applied approach is in line with recent recommendations for item reduction (25). When reviewing the available item reduction literature, the authors found that 55% of the studies had preserved scale structure and the median proportion of reduction was 57% (range 21–88%). The present study is close to this median reduction (from 85 to 39 items, i.e., 54% reduction). In 62% of the studies, only the long form was administered. Use of IRT methods was recommended as advantageous in the suggested guidelines, but was only applied in 11% of the studies.

The two-level scale-scoring approach, where both a composite score and the underlying more detailed subscales can be scored for the well-being and function scales, has also been adopted in previous studies. The most prominent is the most widely used short-form measure, the SF-36 Health Survey (59). Based on SF-36, eight domain scores as well as two Component Summaries can be derived (60), depending on the level of detail required in reporting. The scoring of the SF-36 summaries is based on results from principal component analyses, in contrast to the present study, where a simple summation approach was adopted for ease of scoring and reporting.

The short-form offers an advantageous measure, when reduction of respondent burden, potentially increasing response rates, is considered to outweigh the theoretical reduction of content validity and measurement detail. In an ongoing clinical trial among patients with Graves’ disease (19,61), time to completion of 39 ThyPRO items was short (median 4 minutes; interquartile range 3–5 minutes), according to time-stamped electronic responses. This may be particularly relevant in longitudinal studies with multiple measurement points and when studying larger samples. Reporting the well-being and function scales as the composite score is recommended, when simplicity of reporting, combined with small measurement intervals and high precision, is the primary goal. When a detailed evaluation of physical, emotional, and social well-being and function is warranted, reporting the individual scales is recommended. In general, administering ThyPRO-39 as a whole is recommended to enhance content validity and comparability. However, it may be relevant to omit entire scales (not individual items), if considered irrelevant for a specific future study. For example, the Eye Symptoms scale may not be administered in a trial among patients with nontoxic goiter. Similarly, scales from the full-length ThyPRO may be selectively added to the ThyPRO-39, if considered of particular importance, for example the Sex Life scale, which is not included in ThyPRO-39.

ThyPRO-39 can be implemented in daily clinical practice (62). Patients may respond to the instrument, for example, prior to their appointment, either from home via an e-mail sent in advance, or in the waiting room. Scale scores might then be transferred to the electronic medical record and evaluated by the clinician, similarly to evaluation of, for example, thyroid function tests. These data can then be used for monitoring and communication purposes. Relevant problems (or lack thereof) may be rapidly identified and addressed, ideally with established thresholds and recommended actions and interventions, including referral to psychosocial intervention. For an ongoing clinical trial, real-time automatic monitoring of responses to ThyPRO has been implemented (61). In response to scores above (i.e., worse than) preset thresholds, e-mail alerts are generated and sent to clinical staff. A similar system could monitor responses in clinical practice. However, there is still a requirement for further research on this, for example evaluating how to communicate these results meaningfully to patients; establishment of alert thresholds; identification and evaluation of effectiveness of relevant interventions, among others.

In conclusion, this study has developed an abbreviated 39-item version of the thyroid-related QoL measurement instrument ThyPRO, with good measurement properties. It has high agreement with the long-form original version, and score levels on one form are comparable to score levels on the other. Function and well-being may be reported as a composite score or as individual scale scores. This abbreviated version, named the ThyPRO-39, is recommended for use in clinical studies, as a possible alternative to the original version. Scoring programs for use on various platforms are available from the first author for both versions.

Acknowledgments

This study has been supported by grants from the Danish Agency for Science, Technology, and Innovation: Council for Strategic Research and Council for Independent Research, Genzyme Corporation, Novo Nordisk Foundation by unrestricted research grants and Agnes and Knut Mørk’s Foundation. Warm thanks to Sofie Larsen Rasmussen and Lene Frydenberg for help with identification and inclusion of patients; to Sofie Larsen Rasmussen and Kira Bang Bové for help with follow-up; to Sofie Larsen Rasmussen, Kim Åbels, Emilie Birch, Thea Christoffersen, and Sara Kehlet Watt for help with generation of clinical data; and to Selma Watt and Laura Siim Magnussen for data entry. Researchers interested in using the ThyPRO or the ThyPRO-39 may contact the corresponding author.

Author Disclosure Statement

None of the authors has any financial conflicts of interest to declare.

References


31. Orlando M, Thissen D 2003 Further investigation of the performance of S-X 2: an item fit index for use with
38. McDonald RP 1999 Test Theory. A Unified Treatment. Lawrence Erlbaum, Mahwah, NJ.