THE JOURNAL OF
SEXUAL MEDICINE

Psychometric Evaluation of the Hypogonadism Impact of Symptoms Questionnaire

CrossMark

Heather L. Gelhorn, PhD,[1] Ebony Dashiell-Aje, PhD,[1] Michael G. Miller, PharmD,[2] Leonard R. DeRogatis, PhD,[3] Adrian Dobs, MD, MHS,[4] Allen D. Seftel, MD,[5] Stanley E. Althof, PhD,[6] Meryl Brod, PhD,[7] and Dennis A. Revicki, PhD[1]

ABSTRACT

Introduction: The Hypogonadism Impact of Symptoms Questionnaire (HIS-Q) is a patient-reported outcome measurement designed to comprehensively evaluate the symptoms of hypogonadism and to detect changes in these symptoms in response to treatment.

Aim: To conduct item analysis and reduction, evaluate the psychometric properties of the HIS-Q, and provide guidance on interpreting the instrument score.

Methods: A 12-week observational, longitudinal study of hypogonadal men was conducted. Participants completed the HIS-Q every 2 weeks. Blood samples were collected to evaluate testosterone levels. Participants also completed the Aging Male's Symptoms Scale, the International Index of Erectile Function, the Short Form-12 Health Survey, and the Patient-Reported Outcomes Measurement Information System Sexual Activity, Satisfaction with Sex Life, Sleep Disturbance, and Applied Cognition Scales (at baseline and weeks 6 and 12). Clinicians completed the Clinical Global Impression of Severity and Change measurements and a clinical form.

Main Outcome Measures: Individual item performance was evaluated using descriptive statistics and Rasch analyses. Reliability (internal consistency and test-retest), validity (concurrent and know groups), and responsiveness were assessed.

Results: In total, 177 men participated in the study (mean age = 54.1 years, range = 23−83). The original 53-item draft HIS-Q was reduced to 28 items; the final instrument included five domains (sexual, energy, sleep, cognition, and mood) with two sexual subdomains (libido and sexual function). For all domains, test-retest reliability was acceptable (intraclass correlation coefficients > 0.70), construct validity was good ($|r > 0.30|$ for all comparisons). Known-groups validity was demonstrated for all HIS-Q domain scores, subdomain scores, and the total score as measured by the Clinical Global Impression of Severity, and total testosterone level at baseline ($P < .05$ for all comparisons). All domains and subdomains were responsive to change based on patient-rated anchor questions ($P < .05$ for all comparisons).

Conclusion: The final 28-item HIS-Q is reliable, valid, and responsive. The HIS-Q is suitable for inclusion in future clinical trials to help characterize the effects of testosterone replacement therapy.

J Sex Med 2016;13:1737−1749. Copyright © 2016, The Authors. Published by Elsevier Inc. on behalf of the International Society for Sexual Medicine. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Key Words: Hypogonadism; Patient-Reported Outcome; Hypogonadism Impact of Symptoms Questionnaire; Psychometric Properties; Reliability; Validity; Responsiveness

INTRODUCTION

Hypogonadism is associated with a range of symptoms, including decreased libido, erectile dysfunction, decreased energy, sleep disturbance, and changes in mood.[1] Many of these symptoms are difficult to evaluate clinically and are best assessed through patient-reported outcome (PRO) measurements. There are several existing PRO instruments that have been used to evaluate the symptoms and effects of hypogonadism, such as the Psychosexual Daily Questionnaire,[2] the Aging Males Symptoms Scale (AMS),[3] the Sexual Arousal, Interest, and Drive Scale, and the Hypogonadism Energy Diary.[4] Although each of these instruments is useful in specific contexts, and some are well known, each is limited in at least one way. Specifically, all were not designed to assess hypogonadism comprehensively, have not been validated, and/or have not been developed according to regulatory standards.

The Hypogonadism Impact of Symptoms Questionnaire (HIS-Q) is a PRO instrument that was designed to assess changes in symptoms in hypogonadal men in response to testosterone replacement therapy (TRT).[5] The instrument was developed primarily for use in clinical trial settings, has been developed in accordance with the Food and Drug Administration's PRO guidance for industry, and addresses the limitations of existing instruments.[6] Development was informed by a literature review, input from clinicians, and qualitative studies with patients aimed at comprehensively identifying the symptoms of the condition.[5]

The original draft HIS-Q was a 53-item questionnaire with a conceptual framework that included sexual, physical signs and symptoms, energy, sleep, cognitive, and mood domains. Results from the initial qualitative work in hypogonadal patients indicated that the HIS-Q was a comprehensive measurement of hypogonadal symptoms in men. The HIS-Q reflected the varied symptom experiences of patients, and the content validity of the instrument was confirmed.[5]

METHODS

Aims

The present study was designed as a stand-alone, 12-week, longitudinal, observational validation study of hypogonadal men for informing item reduction, developing scoring and scaling, and evaluating the psychometric properties of the HIS-Q. The primary objective of this study was to develop a final version of the HIS-Q of appropriate length for use in clinical trials, with an established scoring algorithm and acceptable psychometric properties (ie, reliability, validity, and responsiveness).

Participants

Twenty U.S. clinical sites participated in the study. Eligible participants who signed informed consents included men who were at least 18 years old; diagnosed with hypogonadism (serum total testosterone concentration < 300 ng/dL before enrollment);

switching TRT treatments or on maintenance therapy or were treatment naïve; and able to understand English. Patients were excluded if they had non-stabilized depression (stabilized was defined as on the same antidepressant medication at the same dosage for ≥ 3 months), severe psychiatric illness, or addictions; history of or current obstructive sleep apnea; a clinically significant medical condition; or taking a concurrent medication that would affect hormonal balance or sexual functioning (eg, phosphodiesterase type 5 inhibitors) or interfere with participants' participation in the study.

Procedures and Measurements

The study protocol and procedures used were reviewed and approved by the appropriate institutional review committee (Ethical and Independent Review Services Institutional Review Board, September 25, 2013, protocol 13110-01). All study staff at every site were trained using a standardized training protocol on the purpose and procedures for the study. Participants completed in-person study visits at baseline, week 2, and week 12. The first assessment was completed during their first in-clinic visit (baseline), and subsequent assessments were completed at home every 2 weeks from week 2 through week 12. All PRO assessments were completed on an electronic PRO device. A subset of participants also completed a daily diary from baseline to day 28. The schedule of study assessments and a description of each is presented in Table 1.[3,7−11] Testosterone levels were assessed through blood draws from all participants at baseline and from participants who were beginning treatment or switching to a different treatment at weeks 2 and 12. Blood serum samples were processed and analyzed by a central laboratory (Total Testosterone: AbbVie, Inc., North Chicago, IL, USA; Free Testosterone: PPD Laboratories, Middleton, WI, USA) using liquid chromatography and tandem mass spectrometry and equilibrium dialysis to assess total and free testosterone levels, respectively.

Statistical Analyses

Descriptive statistics were used to characterize the sociodemographic and clinical characteristics of the sample. Item-level descriptive statistics (means, SD, median, range, frequency) were used to evaluate the performance of individual HIS-Q items. Then, exploratory factor analyses were used to analyze the factor structure of the instrument. Rasch analyses[12] were used to evaluate individual item and subscale properties. Exploratory factor analyses and Rasch analyses were conducted using SAS (SAS Institute, Cary, NC, USA) and RUMM2030,[13] respectively.

These analyses were conducted iteratively and the results were used to reduce the item pool and establish subscales and a scoring algorithm for the reduced HIS-Q. Items were flagged for consideration as deletion candidates based on descriptive item analyses (>50% floor and ceiling effects, high inter-item correlations > 0.80 and/or low item-total correlations < 0.20), factor analyses (weak factor loadings < 0.40, poor model fit), and

Table 1. Study events schedule

Study events	Mode	Screening	Visit 1, baseline	Visit 2, week 2*	Week 4†	Week 6†	Week 8†	Week 10†	Visit 3, week 12*	Items, n	Concepts measured	Interpretation guidelines
Investigator/site completed												
Clinical form	Paper		✔								Patients' clinical characteristics	—
CGI-S	Paper		✔	✔					✔	Single item	Physician's overall impression of patient's hypogonadism symptoms	Higher scores indicate greater severity
CGI-C	Paper			✔					✔	Single item	Clinician's perception of change in symptoms between study visits	—
Serum testosterone testing	Blood draw		✔	✔‡					✔‡		—	—
Medical report form	Paper	Completed as needed to report any changes to a patient's treatment or testosterone levels§									—	—
Patient completed												
HIS-Q			✔	✔	✔	✔	✔	✔	✔	53	Sexual and physical signs and symptoms, energy, sleep, cognition, and mood	Lower scores indicate better function and fewer symptoms
AMS[3]	Electronic		✔			✔			✔	17	Sleep difficulty, low energy, physical symptoms, effects on sexual functioning and mood	Lower scores indicate better functioning
IIEF[7]	Electronic		✔			✔			✔	15	Domains of sexual function (erectile dysfunction, orgasmic function, sexual desire, intercourse satisfaction, and overall satisfaction)	Higher scores indicate less dysfunction
PROMIS Interest in Sexual Activity Scale[8]	Electronic		✔			✔			✔	4	Sexual function, including desire in past 30 d	Higher scores indicate more interest in sexual activity
PROMIS Global Satisfaction With Sex Life Scale[8]	Electronic		✔			✔			✔	7	Satisfaction with sex life in past 30 d	Higher scores indicate greater satisfaction

(continued)

Table 1. Continued

Study events	Mode	Screening	Visit 1, baseline	Visit 2, week 2*	Week 4†	Week 6†	Week 8†	Week 10†	Visit 3, week 12*	Items, n	Concepts measured	Interpretation guidelines
PROMIS Sleep Disturbance Scale[9]	Electronic		✔			✔			✔	8	Perceptions of sleep quality, sleep depth, and restoration associated with sleep	Higher scores indicate greater sleep disturbance
PROMIS Applied Cognition—Abilities: (SF-8a)[10]	Electronic		✔			✔			✔	8	Perceptions of cognitive functioning and changes in cognitive abilities (eg, concentration, memory) over 7-d period	Higher scores indicate better cognitive functioning
SF-12[11]	Electronic		✔			✔			✔	12	Functional health and well-being, physical and mental health, and health utility during a typical day and past 4 wk	Higher scores indicate better health status
Anchor questions	Electronic		✔	✔	✔	✔			✔	9	Sexual activity, libido, erectile functioning, overall sexual function, tiredness, mood, cognitive functioning, sleep, and overall hypogonadism condition	Higher scores indicate better functioning and outcomes
Daily diary‖	Electronic		Daily from baseline to day 28							12	Sexual activity, erectile function, energy, sleep, cognition, and mood	Higher scores indicate greater frequency or more symptoms

AMS = Aging Male's Symptoms Scale; CGI-C = Clinical Global Impression—Change; CGI-S = Global Impression—Severity; HIS-Q = Hypogonadism Impact of Symptoms Questionnaire; IIEF = International Index of Erectile Function; PROMIS = Patient-Reported Outcomes Measurement Information System; SF-12 = Short Form-12 Health Survey.

*At visits 1, 2, and 3, patients attended a study visit; however, all patient-reported outcome measurements for visits 2 and 3 were completed by patients at home, before the visit, at their regularly scheduled time on the electronic patient-reported outcomes device.

†Patients did not come to the sites for study-related visits during these weeks but did complete study questionnaires on the electronic patient-reported outcomes device at home at each of these time points.

‡Only patients on new treatment had study-related blood draws at visits 2 and 3.

§Sites reported any additional serum testosterone blood sample results that were independent of the study blood draws at each visit.

‖A subset of 60 patients completed a daily diary of questions related to the sexual activity domain and the other domains (energy, sleep, cognitive, mood) from baseline to day 28.

Rasch analyses (item misfit, disordered thresholds). Item reduction decisions also considered previous qualitative research findings (ie, concept elicitation, content validity, etc) and input from expert clinicians on the clinical relevance and importance of items.

After item pool reduction and scoring algorithm establishment, the psychometric properties of the HIS-Q were evaluated. Internal consistency reliability of the HIS-Q domains was evaluated using the Cronbach α coefficient[14] at baseline, week 6, and week 12, with reliability values of at least 0.70 denoting a more homogeneous scale.[15] Test-retest reliability was assessed using intraclass correlation coefficients and paired-sample t-tests for all HIS-Q domain scores and the total score to examine HIS-Q stability over time within a stable subsample. Stable subjects were those with "no change" in relevant anchor questions and Clinical Global Impression (CGI) for Change from baseline to week 2; intraclass correlation coefficients of at least 0.7 indicate good test-retest reliability, coefficients from 0.4 to 0.7 indicate moderate test-retest reliability, and coefficients less 0.4 indicated low test-retest reliability.[15,16]

Convergent and divergent validity of the HIS-Q was assessed using Pearson product-moment and Spearman rank correlation coefficients to estimate the relation between HIS-Q domains and other conceptually relevant measurements (eg, higher correlations among HIS-Q sexual domains and Patient-Reported Outcomes Measurement Information System [PROMIS] and AMS sexual scales, HIS-Q energy domain and AMS Somato-Vegetative Scale and Short Form-12 Health Survey vitality item, HIS-Q sleep domain and PROMIS sleep, HIS-Q cognition domain and PROMIS cognition, HIS-Q mood domain and AMS psychological). To assess known-groups validity, mean domain scores on the HIS-Q were analyzed by disease severity (CGI for Severity) and total testosterone was analyzed by analysis of variance.

Responder definitions were identified for the HIS-Q domains, subdomains, and total score using anchor-based and distribution-based methods. Mean scores for each HIS-Q domain and subdomain for participants who improved by one point on each concept-specific anchor question were used to establish anchor-based responder definitions. The SD at baseline (0.2, 0.3, and 0.5 times estimates) and the standard error of measurement[17,18] were used to establish distribution-based responder definitions. A triangulation process considering anchor- and distribution-based definitions was used to derive final responder definitions of clinically meaningful change for each of the HIS-Q domain and subdomain scores.

RESULTS

In total, 196 patients were recruited into the study; a final analysis sample of 177 patients included all participants who had HIS-Q and clinical data at baseline (baseline, n = 177; week 2, n = 163; week 4, n = 166; week 6, n = 162; week 8, n = 157;

Table 2. Sociodemographic characteristics (N = 177)

Age (y)	
Mean (SD)	54.1 (11.4)
Median (range)	55.0 (23.0–83.0)
Missing, n (%)	1 (0.6)
Race, n (%)*	
Black or African American	32 (18.1)
White	131 (74.0)
Other†	7 (4.0)
Missing	7 (4.0)
Ethnicity, n (%)	
Hispanic or Latino	9 (5.1)
Not Hispanic or Latino	166 (93.8)
Missing	2 (1.1)
Employment status, n (%)	
Employed, full time	105 (59.3)
Employed, part time	12 (6.8)
Student	2 (1.1)
Unemployed, disabled, retired	52 (29.4)
Other‡	5 (2.8)
Missing	1 (0.6)
Education, n (%)	
Secondary, high school, some college, trade school	88 (49.7)
College degree	58 (32.8)
Postgraduate degree	30 (16.9)
Missing	1 (0.6)
Currently in intimate relationship, n (%)	
Yes	146 (82.5)
No	30 (16.9)
Missing	1 (0.6)

*Categories are not mutually exclusive.
†Asian (n = 3), Native Hawaiian/Pacific Islander (n = 1), Hispanic (n = 1), Haitian (n = 1), and Jamaican (n = 1).
‡Self-employed (n = 4) and sales (n = 1).

week 10, n = 160; week 12, n = 156). Participants were recruited from 20 clinical sites in 14 different states across the United States (mean number of patients per site = 9.9, SD = 4.6). Participants included 47 men who were new to TRT, 41 men who were switching from one TRT to another, and 89 maintenance patients who were on TRT at the time of enrollment and had no plans to change their treatment.

Demographics and Clinical Characteristics

The mean age of study participants was 54.1 years (SD = 11.4), and most participants were white (74.0%) and were involved in an intimate relationship (82.5%; Table 2). Average time since hypogonadism diagnosis was 2.8 years (SD = 2.2), with almost half the study participants classified by their providers as having a primary acquired etiology (45.8%; Table 3). Average body mass index of the study sample was 30.2 (SD = 5.2). Baseline mean total and free testosterone levels for all tested participants were 507.6 ng/dL (SD = 495.4) and 15.0 ng/dL (SD = 15.9), respectively.

Table 3. Baseline participant clinical characteristics—site reported (N = 177)

Time since initial diagnosis of hypogonadism (y)	
Mean (SD) [range]	2.2 (3.2) [0.0–20.56]
Unknown, n (%)	1 (0.6)
Provider-reported hypogonadism etiology, n (%)	
Primary congenital	23 (13.0)
Primary acquired	81 (45.8)
Secondary congenital	0 (0.0)
Secondary acquired	28 (15.8)
Combined	7 (4.0)
Unknown	38 (21.5)
Specific suspected etiology or diagnosis, n (%)	
Pituitary adenoma or disorder	2 (1.1)
Testicular trauma or disorder	2 (1.1)
Other*	19 (10.7)
Unknown	154 (87.0)
Chief complaint or presenting symptom, n (%)	
Erectile dysfunction	46 (26.0)
Low libido	45 (25.4)
Tiredness	29 (16.4)
Fatigue	48 (27.1)
Other†	5 (2.8)
Unknown	4 (2.3)
History of testosterone replacement medications, n (%)	
No history	52 (29.4)
Buccal	2 (1.13)
Topical	69 (39.0)
Patch	6 (3.4)
Subcutaneous pellet	20 (11.3)
Injection	69 (39.0)
Missing	1 (0.6)
Calculated BMI, mean (SD) [range]‡	30.2 (5.2) [21.5–53.2]
Baseline serum total testosterone concentration, n§	172
Concentration (ng/dL), mean (SD) [range]	507.62 (495.37) [19.70–4,160.00]
Missing, n (%)	5 (2.8)
Baseline free testosterone concentration, n	149
Concentration (ng/dL), mean (SD) [range]	15.00 (15.89) [0.53–126.00]
Missing, n (%)	28 (15.8)

BMI = body mass index.
*Senescence (n = 13), obesity (n = 2), aging (n = 3), and testicular failure (n = 1).
†No symptom (n = 1), poor concentration (n = 1), low energy (n = 1), weakness (n = 1), and weight gain (n = 1).
‡BMI = (weight [pounds] × 703)/(height [inches])^2.
§Baseline serum total testosterone concentration lower than 300 ng/dL (n = 68).

Baseline AMS scores indicated that participants on maintenance therapy had moderate levels of sexual impairment, whereas patients who were switching medications or were treatment naïve had severe sexual impairment (overall mean = 11.7, SD = 4.3, range = 5–23). Baseline AMS scores indicated moderate psychological impairment (overall mean = 8.7, SD = 3.8) and moderate somatic impairment (overall mean = 15.1, SD = 5.2). Mean International Index of Erectile Function domain scores indicated moderate to high dysfunction, with the highest dysfunction demonstrated for the orgasmic function subscale (mean = 3.6 ± 2.8, range = 0–10); moderate dysfunction was observed for all other subscales (ie, erectile function, sexual desire, intercourse satisfaction, and overall satisfaction). The mean CGI for Severity value for the full sample at baseline was 3.5 (SD = 1.1, range = 1–6), indicating mild to moderate hypogonadal symptom severity.

HIS-Q Item Evaluation and Item Reduction

The individual item-level analyses demonstrated acceptable distribution of the HIS-Q item responses across response categories and good distributional characteristics. Minor issues (eg, floor effects, high inter-item correlations, disordered thresholds, poor fit to the Rasch model) were noted for some items and taken into consideration during item reduction. For example, because expected items 15 (sexual activities enjoyable) and 16 (sexual activities satisfying) and items 22 (tired) and 23 (physically tired) were each highly correlated with each other. Thus, additional information (ie, factor loadings and Rasch results) was used to select the best-performing item from each pair. Other items performed poorly based on most item evaluation criteria and were eliminated.

Initially, a 30-item revised version of the instrument was established based on the item analyses; this included five numerical response items on frequency of sexual activities and an additional 25 ordinal response scale items in the sexual (n = 7), physical signs and symptoms (n = 2), energy (n = 3), sleep (n = 3), cognition (n = 3), and mood (n = 7) domains. However, psychometric analyses on this 30-item version of the HIS-Q indicated very poor psychometric characteristics for the physical signs and symptoms domain, and the two items from this domain were subsequently removed, yielding the final 28-item HIS-Q (Supplementary Appendix A). This decision was supported by the clinical experts who assessed aspects of this domain using direct clinical measurements.

Factor Structure and Scoring for the Final 28-Item HIS-Q

Exploratory factor analyses (n = 133) suggested that the final HIS-Q is multifactorial. A five-factor solution yielded generally high factor loadings for the items within the following domains and subdomains (factor loading ranges: libido = 0.50–0.68; sexual function = 0.51–0.87; energy = 0.71–0.80; sleep = 0.37–0.68; and mood and cognition = 0.31–0.68;

Supplementary Table 1). The cognition domain did not emerge as a distinct factor in this five-factor solution, but rather the cognition items loaded with the mood items.

Instrument scoring includes each of the 23 ordinal response scale items and yields five domain scores (sexual; energy, sleep, cognition, and mood) and two sexual subdomain scores (libido and sexual function). All 23 ordinal response scale items also can be used to yield a HIS-Q total score. The domain, subdomain, and total scores range from 0 to 100, with higher scores indicating greater dysfunction. No scoring is needed for the first five questions, because the items capture numerical data related to the frequency of sexual activity.

Psychometric Evaluation of the Final 28-Item HIS-Q

After item selection and development of the scoring algorithm, the psychometric properties of the final 28-item HIS-Q were evaluated.

Rasch Analyses

Rasch analyses were conducted on each of the HIS-Q domains and subdomains (n = 117—171). Item performance was very good, with all items demonstrating fit to the Rasch model ($P > .05$ for all comparisons) and good distributions of item thresholds (libido, $\beta = -3.2$ to 3.4; sexual function, $\beta = -2.4$ to 7.49; energy, $\beta = -6.0$ to 7.4; sleep, $\beta = -2.6$ to 2.1; cognition, $\beta = -3.2$ to 2.1; mood, $\beta = -3.6$ to 2.7). These item thresholds were well matched to the distributions of individuals within each domain. There were a few minor issues with disordered thresholds (ie, some item response categories did not accurately distinguish between participants with different levels of hypogonadism severity) for three items, including "difficult achieving erections," "difficulty ejaculating," and "sad."

Reliability

The internal consistency reliability (ie, Cronbach α) was very good for the sexual symptoms domain (baseline = 0.79, week 6 = 0.80, week 12 = 0.81), libido subdomain (baseline = 0.74, week 6 = 0.73, week 12 = 0.76), sexual function subdomain (baseline = 0.78, week 6 = 0.80, week 12 = 0.78), energy domain (baseline = 0.91, week 6 = 0.90, week 12 = 0.90), mood domain (baseline = 0.85, week 6 = 0.87, week 12 = 0.89), and HIS-Q total score (baseline = 0.89, week 6 = 0.91, week 12 = 0.92). Internal consistency reliability was slightly lower for the cognition (range = 0.65—0.72) and sleep (range = 0.58—0.64) domains.

Test-retest reliability was assessed in those who were defined as stable based on patient-reported anchor questions from baseline to week 2. The ICCs were good for the sexual symptoms domain (0.81), libido subdomain (0.77), sexual function subdomain (0.80), energy domain (0.73), mood domain (0.87), and total score (0.80). Test-retest reliability was acceptable for the sleep (0.68) and cognition (0.68) domains.

Validity

There was strong evidence for convergent validity across all scales. The sexual domain and libido and sexual function subdomains were strongly correlated with the AMS Sexual Scale (r = 0.65, 0.55, 0.55, respectively, $P < .0001$ for all comparisons), the International Index of Erectile Function Overall Satisfaction Score (r = 0.68, 0.44, 0.67, $P < .0001$ for all comparisons), the PROMIS Sexual Activity Score (r = -0.60, -0.74, -0.39, $P < .0001$ for all comparisons), and the PROMIS Global Satisfaction with Sex Life Score (r = -0.65, -0.45, -0.60, $P < .0001$ for all comparisons). The energy domain was strongly correlated with the AMS Somato-Vegetative Scale (r = 0.67, $P < .0001$) and the Short Form-12 Health Survey Vitality item (r = -0.62, $P < .0001$). The sleep domain was strongly correlated with the PROMIS Sleep Disturbance score (r = 0.75, $P < .0001$), the cognition domain was strongly correlated with the PROMIS Applied Cognition score (r = -0.72, $P < .0001$), and the mood domain strongly correlated with the AMS Psychological Scale (r = 0.78, $P < .0001$). Divergent validity was demonstrated through low and non-significant correlations among conceptually unrelated scales (eg, HIS-Q energy, sleep, cognition, and mood domains with International Index of Erectile Function Orgasmic Function Scale; r = 0.11—0.16).

The HIS-Q scores also were correlated with testosterone levels. At baseline, there were moderate correlations between total testosterone and the sexual domain (r = -0.36, $P < .0001$), libido subdomain (r = -0.29, $P < .001$), sexual function domain (r = -0.32, $P < .0001$), energy domain (r = -0.25, $P < .001$), and HIS-Q total score (r = -0.34, $P < .0001$). Smaller but significant correlations were observed between total testosterone and sleep (r = -0.17, $P < .05$), cognition (r = -0.19, $P < .05$), and mood (r = -0.18, $P < .05$) domains. A similar pattern of results with moderate correlations was observed between free testosterone levels and sexual, libido, sexual function, energy, and total scores (r = -0.37 to -0.29, $P < .001$ for all comparisons).

Results demonstrated good known-group validity. All HIS-Q scores were significantly different when grouped according to clinician ratings of severity ($P < .05$ for all comparisons), and groups with greater clinician-rated severity also had significantly higher HIS-Q scores (Table 4). Similarly, when grouped by total testosterone levels, significantly higher HIS-Q scores were observed in men with lower testosterone levels ($P < .05$ for all comparisons; Table 4).

Responsiveness

Changes in each of the patient-reported anchor questions were reflected by significant changes in the expected direction for all the HIS-Q scales ($P < .05$ for all comparisons; Table 5). Responsiveness also was assessed using changes based on the CGI for Severity; the sexual domain ($P < .01$), sexual function subdomain ($P < .01$), energy domain ($P < .05$), and HIS-Q total score ($P < .01$) showed significant trends in the expected direction from baseline to week 2 and from baseline to week 12 (Table 6).

Table 4. Known-groups validity

HIS-Q score	CGI-S symptom severity categories (N = 177)									Overall F-test*		Pairwise comparison (P value)[†]
	No or very mild symptoms		Mild		Moderate		Severe					
	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	F	P value	
Sexual domain	34	30.99 (3.26)	35	41.73 (3.21)	73	46.78 (2.22)	30	57.86 (3.47)	11.29	<.0001	2[‖], 3[#], 5[‖]
Libido subdomain	34	38.24 (3.43)	36	46.99 (3.33)	73	48.74 (2.34)	30	56.94 (3.65)	4.76	.0033	3[‖]
Sexual function subdomain	35	26.61 (4.26)	36	37.85 (4.20)	73	45.21 (2.95)	30	58.54 (4.60)	9.37	<.0001	2[‖], 3[#], 5[‖]
Energy domain	35	31.19 (4.02)	37	41.89 (3.91)	73	48.63 (2.78)	30	57.22 (4.34)	7.35	.0001	2[‖], 3[¶]
Sleep domain	35	31.67 (3.19)	37	33.56 (3.10)	71	36.85 (2.24)	30	44.72 (3.45)	2.97	.0334	
Cognition domain	35	25.24 (2.75)	36	30.32 (2.72)	73	36.76 (1.91)	30	43.33 (2.97)	7.95	<.0001	2[‖], 3[¶], 5[‖]
Mood domain	35	23.16 (2.85)	37	31.18 (2.77)	72	35.81 (1.99)	30	41.67 (3.08)	7.40	.0001	2[‖], 3[¶]
HIS-Q total score	34	28.01 (2.21)	35	35.88 (2.18)	71	41.11 (1.53)	30	49.24 (2.36)	15.83	<.0001	2[#], 3[#], 5[¶], 6[‖]

HIS-Q Score	Total testosterone categories (n = 172)						Overall F-test*		Pairwise comparison (P value)[‡]
	<300 ng/dL		300−500 ng/dL		>500 ng/dL				
	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	F	P value	
Sexual symptoms domain	68	51.31 (2.39)	55	45.08 (2.65)	45	35.08 (2.93)	9.21	.0002	2[#], 3[‖]
Libido subdomain	68	54.04 (2.42)	56	47.62 (2.66)	45	39.81 (2.97)	6.94	.0013	2[¶]
Sexual function subdomain	68	49.26 (3.15)	55	43.75 (3.51)	47	31.52 (3.79)	6.56	.0018	2[¶]
Energy symptoms domain	68	51.59 (2.95)	56	42.86 (3.25)	47	37.94 (3.55)	4.69	.0104	2[‖]
Sleep symptoms domain	67	39.18 (2.31)	55	39.09 (2.54)	47	29.96 (2.75)	4.01	.0199	2[‖]
Cognition symptoms domain	68	37.99 (2.06)	55	33.64 (2.29)	47	29.61 (2.47)	3.45	.0342	2[‖]
Mood symptoms domain	68	36.71 (2.13)	55	33.77 (2.36)	47	27.89 (2.56)	3.55	.0310	2[‖]
HIS-Q total score	67	43.67 (1.68)	54	38.78 (1.87)	45	32.15 (2.05)	9.43	.0001	2[#]

CGI-S = Global Impression−Severity; HIS-Q = Hypogonadism Impact of Symptoms Questionnaire; LS = least-squares; SE = standard error.
[‖]P < .05; [¶]P < .001; [#]P < .0001.
*General linear model (PROC GLM). Pairwise comparisons between LS means were performed using the Scheffe test adjusting for multiple comparisons.
[†]Pairwise comparisons: 1 = no or very mild vs mild symptoms; 2 = no or very mild vs moderate symptoms; 3 = no or very mild vs severe symptoms; 4 = mild vs moderate symptoms; 5 = mild vs severe symptoms; 6 = moderate vs severe symptoms.
[‡]Pairwise comparisons: 1 = <300 vs 300−500 ng/dL; 2 = <300 vs >500 ng/dL; 3 = 300−500 vs >500 ng/dL.

Table 5. Responsiveness and anchor-based interpretation: HIS-Q score change by concept-specific anchor question score change

| HIS-Q score change | Change in anchor question | | | | | | | | Overall F-test* | | Pairwise comparison (P value)[†] |
| | Decline | | Stable | | Improved by 1 point | | Improved by ≥2 points | | | | |
	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	F	P value	
Sexual symptoms domain by sexual activity anchor											
Baseline to week 6	29	8.00 (2.61)	65	−1.48 (1.74)	42	−11.14 (2.17)	18	−33.33 (3.31)	36.46	<.0001	1‡, 2‖, 3‖, 4‡, 5‖, 6‖
Baseline to week 12	25	6.14 (2.97)	62	−1.73 (1.88)	40	−10.36 (2.35)	20	−26.07 (3.32)	20.64	<.0001	2§, 3‖, 4‡, 5‖, 6‡
Sexual symptoms domain by overall sexual function anchor											
Baseline to week 6	30	9.88 (2.69)	54	−2.91 (2.00)	41	−8.54 (2.30)	29	−24.88 (2.73)	28.76	<.0001	1‡, 2‖, 3‖, 5‖, 6§
Baseline to week 12	23	6.06 (3.10)	63	−0.96 (1.88)	27	−8.60 (2.86)	34	−21.64 (2.55)	20.22	<.0001	2‡, 3‖, 5‖, 6‡
Libido subdomain by libido anchor											
Baseline to week 6	39	8.55 (2.22)	64	−4.30 (1.73)	42	−12.10 (2.13)	9	−27.78 (4.61)	24.29	<.0001	1§, 2‖, 3‖, 4‡, 5‖, 6‡
Baseline to week 12	34	3.68 (2.91)	68	−1.23 (2.06)	33	−8.59 (2.95)	13	−25.00 (4.70)	10.38	<.0001	2‡, 3‖, 5§, 6‡
Sexual function subdomain by erectile function anchor											
Baseline to week 6	31	1.21 (4.43)	79	−4.11 (2.77)	26	−14.66 (4.84)	19	−21.05 (5.66)	4.40	.0054	3‡
Baseline to week 12	35	1.79 (3.71)	66	−5.49 (2.70)	26	−15.14 (4.30)	21	−22.32 (4.79)	6.51	.0004	2‡, 3‡, 5‡
Sexual function subdomain by overall sexual function anchor											
Baseline to week 6	31	14.92 (3.81)	54	−4.05 (2.89)	41	−10.37 (3.31)	29	−30.60 (3.94)	23.72	<.0001	1‡, 2‖, 3‖, 5‖, 6‡
Baseline to week 12	24	7.81 (4.03)	63	−0.69 (2.49)	27	−14.81 (3.80)	34	−26.65 (3.39)	19.16	<.0001	2‡, 3‖, 4‡, 5‖
Energy symptom domain by tiredness anchor											
Baseline to week 6	22	10.98 (4.32)	70	−3.10 (2.42)	52	−18.59 (2.81)	11	−37.88 (6.10)	20.48	<.0001	1‡, 2‖, 3‖, 4§, 5‖, 6‡
Baseline to week 12	26	9.94 (3.77)	56	−5.65 (2.57)	49	−20.24 (2.74)	16	−46.35 (4.80)	33.38	<.0001	1§, 2‖, 3‖, 4‡, 5‖, 6‡
Mood symptom domain by mood anchor											
Baseline to week 6	46	11.72 (1.87)	64	1.45 (1.58)	33	−8.86 (2.20)	12	−22.32 (3.65)	31.20	<.0001	1§, 2‖, 3‖, 4‡, 5‖, 6‡
Baseline to week 12	37	8.78 (2.14)	59	1.03 (1.70)	36	−6.75 (2.17)	16	−21.88 (3.26)	23.36	<.0001	1‡, 2‖, 3‖, 5‖, 6‡
Cognition symptom domain by cognition anchor											
Baseline to week 6	43	9.50 (2.15)	69	−0.36 (1.69)	29	−7.18 (2.61)	14	−19.05 (3.76)	17.53	<.0001	1‡, 2‖, 3‖, 5§
Baseline to week 12	30	8.33 (2.95)	75	−0.67 (1.86)	29	−10.92 (3.00)	14	−18.45 (4.31)	12.02	<.0001	2§, 3‖, 4‡, 5‡
Sleep symptom domain by sleep anchor											
Baseline to week 6	51	8.50 (2.12)	58	−2.16 (1.99)	39	−9.83 (2.42)	6	−31.94 (6.18)	19.47	<.0001	1‡, 2‖, 3‖, 5§, 6‡

(continued)

Table 5. Continued

HIS-Q score change	Change in anchor question											
	Decline		Stable		Improved by 1 point		Improved by ≥2 points		Overall F-test*		Pairwise comparison (P value)†
	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	F	P value	
Baseline to week 12	34	8.58 (2.66)	54	−3.55 (2.11)	48	−9.38 (2.24)	10	−33.33 (4.91)	21.15	<.0001	1‡, 2||, 3||, 5||, 6§
HIS-Q total score by overall hypogonadism anchor											
Baseline to week 6	32	7.81 (1.85)	45	1.38 (1.56)	46	−7.56 (1.55)	27	−17.41 (2.02)	33.79	<.0001	2||, 3||, 4‡, 5||, 6‡
Baseline to week 12	29	6.86 (2.16)	36	−1.09 (1.94)	48	−8.67 (1.68)	30	−15.11 (2.12)	20.60	<.0001	2||, 3||, 4‡, 5||

HIS-Q = Hypogonadism Impact of Symptoms Questionnaire; LS = least-squares; SE = standard error.
‡P < .05; §P < .001; ||P < .0001.
*General linear model (PROC GLM). Pairwise comparisons between LS means were performed using the Scheffe test adjusting for multiple comparisons.
†Group comparisons: 1 = decline vs stable group; 2 = decline vs improvement group 1; 3 = decline vs improvement group 2; 4 = stable vs improvement group 1; 5 = stable vs improvement group 2; 6 = improvement group 1 vs 2.

Responder Definition

Responder definitions were defined using anchor- and distribution-based methods. Anchor-based estimates were obtained from the mean scores for participants who improved by one point on the anchor questions for each domain (Table 5). The anchor-based and distribution-based responder estimates are presented in Figure 1, as is the final responder definition, which was determined by triangulating across all estimates (sexual = 7.0; libido = 8.3; sexual function = 12.5; energy = 16.6; sleep = 8.3; cognition = 8.3; mood = 7.0; Figure 1). Interpretation of the HIS-Q total score is complex, and the psychometric support for the contributing items and subscales varies; the responder definition for the HIS-Q total score is 7.0 (Figure 1).

DISCUSSION

The HIS-Q is a 28-item PRO instrument to assess changes in hypogonadal symptoms in response to TRT. The present study provides strong support for the psychometric properties of this newly developed instrument including the reliability, validity, and responsiveness in addition to information on the interpretation of the HIS-Q. To our knowledge, this is the first PRO that has been specifically developed for use in populations of hypogonadal men that comprehensively evaluates all the relevant domains, as identified through direct input from patients.

The HIS-Q demonstrated good correlations with other instruments designed to measure relevant domains of interest (eg, sexual, energy, mood, etc). These strong correlations and the comprehensive nature of the instrument suggest that it might be a useful end point in future studies evaluating TRT. The HIS-Q had moderate correlations with testosterone levels. Correlations between PRO and clinical measurements are often small to moderate in magnitude; this is expected because the value of a PRO measurement lies in gathering information from the patient that cannot be assessed through clinical measurements or other means.

It is important to note that the objective of this research was not to establish expected TRT treatment effects or estimate effects sizes for the instrument, but to ensure that the selected items performed appropriately for instrument finalization. Because of the observational study design and cohort selected for the present research (eg, many had relatively normal testosterone levels at baseline because the study required a "maintenance" group in which HIS-Q changes were not expected), reliable estimates of TRT treatment effects will need to be determined in future well-controlled studies.

Although it is possible to calculate a total score for the HIS-Q, this score should be carefully considered, adequately justified for specific research contexts, and interpreted with caution. Qualitative work results suggest that hypogonadism is highly heterogeneous in the symptoms experienced[5] and the psychometric properties vary across different HIS-Q domains. The HIS-Q domains measure distinct constructs and use of the composite

Table 6. Responsiveness: HIS-Q score change by CGI-S score change

| HIS-Q score change | CGI-S score change | | | | | | Overall F-test* | | Pairwise comparison (P value)[†] |
| | Decline (≥1) | | Stable (0) | | Improvement (≤−1) | | | | |
	n	LS mean (SE)	n	LS mean (SE)	n	LS mean (SE)	F	P value	
Sexual symptoms domain									
Baseline to week 2	13	7.69 (4.48)	76	−0.52 (1.85)	55	−8.31 (2.18)	6.71	.0016	1[‡], 2[‡]
Baseline to week 12	13	3.02 (4.73)	57	0.63 (2.26)	73	−11.99 (2.00)	10.57	<.0001	1[§], 2[‡]
Libido subdomain									
Baseline to week 2	13	3.21 (4.59)	78	−2.14 (1.87)	55	−5.61 (2.23)	1.70	.1861	
Baseline to week 12	14	−5.95 (4.81)	57	1.02 (2.38)	73	−8.33 (2.11)	4.39	.0142	1[‡]
Sexual function subdomain									
Baseline to week 2	13	11.06 (6.21)	78	0.08 (2.54)	55	−10.34 (3.02)	6.25	.0025	1[‡], 2[‡]
Baseline to week 12	14	4.02 (6.17)	58	0.54 (3.03)	73	−14.64 (2.70)	8.67	.0003	1[‡], 2[‡]
Energy symptoms domain									
Baseline to week 2	13	7.69 (5.91)	80	−7.40 (2.38)	55	−11.82 (2.87)	4.43	.0136	2[‡]
Baseline to week 12	16	1.04 (6.25)	57	−8.33 (3.31)	73	−17.01 (2.93)	4.24	.0162	2[‡]
Sleep symptoms domain									
Baseline to week 2	13	2.56 (4.27)	79	−1.69 (1.73)	55	−4.70 (2.08)	1.37	.2566	
Baseline to week 12	16	2.08 (4.65)	57	−2.78 (2.46)	72	−7.75 (2.19)	2.34	.0997	
Cognition symptoms domain									
Baseline to week 2	13	2.56 (4.50)	80	−0.10 (1.81)	55	−3.18 (2.19)	0.93	.3952	
Baseline to week 12	16	6.77 (4.38)	57	−1.02 (2.32)	73	−4.45 (2.05)	2.80	.0640	

CGI-S = Global Impression—Severity; HIS-Q = Hypogonadism Impact of Symptoms Questionnaire; LS = least-squares; SE = standard error.
[‡]P < .05; [§]P < .001.
*General linear model (PROC GLM). Pairwise comparisons between LS means were performed using the Scheffe test adjusting for multiple comparisons.
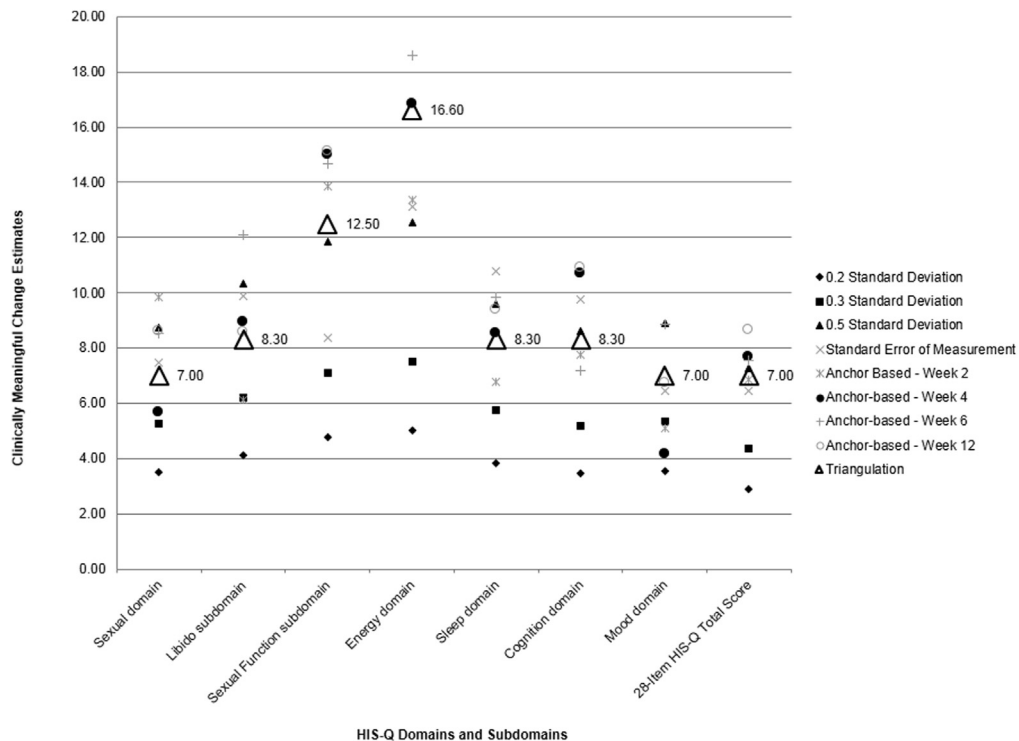[†]Improvement vs Stable; 2: Improvement vs Decline; 3: Stable vs Decline.

Figure 1. Summary of distribution- and anchor-based estimates of clinically meaningful change for the Hypogonadism Impact of Symptoms Questionnaire (HIS-Q).

HIS-Q total score as an end point might not capture treatment effects precisely. Analyzing specific domains of interest is recommended over using the total score. The sexual domain, libido and sexual function subdomains, and energy domain have the strongest psychometric properties and are likely the best outcome measurements for studies aiming to characterize TRT treatment effects.

The longer 2-week recall period was selected for the HIS-Q sexual domain because hypogonadal men often have lower levels of sexual activity and interest; thus, a longer period is appropriate to allow the opportunity for men to have relevant sexual experiences on which they can report. Shorter 1-week recall periods were selected for other domains. The recall periods that were selected for the HIS-Q domains are intended to strike a balance among accuracy of recall, respondent burden, and consideration for an adequate opportunity to engage in activities that form the basis for certain responses.

There were several limitations that should be noted. The etiology of hypogonadism for each patient participating in the study was not always known. In particular, the study team found that it was difficult to recruit patients with congenital secondary hypogonadism and those who were untreated. In addition, the men who participated in this study had testosterone levels that were not as low as expected at baseline, although the study was strengthened by the use of a central laboratory for the liquid chromatographic and tandem mass spectrometric testosterone testing. Moreover, some analyses in the present study (eg, exploratory factor analysis) could have benefited from a larger sample. Future studies of larger samples of men with hypogonadism will be used to further evaluate and confirm the factor structure of the HIS-Q. Future studies also will be aimed at further examining the performance properties of the HIS-Q and the relations of the instrument with clinical measurements in samples of symptomatic hypogonadal men with unequivocally low testosterone levels. It is important to note that the HIS-Q was neither designed nor tested as a screener for hypogonadism and its performance characteristics for this purpose are unknown.

CONCLUSIONS

The HIS-Q, a newly developed PRO instrument for the evaluation of hypogonadal symptoms, has demonstrated good reliability, validity, and responsiveness. The measurement has been developed in line with current Food and Drug Administration PRO guidance, and the final version of the measurement can be incorporated into TRT clinical trials. A shorter version of the instrument that lessens respondent burden and might be useful in clinical practice is currently under development.

ACKNOWLEDGMENTS

Corresponding Author: Heather L. Gelhorn, PhD, Evidera, 7101 Wisconsin Avenue, Suite 1400, Bethesda, MD 20814, USA; E-mail: heather.gelhorn@evidera.com

STATEMENT OF AUTHORSHIP

Category 1

(a) Conception and Design
Heather L. Gelhorn; Ebony Dashiell-Aje; Michael G. Miller; Dennis A. Revicki

(b) Acquisition of Data
Heather L. Gelhorn; Ebony Dashiell-Aje; Michael G. Miller; Dennis A. Revicki

(c) Analysis and Interpretation of Data
Heather L. Gelhorn; Ebony Dashiell-Aje; Michael G. Miller; Leonard R. DeRogatis; Adrian Dobs; Allen D. Seftel; Stanley E. Althof; Meryl Brod; Dennis A. Revicki

Category 2

(a) Drafting the Article
Heather L. Gelhorn; Ebony Dashiell-Aje

(b) Revising It for Intellectual Content
Michael G. Miller; Leonard R. DeRogatis; Adrian Dobs; Allen D. Seftel; Stanley E. Althof; Meryl Brod; Dennis A. Revicki

Category 3

(a) Final Approval of the Completed Article
Heather L. Gelhorn; Ebony Dashiell-Aje; Michael G. Miller; Leonard R. DeRogatis; Adrian Dobs; Allen D. Seftel; Stanley E. Althof; Meryl Brod; Dennis A. Revicki

REFERENCES

1. Novak A, Brod M, Elbers J. Andropause and quality of life: findings from patient focus groups and clinical experts. Maturitas 2002;43:231-237.

2. Lee KK, Berman N, Alexander GM, et al. A simple self-report diary for assessing psychosexual function in hypogonadal men. J Androl 2003;24:688-698.

3. Heinemann LAJ, Zimmerman T, Vermeulen A, et al. A new 'Aging Males' Symptoms' (AMS) rating scale. Aging Male 1992;2:105-114.

4. Hayes RP, Henne J, Kinchen KS. Establishing the content validity of the Sexual Arousal, Interest, and Drive Scale and the Hypogonadism Energy Diary. Int J Clin Pract 2015; 69:454-465.

5. Gelhorn HL, Vernon MK, Stewart KD, et al. Content validity of the Hypogonadism Impact of Symptoms Questionnaire (HIS-Q): a patient-reported outcome measure to evaluate symptoms of hypogonadism. Patient 2016;9:181-190.

6. Food and Drug Administration. Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Fed Regist 2009;74:65132-65133.

7. Rosen RC, Riley A, Wagner G, et al. The International Index of Erectile Function (IIEF): a multidimensional scale for assessment of erectile dysfunction. Urology 1997;49:822-830.

8. Flynn KE, Jeffery DD, Keefe FJ, et al. Sexual functioning along the cancer continuum: focus group results from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)). Psychooncology 2011;20:378-386.

9. Yu L, Buysse DJ, Germain A, et al. Development of short forms from the PROMIS Sleep Disturbance and Sleep-Related Impairment item banks. Behav Sleep Med 2011;10:6-24.

10. Becker H, Stuifbergen A, Morrison J. Promising new approaches to assess cognitive functioning in people with multiple sclerosis. Int J MS Care 2012;14:71-76.

11. Ware JE, Kosinski M, Turner-Bowker DM, et al. How to score version 2 of the SF-12 Health Survey. Lincoln, RI: Quality Metrics; 2002.

12. Jones PW, Chen WH, Wilcox TK, et al; Group E-PS. Characterizing and quantifying the symptomatic features of COPD exacerbations. Chest 2011;139:1388-1394.

13. RUMM. 2030: Rasch unidimensional measurement models. Australia: RUMM Laboratory Pty Ltd; 2010.

14. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;163:297-334.

15. Hays RD, Revicki DA. Reliability and validity, including responsiveness. In: Fayers PM, Hays RD, eds. Assessing quality of life in clinical trials. 2nd ed. New York: Oxford University Press; 2005. p. 25-39.

16. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

17. Wyrwich KW, Nienaber NA, Tierney WM, et al. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469-478.

18. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52:861-873.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jsxm.2016.09.006.