Quality of Life in European Patients with Addison's Disease: Validity of the Disease-Specific Questionnaire AddiQoL

Marianne Øksnes, Sophie Bensing, Anna-Lena Hulting, Olle Kämpe, Annika Hackemann, Gesine Meyer, Klaus Badenhoop, Corrado Betterle, Anna Parolo, Roberta Giordano, Alberto Falorni, Lucyna Papierska, Wojciech Jeske, Anna A. Kasperlik-Zaluska, V. Krishna K. Chatterjee, Eystein S. Husebye, and Kristian Løvås

Context: Patients with Addison's disease (AD) self-report impairment in specific dimensions on well-being questionnaires. An AD-specific quality-of-life questionnaire (AddiQoL) was developed to aid evaluation of patients.

Objective: We aimed to translate and determine construct validity, reliability, and concurrent validity of the AddiQoL questionnaire.

Methods: After translation, the final versions were tested in AD patients from Norway (n = 107), Sweden (n = 101), Italy (n = 165), Germany (n = 200), and Poland (n = 50). Construct validity was examined by exploratory factor analysis and Rasch analysis, aiming at unidimensionality and fit to the Rasch model. Reliability was determined by Cronbach's coefficient-α and Person separation index. Longitudinal reliability was tested by differential item functioning in stable patient subgroups. Concurrent validity was examined in Norwegian (n = 101) and Swedish (n = 107) patients.

Results: Exploratory factor analysis and Rasch analysis identified six items with poor psychometric properties. The 30 remaining items fitted the Rasch model and proved unidimensional, supported by appropriate item and person fit residuals and a nonsignificant χ^2 probability. Crohnbach's α-coefficient 0.93 and Person separation index 0.86 indicate high reliability. Longitudinal reliability was excellent. Correlation with Short Form-36 and Psychological General Well-Being Index scores was high. A shorter subscale comprising eight items also proved valid and reliable. Testing of AddiQoL-30 in this large patient cohort showed significantly worse scores with increasing age and in women compared with men but no difference between patients with isolated AD and those with concomitant diseases.

Conclusion: The validation process resulted in a revised 30-item AddiQoL questionnaire and an eight-item AddiQoL short version with good psychometric properties and high reliability. ***(J Clin Endocrinol Metab* 97: 568–576, 2012)**

Primary adrenal insufficiency [Addison's disease (AD)] is a rare chronic disease, treated with glucocorticoid and mineralocorticoid replacement (1); additional replacement of the adrenal androgen dehydroepiandrosterone is debated (2–5). Novel treatment strategies such as modified-release hydrocortisone tablets or continuous sc hydrocortisone infusion are under investigation (6–9). There is no gold standard for assessment of treatment, but

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in U.S.A.

doi: 10.1210/jc.2011-1901 Received June 30, 2011. Accepted October 17, 2011.
First Published Online November 16, 2011

* Author affiliations are shown at the bottom of the next page.

Abbreviations: AcroQoL, Acromegaly Quality of Life Questionnaire; AD, Addison's disease; AddiQoL, AD-specific quality-of-life questionnaire; AGHDA, Assessment of Growth Hormone Deficiency in Adults; CushingQoL, disease-specific questionnaire for evaluating QoL in Cushings syndrome; DIF, differential item functioning; EFA, exploratory factor analysis; HRQoL, Health-Related Quality of Life; PGWB, Psychological General Well-Being Index; PSI, person separation index; SF-36, Short-Form-36.

a clinical scoring system has been proposed (10). Patient surveys reproducibly report impairment in particular dimensions of general well-being questionnaires (3, 11, 12). Generic questionnaires have been applied to study differences between subgroups of patients with AD (13, 14) and other autoimmune endocrinopathies (15). However, questionnaires containing disease-specific items are likely to be more sensitive to effects that clinicians wish to monitor (16). Recently, the disease-specific quality-of-life questionnaire (AddiQoL) was developed as an evaluative tool in AD (17), which might facilitate the detection of changes in well-being in future clinical trials and during regular follow-up of patients.

Validity is the process of demonstrating that an instrument quantifies what it seeks to measure and that it is useful for this purpose. Construct validity aligns a questionnaire to a theorized underlying trait and involves testing of correlation between the items. Here we used Rasch analysis to explore the psychometric properties of AddiQoL. Rasch analysis is a mathematical item response model increasingly used in somatic medicine and endocrinology and has been used in validation of Quality of Life-Assessment of Growth Hormone Deficiency in Adults (AGHDA) (18), Acromegaly Quality of Life Questionnaire (AcroQoL) (19), and the disease-specific questionnaire for evaluating QoL in Cushings syndrome (CushingQoL) (20). The objective is to test how well the observed data fit with the expectations of the mathematic measurement model (21, 22).

Reliability implies the degree to which an instrument is free from random error. The traditional reliability coefficient (Cronbach's-α) indicates how well an individual item correlates with the other items in a questionnaire. In Rasch analysis, the person separation index (PSI) is equivalent to Crohnbach's-α and represents the power of the construct to discriminate between respondents, giving an indication of how precisely patients have been spread out along the continuum (23). Test-retest reliability or repeatability is the correlation between scores from the same individual assessed on two separate occasions, given that their clinical condition is stable.

Concurrent validity measures how the questionnaire performs against a gold standard instrument, usually by exploring correlation of questionnaire scores. The AddiQoL is, to our knowledge, the first disease-specific Health-Related Quality of Life (HRQoL) questionnaire in AD, such that no gold standard exists. In other endocrine disorders, the Short-Form-36 (SF-36) and the Psychological General Well-Being Index (PGWB) have been used to validate the AGHDA (24, 25), the AcroQoL (26), and the CushingQoL questionnaires (20).

Validation of a questionnaire requires responses from a large number of subjects, which is difficult to achieve from a single country, for a rare disorder such as AD. Hence, in the current study, we translated the original English AddiQoL into Norwegian, Swedish, Italian, German, and Polish versions; these were administered to large cohorts of patients with AD in each country for evaluation of construct validity and reliability. Test-retest reliability was tested in patient subgroups in Norway, Italy, and Sweden. Concurrent validity was investigated by examining correlation between the AddiQoL scores and results of simultaneously administered SF-36 and PGWB questionnaires in Norway and Sweden. We also sampled AddiQoL data from a random population group in Norway. The final revised questionnaire was ultimately used to assess HRQoL in different subgroups of this large cohort.

Materials and Methods

Design and subjects

First, the AddiQoL was translated from English into Norwegian, Swedish, German, Polish, and Italian versions, following international recommendations (27). Second, patients with verified AD were recruited from patient registries or consecutively from outpatient clinics. The diagnostic criteria for inclusion in the European AD database (Euradrenal) are one or more of the following: 1) low serum cortisol and high ACTH, 2) positive ACTH stimulation test, and/or 3) chronic replacement therapy with glucocorticoids and fludrocortisone. The patients received an invitation letter containing study information and AddiQoL; by returning the precoded questionnaires, they were included in the study. In addition, patients in Norway and Sweden received the SF-36 and the PGWB for analysis of concurrent validity. There were no exclusion criteria. The AddiQoL subject codes were used to retrieve patient characteristics such as age, sex, and concurrent autoimmune diseases from registries or via an additional precoded registration form. For analysis of longitudinal reliability, a subgroup of at least 20 clinically stable patients from Norway, Sweden, and Italy completed AddiQoL a second time, 2–6 wk after the first questionnaire. A random sample of 2000 persons with even sex, age, and geographical distribution was drawn from the Norwegian People Registry. They received a letter with invitation to participate as control

Institute of Medicine (M.Ø., K.L., E.S.H.), University of Bergen, and Department of Medicine (K.L., E.S.H.), Haukeland University Hospital, N-5021 Bergen, Norway; Department of Molecular Medicine and Surgery (A.L.H., S.B.), Karolinska Institutet, SE-171 77 Stockholm, Sweden; Department of Medical Sciences (S.B., O.K.), Uppsala University, SE-751 05 Uppsala, Sweden; Division Endocrinology and Diabetology (K.B., A.H., G.M.), Medical Department 1, University Hospital of the Goethe-University Frankfurt, 60325 Frankfurt am Main, Germany; Endocrine Unit (C.B., A.P.), Department of Medical and Surgical Sciences, University of Padua, 35122 Padua, Italy; Department of Clinical and Biological Sciences (R.G.), University of Turin, 10126 Turin, Italy; Department of Internal Medicine (A.F.), University of Perugia, 06126 Perugia, Italy; Endocrinology Clinic (A.A.K.-Z., L.P., W.J.), The Medical Centre of Postgraduate Education, Ceglowska 80, Warsaw, Poland; and Institute of Metabolic Science (V.K.K.C.), University of Cambridge, Cambridge CB20QQ, United Kingdom

subjects by returning the completed anonymized AddiQoL questionnaire, with registration of age and sex only. Third, the responses from the patients were analyzed by exploratory factor analysis (EFA) and Rasch analysis for assessment of validity and reliability and hence for amendment of the questionnaire. The study was approved by regional ethics committees in each country.

Translation

The forward translation was performed by a minimum of three native speakers of the target language, who had good knowledge of English. Translations were performed locally by the study group in each country. Their preliminary versions were discussed locally by a panel of experts, *i.e.* clinicians in endocrinology, agreeing on versions to be evaluated further. For quality control, these versions were assessed by two professional translators (Lionbridge Technologies Inc., Waltham, MA), who evaluated the conceptual equivalence with the original, clarity and use of a familiar register. Thus, two adjusted versions of each AddiQoL translation were generated, which were reevaluated by the study groups in each country, who decided on a final version.

Questionnaires

The original AddiQoL is a 36-item questionnaire; each item contains six scoring categories. Twenty-five items are negative HRQoL statements that need to be reversed for questionnaire scoring; thus, a higher score indicates a higher level of HRQoL. The questionnaire was developed in the English language and initial statistical analysis performed in 85 patients form the United Kingdom (17). The SF-36 is a generic HRQoL questionnaire, widely used and thoroughly validated (28). The SF-36 is translated into many languages and has been used in previous studies of HRQoL in AD (11, 12). The PGWB is a validated 22-item generic HRQoL questionnaire that has been translated into several languages, intended to measure the subjective feeling of psychological well-being (29).

Construct validity and Rasch analysis

The Rasch model rests on the idea that useful measurement involves examination of only one human attribute at a time (unidimensionality). The model allows quantitative assessment (additivity of items) from data that are ordinal, based on logistic transformation of the item responses (22). First, EFA was used to examine dimensionality of AddiQoL. Second, we applied Rasch analysis [RUMM 2020 software (30)] to further identify and eliminate items with poor fit to the Rasch model (31).

Overall fit statistics include item-person interaction statistics, calculated as mean item location and mean person location. Both person fit and item fit is transformed by RUMM to approximate a Z-score; this represents a standardized normal distribution. Therefore, if the items and persons fit the model perfectly, the mean fit residual is expected to be zero with SD around 1. Overall fit statistics also includes χ^2 statistics for item-trait test of fit to the model. This tests whether the items work as expected at group level along the range of the scale. A nonsignificant χ^2 probability implies that the hierarchical ordering of items and persons do not vary across the range of the scale.

Generally, any item with a fit residual greater than ± 2.5 is a cause for concern; a high positive item fit residual indicates that the item does not separate well between high and low person ability, and a high negative item fit residual indicates redundancy or local dependency (see below) of the item. Other causes of misfit to the Rasch model are disordered thresholds and differential item functioning (DIF; item bias). A threshold is the point at which the probability of endorsing two neighboring response alternatives is equal; one threshold exists for each transition between one scoring alternative to the next. To obtain ordered thresholds, each item and response alternative was assessed and collapsed or rescored when necessary. DIF analysis explores item performance and instrument performance across different patient groups. DIF exists if one patient group scores significantly different on an item compared with another patient group with similar overall HRQoL level (21). Here we performed DIF analysis for patient sex, age, concurrent disease, country, and time point (for test-retest reliability analysis).

Each item's difficulty (item location) and each person's ability (person location) are organized in ordered hierarchies (22, 32). By plotting item location and person location on the same scale, the targeting of the items to the sample population can be explored. A perfect targeting is indicated if average person location is zero.

Fit to the model and an absence of a significant pattern among the fit residuals supports the scale being unidimensional, *i.e.* there is only one concept being measured. Local dependency exists when there is covariance between the response patterns of items, and this is considered a breach of the strict unidimensionality that the Rasch model requires. This can be corrected for by grouping items with covariance together, *i.e.* treating the item group mathematically as a single combined item (33).

Ultimately, if fit to the model and unidimensionality are present, the individual Person location can then be used as a psychometrically valid total score.

Reliability

PSI is calculated as the ratio of true variance to observed variance and represents the proportion of variance that is not due to error (23, 32). A PSI of 0.85 or more is generally required if the scale is to be used on the individual level. For longitudinal reliability, a test-retest DIF analysis was performed for patient subgroups with stable clinical condition in Norway, Italy, and Sweden over 2- to 6-wk intervals.

Concurrent validity and normative data

AddiQoL scores were compared with SF-36 scores and PGWB score in Norway and Sweden. Spearman's rho with two-tailed significance was calculated for the correlation analysis. A Mann-Whitney *U* test was used to compare Norwegian patients' AddiQoL scores with AddiQoL results from a random Norwegian population sample. However, because the items are optimized for patients with AD, the comparison of responses between patients and healthy controls must be interpreted with caution. Comparison of AddiQoL-30 and the subset AddiQoL-8 (see below) scores in different subgroups of patients was performed by multiple linear regression analysis with sex, age, country, and comorbidity as independent variables.

Results

Subjects

A total of 615 patients were recruited from Norway (n = 107), Italy (n = 157), Germany (n = 200), Sweden

TABLE 1. Patient characteristics

	Norway (n = 107)	Italy (n = 157)[a]	Germany (n = 200)	Sweden (n = 101)	Poland (n = 50)[b]
Sex					
Male (%)	39 (36.4)	54 (34.8)	53 (26.5)	36 (35.6)	10 (20)
Female (%)	68 (63.6)	101 (65.2)	147 (73.5)	65 (64.4)	40 (80)
Age (yr)					
18–29 (%)	6 (5.6)	16 (10.5)	15 (7.5)	6 (5.9)	NA
30–39 (%)	17 (15.9)	42 (27.5)	35 (17.5)	18 (17.8)	NA
40–49 (%)	29 (27.1)	43 (28.1)	58 (29.0)	22 (21.8)	NA
50–59 (%)	32 (29.9)	27 (17.6)	42 (21)	27 (26.7)	NA
60–69 (%)	22 (20.6)	17 (11.1)	29 (14.5)	17 (16.8)	NA
>70 (%)	1 (0.9)	8 (5.2)	21 (10.5)	11 (10.9)	NA
Comorbidity (autoimmune)					
Present (%)	70 (65.4)	106 (68.4)	153 (76.5)	60 (59.4)	NA
Thyroid disease (%)	55 (50.9)	83 (52.5)	125 (62.5)	45 (44.6)	NA
Type 1 diabetes (%)	13 (12.0)	6 (3.8)	20 (10)	9 (8.9)	NA
Other (%)[c]	33 (30.6)	21 (13.4)	62 (31)	25 (24.8)	NA

NA, Not available.

[a] For Italy, 16 patients were classified as Autoimmune Polyendocrine Syndrome type 2 (APS2); these are included in comorbidities but excluded from the thyroid and diabetes numbers.

[b] Data regarding age and comorbidity are missing from Poland and from a few Italian patients (sex, n = 2; age, n = 4; and comorbidities, n = 2).

[c] Includes celiac disease, hypoparathyroid disease, pernicious anemia, and primary ovarian failure.

(n = 101), and Poland (n = 50). The Polish data were omitted from some of the analyses due to low numbers and missing data. The original U.K. data were also included in the pooled data analysis. Patient characteristics are presented in Table 1. Information on patient comorbidities was available for Norway, Italy, Germany, and Sweden. Test-retests were available from clinically stable patients, on appropriate substitution therapy, in Norway (n = 37), Italy (n = 25), and Sweden (n = 29). SF-36 and PGWB scorings were available from Norway (n = 107) and Sweden (n = 101), in which the patient response rates were 65 and 80%, respectively. In the Norwegian normative sample, 539 of 2000 persons responded, producing a response rate of 28%. Of the respondents 54% (283) were female and 56% (300) were below the age of 50 yr (18–39 yr, n = 166; 40–49 yr, n = 134; 50–59 yr, n = 123; >60 yr, n = 107).

Translation and quality of questionnaire responses

Overall, the evaluations from the professional translators were favorable; mostly minor errors were noted. The item "I feel lightheaded" proved difficult to translate. Overall, the rate of missing responses was below 1%. There was a tendency toward missing responses on page 2 of the questionnaire. Items regarding sexuality showed the most missing responses, *i.e.* "I am satisfied with my sex-life," 7%, and "I have lost interest in sex," 4.9%. This was not evenly distributed among countries because 22% of patients in Poland did not score the former, whereas this figure for Norway was only 0.9% (United Kingdom, 7%; Italy, 7.3%; Germany, 5.5%; and Sweden, 6.9%).

Construct validity

Initially, the 36-item questionnaire showed misfit to the Rasch model. The EFA identified four subdimensions of AddiQoL, which we denoted fatigue (8 items), emotions (8 items), symptoms (11 items), and miscellaneous (sleep, sexuality, and impact of intercurrent disease, six items) (for item overview, see Supplemental Table 1, published on The Endocrine Society's Journals Online web site at http://jcem.endojournals.org). The three items, nocturia, dry skin, and gaining weight, did not belong to any subdimension and had poor discriminating properties in the Rasch analysis. Further analysis revealed that the fatigue domain fitted the model and achieved unidimensionality in all countries, and no significant DIF was present for sex, age, or comorbidity. Therefore, the fatigue subscale was further tested as a potential AddiQoL short version (AddiQoL-8). The emotions subdimension showed overall good fit to the Rasch model, but there was multidimensionality in the Norwegian data. The item, "Emotional stress makes me exhausted," had high residual correlation with the item, "I cope well in emotional situations," and displayed DIF in the Swedish data; hence, the former item was discarded. In the symptoms subdimension, the item, "I have salt cravings," displayed misfit in nearly all countries (Fig. 1A), although certainly a clinically relevant item. Elimination of this item improved fit to the model, but χ^2 probability remained significant for symptoms in the Italian and German data. The miscellaneous subdimension showed overall good fit and unidimensionality, but the item, "I have lost interest in sex,"

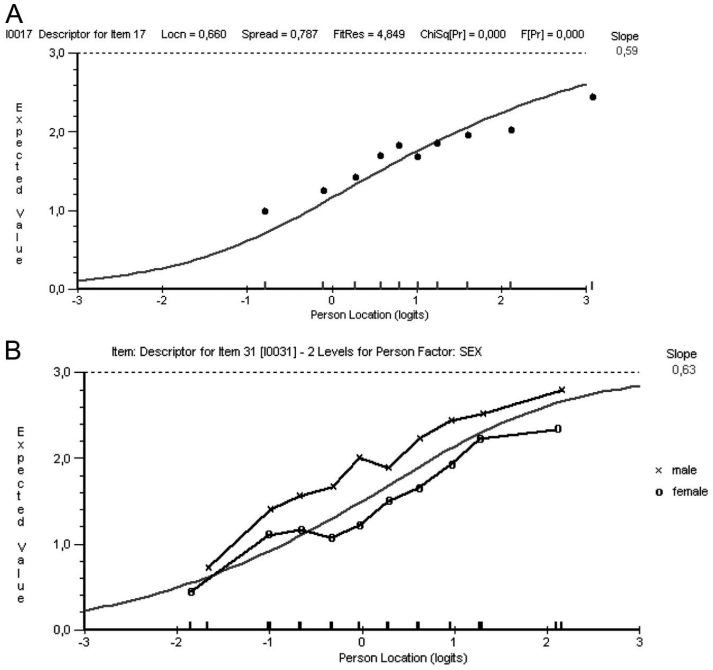

FIG. 1. A, Fit to the Rasch model for the item, "I have salt cravings." The *gray line* depicts the expected scoring pattern as estimated by the model. The *black dots* are actual scoring from groups of patients with similar HRQoL levels (class intervals). The patients with the highest HRQoL level (*far right*) scores less than expected, the patients with the lowest HRQoL level (*far left*) scores better than expected, *i.e.* this item does not separate well between high and low HRQoL. B, DIF sex for the item, "I have lost interest in sex." The *gray line* depicts the expected scoring pattern estimated from the Rasch model. Men score better than expected, females worse ($P < 0.01$). This item showed similar results for DIF age. Patients younger than 50 yr scored worse than expected, patients older than 50 yr, better than expected ($P < 0.01$).

displayed DIF by sex in the Norwegian, Italian, and German data and DIF by age in the Italian and German data (Fig. 1B). Removal of this item improved fit.

Disordered thresholds were present for many items, indicating that the subjects had difficulties differentiating between some response alternatives. We found that rescoring the original six response alternatives (123456) to four (122334) by collapsing the scoring categories "a little of the time"/"some of the time," "a good bit of the time"/"most of the time," "agree"/"slightly agree," and "disagree"/"slightly disagree" improved fit and produced ordered thresholds.

The 30 remaining items, rearranged in the four revised subdimensions as superitems, fitted the Rasch model; this was supported by a nonsignificant item-trait interaction ($\chi^2 = 0.56$) in the pooled data. Also, this item solution proved unidimensional. Supplemental Table 2 displays the overall fit statistics for the pooled data and for individual countries. There was no significant DIF between the genders and no DIF when comparing the results from patients with isolated AD with patients with autoimmune polyendocrine syndromes. Significant DIF for age was present in the emotions subdimension in the Swedish and the pooled data. Significant DIF for country was present in the fatigue, the symptom, and the miscellaneous subdimensions. Based on these results, we elected to go ahead with validation of the revised 30-item questionnaire (AddiQoL-30) and the eight-item subset fatigue (AddiQol-8).

Targeting of the items to the total patient population is shown in Fig. 2. Mean person location was 0.21, indicating that mean patient score was slightly higher than the HRQoL level targeted by the mean of the items (set at zero). Mean person location for individual countries was −0.04 (Poland), 0.09 (Germany), 0.14 (United Kingdom), 0.21 (Italy), 0.23 (Norway), and 0.27 (Sweden), indicating good targeting in all countries.

Reliability

AddiQol-30 demonstrated good reliability as indicated by Cronbach's-α 0.93 and PSI 0.86. PSI for individual countries are presented in Supplemental Table 2. For the AddiQoL-8, PSI ranged from 0.89 to 0.91 in individual countries, indicating excellent reliability as a separate scale. A total of 91 clinically stable patients from Norway, Sweden, and Italy performed test-retest 2–6 wk after the first evaluation. Longitudinal reliability was excellent because no significant DIF between separate time points was detected.

Concurrent validity, patient scores, and normative data

Rasch-transformed AddiQoL-30 and AddiQoL-8 scores were compared with SF-36 scores and PGWB scores in the Norwegian (n = 107) and Swedish (n = 101) patients. Results from the correlation analyses are given in Table 2. Rasch-transformed scores from all countries are shown in Fig. 3. Regression analysis adjusting for age, sex, country, and comorbidity showed that women scored significantly worse than men (AddiQoL-30, $P < 0.001$; AddiQoL-8, $P = 0.001$) and demonstrated worse scores with increasing age (AddiQoL-30, $P <$ 0,001; AddiQoL-8, $P =$ 0.001). No statistical difference was found between patients with isolated AD and those with autoimmune poly-

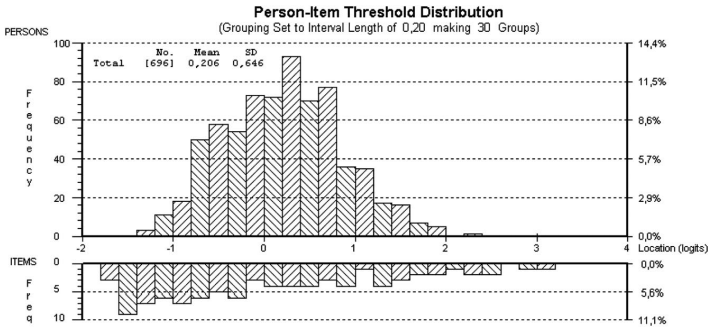

FIG. 2. Person-item targeting for the revised 30-item AddiQoL. The *upper half of the figure* displays spread in AddiQoL scores (person location) for all patients. The *lower half* depicts item threshold distribution (item location). The item thresholds cover the range of HRQoL scores obtained by the patients, hereby minimizing the risk of floor and ceiling effects.

endocrine syndromes. AddiQoL raw scores in patients (median 89, n = 99) were significantly lower than in controls in Norway (median 97, n = 462; U = 14799, z = −5.516, $P < 0.001$, r = 0.23).

Discussion

The validation process resulted in a revised 30-item AddiQoL questionnaire. High reliability, as evidenced by adequate PSI and a high Cronbach's-α, indicates that the items discriminate well between groups of patients with different HRQoL levels. The final revised AddiQoL-30 fitted the stringent Rasch model, implying that basic requirements for a measurement instrument, such as unidimensionality, order, and additivity, are fulfilled. This item solution had the best targeting to the patient sample.

One important aspect of validity is whether individual items work similarly in different patient subgroups, *i.e.* whether item bias exists (34). No significant DIF was found between patients with isolated AD and patients with autoimmune polyendocrine syndromes, and no DIF by sex remained in AddiQol-30. With such reassurance that AddiQoL-30 performs equally in these patient groups, the significantly lower total scores observed in females *vs.* males likely represents a true difference. Similarly, the lack of difference in AddiQol-30 scores between patients with isolated AD and those with polyendocrine syndromes is also reliable. Both findings are consistent with earlier studies using SF-36 in AD (11, 14). We cannot rule out the possibility that the observed DIF by country is due to qualitative differences of the translations. Thus, if the aim was to study HRQoL differences between countries, statistical adjustment of items with DIF country would be required (35). However, DIF country may have negligible clinical impact in evaluating the HRQoL results from a multicenter clinical trial (36).

We demonstrate high correlation between the AddiQoL-30 and AddiQoL-8 scores and SF-36 and PGWB. For SF-36 the correlation was highest with the vitality and general health scales, which were also most affected in previous studies in AD (3, 11, 12, 14). Normative data are not essential in the validation of a disease-specific questionnaire because several of the issues may not be relevant to healthy subjects. Normative data were collected only from Norway; the response rate was low, which could imply selection bias, but the age and sex distribution of healthy subjects resembled that of the patient group. We found a statistically significant difference between the patients and the controls, but the effect size was small. Several of the items showed ceiling effects in the controls,

TABLE 2. Concurrent validity: correlation between AddiQoL-30 scores and AddiQoL-8 scores with SF-36 and PGWB scores

	Norway (n = 107)		Sweden (n = 101)	
	AddiQoL-30	**AddiQoL-8**	**AddiQoL-30**	**AddiQoL-8**
SF-36				
Physical functioning	0.743	0.745	0.692	0.687
Role physical	0.689	0.729	0.717	0.661
Bodily pain	0.604	0.538	0.637	0.545
General health	0.802	0.775	0.768	0.692
Vitality	0.753	0.748	0.837	0.803
Social functioning	0.685	0.676	0.603	0.585
Role emotional	0.433	0.418	0.466	0.411
Mental health	0.585	0.554	0.724	0.675
PGWB				
Total score	0.816	0.797	0.785	0.706

For Spearman's rho, all correlations were significant below the 0.01 level (two tailed).

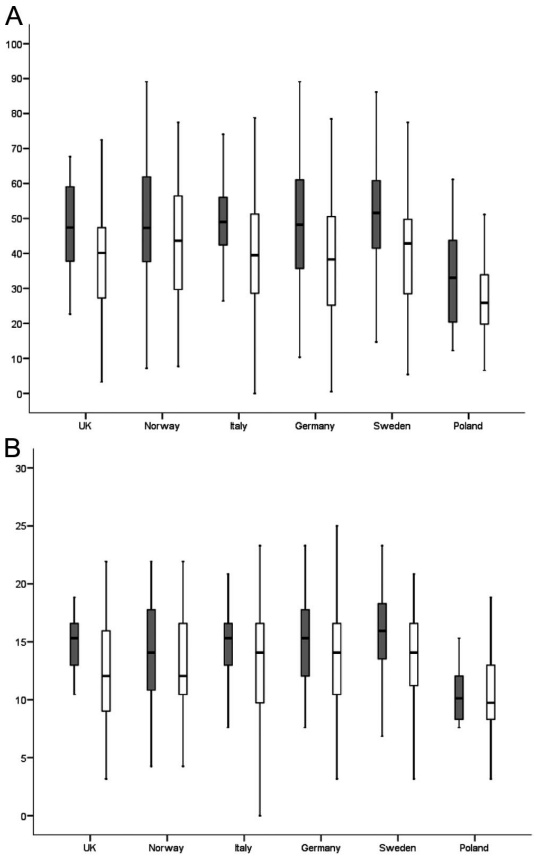

FIG. 3. A, Rasch-transformed AddiQoL-30 scores (range 0–100). B, Rasch-transformed AddiQoL-8 scores (range 0–25). Patient scores for each country [median and interquartile range (boxes)]. Males, *gray*; females, *white*.

which underestimates the effect size. Furthermore, comparison of patients with healthy controls always implies some response bias or a response shift due to adaptation to chronic disease (37).

The results of the Rasch analysis suggest a revised scoring algorithm. The analysis revealed that the six response categories of each item had to be collapsed into four to obtain order, additivity, and fit to the model. The unidimensional structure of the rescored AddiQol-30 suggests that an index based on the algebraic sum of the item (after reversal of negative items) scores will be valid for practical purposes, for instance, in cross-sectional studies. However, AddiQoL was developed primarily as an instrument to evaluate within-individual well-being in clinical trials. For this purpose, Rasch-transformed person location scores will be the optimal psychometric solution.

AddiQoL-30 could possibly be further shortened to reduce the respondents' burden. We found that eight of the items (constituting the fatigue subscale; AddiQoL-8) also had good psychometric properties and a higher reliability than AddiQoL-30 and could be useful as a separate short version of AddiQoL. High reliability and good discriminative questionnaire characteristics usually imply good evaluative properties. On the other hand, a very high reliability coefficient is not necessarily desirable because the same properties that increase the reliability coefficients might also reduce ability to detect change (responsiveness) (38), and the items most sensitive to change might not be the most well-fitting items (38, 39). Whereas the AddiQoL-8 contains items related to fatigue/energy level only, the AddiQoL-30 also includes items concerning other clinically relevant issues in AD, which might be important in evaluating HRQoL changes in response to changes in treatment. This is probably the reason why the AddiQoL-8 scale did not target the whole patient population as well as AddiQoL-30, with a risk of floor and ceiling effects that might compromise responsiveness. Hence, further item reduction will be reevaluated after testing for responsiveness.

In conclusion, the validation process resulted in a revised, 30-item, AddiQoL questionnaire including a separate short version that have high internal consistency and reliability. Its validity as a HRQoL instrument in AD was further substantiated by high correlation with SF-36 subscales and the PGWB Index. Although further studies are necessary to examine its responsiveness to changes in HRQoL over time, this study suggests that the AddiQoL could become a valuable tool in the assessment of subjective health status in patients with Addison's disease.

Acknowledgments

In Norway, we thank Inger Johanne Næss for handling the questionnaires and contributing endocrinologist Martina Moter Erichsen. In Sweden, we thank Lena Ehrenstig for distributing the questionnaires. In Italy, we thank Stefania Marzotti in Perugia and Daniela Forno in Turin for assistance in recruiting patients. In Germany, we thank contributing endocrinologists (Reinhard Santen, Alexander Mann, Holger Willenberg, Stefanie Hahner, Bruno Allolio, Nicole Reisch, and Endokrinologikum Hamburg) and the German Endocrine Society's adrenal division. We also thank the patient organization GLANDULA. We also thank all the patients with Addison's disease who participated.

Address all correspondence and requests for reprints to: Marianne Øksnes, Institute of Medicine, Haukeland University Hospital, Jonas Liesvei 65, N-5021 Bergen, Norway. E-mail: marianne.oksnes@med.uib.no.

This work was supported by the European Union Seventh Framework Program Grant 201167, Euradrenal, and the National Institute of Health Research Cambridge Biomedical Cen-

tre (United Kingdom). The Swedish contribution was also supported by the Swedish Society of Medicine.

Disclosure Summary: No conflict of interest was reported by any of the authors.

References

1. **Arlt W, Allolio B** 2003 Adrenal insufficiency. Lancet 361:1881–1893
2. **Løvås K, Gebre-Medhin G, Trovik TS, Fougner KJ, Uhlving S, Nedrebø BG, Myking OL, Kampe O, Husebye ES** 2003 Replacement of dehydroepiandrosterone in adrenal failure: no benefit for subjective health status and sexuality in a 9-month, randomized, parallel group clinical trial. J Clin Endocrinol Metab 88:1112–1118
3. **Gurnell EM, Hunt PJ, Curran SE, Conway CL, Pullenayegum EM, Huppert FA, Compston JE, Herbert J, Chatterjee VK** 2008 Long-term DHEA replacement in primary adrenal insufficiency: a randomized, controlled trial. J Clin Endocrinol Metab 93:400–409
4. **Arlt W, Callies F, van Vlijmen JC, Koehler I, Reincke M, Bidlingmaier M, Huebler D, Oettel M, Ernst M, Schulte HM, Allolio B** 1999 Dehydroepiandrosterone replacement in women with adrenal insufficiency. N Engl J Med 341:1013–1020
5. **Hunt PJ, Gurnell EM, Huppert FA, Richards C, Prevost AT, Wass JA, Herbert J, Chatterjee VK** 2000 Improvement in mood and fatigue after dehydroepiandrosterone replacement in Addison's disease in a randomized, double blind trial. J Clin Endocrinol Metab 85:4650–4656
6. **Løvås K, Husebye ES** 2007 Continuous subcutaneous hydrocortisone infusion in Addison's disease. Eur J Endocrinol 157:109–112
7. **Johannsson G, Filipsson H, Bergthorsdottir R, Lennernas H, Skrtic S** 2007 Long-acting hydrocortisone for glucocorticoid replacement therapy. Horm Res 68(Suppl 5):182–188
8. **Newell-Price J, Whiteman M, Rostami-Hodjegan A, Darzy K, Shalet S, Tucker GT, Ross RJ** 2008 Modified-release hydrocortisone for circadian therapy: a proof-of-principle study in dexamethasone-suppressed normal volunteers. Clin Endocrinol (Oxf) 68:130–135
9. **Johannsson G, Bergthorsdottir R, Nilsson AG, Lennernas H, Hedner T, Skrtic S** 2009 Improving glucocorticoid replacement therapy using a novel modified-release hydrocortisone tablet: a pharmacokinetic study. Eur J Endocrinol 161:119–130
10. **Arlt W, Rosenthal C, Hahner S, Allolio B** 2006 Quality of glucocorticoid replacement in adrenal insufficiency: clinical assessment vs. timed serum cortisol measurements. Clin Endocrinol (Oxf) 64:384–389
11. **Hahner S, Loeffler M, Fassnacht M, Weismann D, Koschker AC, Quinkler M, Decker O, Arlt W, Allolio B** 2007 Impaired subjective health status in 256 patients with adrenal insufficiency on standard therapy based on cross-sectional analysis. J Clin Endocrinol Metab 92:3912–3922
12. **Løvas K, Loge JH, Husebye ES** 2002 Subjective health status in Norwegian patients with Addison's disease. Clin Endocrinol (Oxf) 56:581–588
13. **Bleicken B, Hahner S, Loeffler M, Ventz M, Allolio B, Quinkler M** 2008 Impaired subjective health status in chronic adrenal insufficiency: impact of different glucocorticoid replacement regimens. Eur J Endocrinol 159:811–817
14. **Erichsen MM, Løvås K, Skinningsrud B, Wolff AB, Undlien DE, Svartberg J, Fougner KJ, Berg TJ, Bollerslev J, Mella B, Carlson JA, Erlich H, Husebye ES** 2009 Clinical, immunological, and genetic features of autoimmune primary adrenal insufficiency: observations from a Norwegian registry. J Clin Endocrinol Metab 94:4882–4890
15. **Ott J, Promberger R, Kober F, Neuhold N, Tea M, Huber JC, Hermann M** 2011 Hashimoto's thyroiditis affects symptom load and quality of life unrelated to hypothyroidism: a prospective case-control study in women undergoing thyroidectomy for benign goiter. Thyroid 21:161–167
16. **Prieto L, Santed R, Cobo E, Alonso J** 1999 A new measure for assessing the health-related quality of life of patients with vertigo, dizziness or imbalance: the VDI questionnaire. Qual Life Res 8:131–139
17. **Løvås K, Curran S, Øksnes M, Husebye ES, Huppert FA, Chatterjee VK** 2010 Development of a disease-specific quality of life questionnaire in Addison's disease. J Clin Endocrinol Metab 95:545–551
18. **Wiren L, Whalley D, McKenna S, Wilhelmsen L** 2000 Application of a disease-specific, quality-of-life measure (QoL-AGHDA) in growth hormone-deficient adults and a random population sample in Sweden: validation of the measure by rasch analysis. Clin Endocrinol (Oxf) 52:143–152
19. **Webb SM, Prieto L, Badia X, Albareda M, Catala M, Gaztambide S, Lucas T, Paramo C, Pico A, Lucas A, Halperin I, Obiols G, Astorga R** 2002 Acromegaly Quality of Life Questionnaire (ACROQOL) a new health-related quality of life questionnaire for patients with acromegaly: development and psychometric properties. Clin Endocrinol (Oxf) 57:251–258
20. **Webb SM, Badia X, Barahona MJ, Colao A, Strasburger CJ, Tabarin A, van Aken MO, Pivonello R, Stalla G, Lamberts SW, Glusman JE** 2008 Evaluation of health-related quality of life in patients with Cushing's syndrome with a new questionnaire. Eur J Endocrinol 158:623–630
21. **Pallant JF, Tennant A** 2007 An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 46:1–18
22. **Tennant A, McKenna SP, Hagell P** 2004 Application of Rasch analysis in the development and application of quality of life instruments. Value Health 7(Suppl 1):S22–S26
23. **Tennant A, Conaghan PG** 2007 The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 57:1358–1362
24. **McKenna SP, Doward LC, Alonso J, Kohlmann T, Niero M, Prieto L, Wíren L** 1999 The QoL-AGHDA: an instrument for the assessment of quality of life in adults with growth hormone deficiency. Qual Life Res 8:373–383
25. **Suzukamo Y, Noguchi H, Takahashi N, Shimatsu A, Chihara K, Green J, Fukuhara S** 2006 Validation of the Japanese version of the Quality of Life-Assessment of Growth Hormone Deficiency in Adults (QoL-AGHDA). Growth Horm IGF Res 16:340–347
26. **Webb SM** 2006 Quality of life in acromegaly. Neuroendocrinology 83:224–229
27. **Acquadro C, Conway K, Hareendran A, Aaronson N** 2008 Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. Value Health 11:509–521
28. **McHorney CA, Ware Jr JE, Raczek AE** 1993 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 31:247–263
29. **Dupyu H ed** 1984 The Psychological General Well-Being (PGWB) Index. New York: Le Jaque Publishing
30. **Andrich L, Sheridan, Luo** 2003 RUMM2020. Perth, Australia: RUMM Laboratory
31. **Smith RM** 2000 Fit analysis in latent trait measurement models. J Appl Meas 1:199–218
32. **Smith Jr EV** 2001 Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. J Appl Meas 2:281–311
33. **Tang K, Beaton DE, Lacaille D, Gignac MA, Zhang W, Anis AH, Bombardier C** 2010 The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis? Qual Life Res 19:1057–1068
34. **Teresi JA** 2006 Overview of quantitative measurement methods.

Equivalence, invariance, and differential item functioning in health applications. Med Care 44:S39–S49

35. **Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sorensen F, Marincek C, Burger H, Phillips S** 2004 Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 42:I37–I48
36. **Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MA** 2009 The practical impact of differential item functioning analyses in a health-related quality of life instrument. Qual Life Res 18:1125–1130
37. **Oort FJ, Visser MR, Sprangers MA** 2009 Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. J Clin Epidemiol 62:1126–1137
38. **Fava GA, Belaise C** 2005 A discussion on the role of clinimetrics and the misleading effects of psychometric theory. J Clin Epidemiol 58:753–756
39. **Wright JG, Feinstein AR** 1992 A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol 45:1201–1218

Supplementary table1: AddiQoL items

Items AddiQoL-36	Sub-dimension	AddiQoL-30
1. I feel good about my health	Fatigue	
2. I can keep going during the day without feeling tired	Fatigue	
3. Normal daily activities make me tired	Fatigue	
4. I have to struggle to finish jobs	Fatigue	
5. I have to push myself to do things	Fatigue	
6. I lose track of what I want to say	Symptom	
7. I sleep well	Miscelleneous	
8. I feel rested when I wake up in the morning	Miscelleneous	
9. I need to get up during the night to pass water	Did not fit any dimension	Eliminated due to misfit
10. I feel unwell first thing in the morning	Symptom	
11. I am satisfied with my sex life	Miscelleneous	
12. I am relaxed	Emotion	
13. I feel low or depressed	Emotion	
14. I am irritable	Emotion	
15. I find it difficult to think clearly	Emotion	
16. I feel lightheaded	Symptom	
17. I have salt cravings	Symptom	Eliminated due to misfit
18. I sweat for no particular reason	Symptom	
19. I get headaches	Symptom	
20. I get nauseous	Symptom	
21. My joints and/or muscles ache	Symptom	
22. I have back pain	Symptom	
23. My legs feel weak	Symptom	
24. I worry about my health	Symptom	
25. My ability to work is limited	Fatigue	
26. I can concentrate well	Emotion	
27. I am happy	Emotion	
28. I feel full of energy	Fatigue	
29. I feel physically fit	Fatigue	
30. Emotional stress makes me exhausted	Emotion	Eliminated due to item bias
31. I have lost interest in sex	Miscelleneous	Eliminated due to item bias
32. I put on weight easily	Did not fit any dimension	Eliminated due to misfit
33. I have dry skin	Did not fit any dimension	Eliminated due to misfit
34. I get ill more easily than others	Miscelleneous	
35. I take a long time to recover from illnesses	Miscelleneous	
36. I cope well in emotional situations	Emotion	

Supplementary table2. Overall fit to the Rasch model; the AddiQoL-30 results

	Norway (n=107)	Sweden (n=101)	Italy (n=156)	Germany (n=200)	Poland (n=50)	All countries (n=696)
Item Fit Residual (SD)	-0.55 (1.15)	-0.57 (1.45)	-0.51 (0.90)	0.01 (1.48)	-0.47 (0.9)	0.00 (2.7)
Person Fit Residual (SD)	-0.52 (1.06)	-0.62 (1.22)	-0.6 (1.08)	-0.48 (1.13)	-0.66 (1.2)	-0.51 (1.15)
χ^2 probability	0.81	0.66	0.88	0.73	0.92	0.73
Person Separation Index	0.84	0.87	0.83	0.91	0.85	0.86
Cronbach's α	ND	ND	ND	ND	ND	0.93

All countries also includes original UK data (n=82). Item Fit Residual: Mean item deviation from model estimations. Person Fit Residual: Mean person deviation from model estimations. A non-significant χ^2 probability implies that the hierarchical ordering of items and persons do not vary across the range of the scale. ND; not done.