Measurement of Ocular Surface Irritation on a Linear Interval Scale with the Ocular Comfort Index

Michael E. Johnson[1,2] *and Paul J. Murphy*[1]

Purpose. To examine the psychometric properties of the Ocular Comfort Index (OCI), a new instrument that measures ocular surface irritation designed with Rasch analysis to produce estimates on a linear interval scale.

Methods. The OCI was self-completed by 452 subjects. Some of them repeated the questionnaire, to aid in determining its reliability and test–retest repeatability. Ten versions were produced to evaluate question order effects. In addition, three construct hypotheses were tested to verify that the OCI was measuring what was intended, concordance with the Ocular Surface Disease Index (OSDI), the relationship with tear break-up time (TBUT), and the change in TBUT after the use of ocular lubricants in individuals with moderate dry eye.

Results. A 12-item OCI was developed with well-functioning items and categories: 95% confidence interval for the intraclass correlation coefficient = 0.81 to 0.91; person separation = 2.66; item separation = 11.12; and 95% repeatability coefficient = 13.1 units (0–100 scale). The ordering of items had no effect on OCI measures ($P = 0.41$). The OCI measure exhibited a positive correlation with the OSDI score ($P < 0.0001$) and a negative correlation with TBUT ($P < 0.0001$) and was able to detect improvement in symptoms of dry eye in individuals before and after treatment ($P < 0.0001$).

Conclusions. The OCI was shown to have favorable psychometric properties that make it suitable for assessing the impact of ocular surface disease on patient well-being and changes in severity brought about by disease progression or therapeutic strategies. (*Invest Ophthalmol Vis Sci.* 2007;48:4451–4458) DOI:10.1167/iovs.06-1253

Measurement of the level of discomfort caused by ocular surface disease is currently limited by the shortcomings of available questionnaires. For example, the McMonnies survey was developed to assist the diagnosis of dry eye syndrome (DES) by considering epidemiologic risk factors, the frequency of symptoms of ocular irritation, and sensitivity to environmental triggers.[1,2] Notwithstanding its usefulness in diagnosis,[3] the McMonnies score cannot be relied on as an indicator of symptom severity, defined in this work as an aggregate function of frequency and intensity, because this component of its tally is combined with unrelated others. The McMonnies survey is also unsuitable for appraising temporal changes, because the responses to its epidemiologic questions are likely to be identical at different time points, and its questions relating to environmental triggers introduce noise if they have not been experienced between replicate testing.

The Ocular Surface Disease Index (OSDI) was developed more recently to grade the severity of DES and is notable among other questionnaires for ocular surface disease for having undergone psychometric testing and having been accepted by the U.S. Food and Drug Administration (FDA) for use in clinical trials.[4–6] This instrument has a 12-item, five-category Likert design with three subscales that sequentially ask about symptoms of ocular irritation and the impact on vision-related functioning and environmental triggers of DES. It cannot be assumed that the difficulty step between each category is constant or that the difficulty of all questions is comparable, from which it follows that its scale may not be additive or linearly related to symptom severity. Instead, the score of the OSDI should be interpreted as a relative ordering of person afflictions.[7,8] This limits what can be inferred from the OSDI, because ordinal rankings are less suited to the estimation of any effect size than are interval data.[9] Also its gains in applicability made by investigating several symptom domains are offset by reductions in its interpretative potential.[10] Furthermore, its handling of missing data, caused by omitted or "not applicable" responses, by expressing the score as a percentage of the maximum possible value of the questions answered, is inadequate; higher percentages will be achieved if difficult items, those that are more likely to score lower, are not answered.

The Ocular Comfort Index (OCI) was conceived in response to deficiencies in existing instruments for use in clinical trials. The OCI was designed and tested with Rasch analysis, which calibrates the "difficulty" of items (ρ) along the latent variable of interest that act as marks on a ruler against which person "ability" (α) can be compared.[11] These terms stem from the technique's origins in aptitude testing. In this context, more difficult items are those that tend to receive lower scores, and more able persons experience a greater degree of discomfort. In Rasch models the probability of an observed response by a person to an item is related to their functional ability, which is defined as the difference between their ability and the item's difficulty ($\alpha - \rho$). For dichotomous 0 or 1 responses the ability of a subject is equated to the difficulty of items with which they have a 50% chance of success: $P(1|\alpha,\rho) = 0.5$ when $\alpha = \rho$. Polytomous items with m categories can be considered to represent $m - 1$ marks on a ruler to which persons taking the item are compared, rather than just the one with binary response structures. Here, a person has a 50% chance of responding with category x rather than category $x - 1$ when his or her ability matches the difficulty of the item summed with the step calibration of the category (τ_x): $P(x|\alpha,\rho) = 0.5$ when $\alpha = \rho + \tau_x$.[12] The theory of Rasch analysis was expounded in a recent review article.[13] The foremost merit of these methods is that derived values meet the requirements of a noninteractive conjoint structure, so item difficulty and person ability can be estimated from observed responses without ambiguity.[14]

The purpose of this study was to develop and test the validity of a new instrument capable of measuring the severity of discomfort caused by ocular surface disease for use in clinical trials.

From the [1]Contact Lens and Anterior Eye Research Group, School of Optometry and Vision Sciences, Cardiff University, Cardiff, United Kingdom; and the [2]Optometry Department, Bristol Eye Hospital, Bristol, United Kingdom.

MEJ is supported by a research scholarship from Ultralase Ltd.

Submitted for publication October 18, 2006; revised December 8, 2006, and March 6, 2007; accepted August 20, 2007.

Disclosure: **M.E. Johnson**, Ultralase, Ltd. (F); **P.J. Murphy**, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Michael E. Johnson, Cardiff University, School of Optometry and Vision Sciences, King Edward VII Avenue, Cardiff, CF10 3NB, Wales, UK; JohnsonM2@cardiff.ac.uk.

METHODS

Item Construction

Areas of questioning were identified from a review of the literature and interviews of patients. Appraisal for face validity and redundancy reduced these to eight areas: one positive (comfort), which was reversed for analysis, and seven negative (dryness, grittiness, itching, pain, stinging, tiredness, and visual stability). From these, 15 items were generated—each negative area was split into two subparts that sequentially inquired about frequency and intensity (Table 1).

Too few category response options reduce information yield, whereas too many exceed discriminatory ability.[15] The optimum number is typically seven, and so this number was initially chosen, but was confirmed empirically in subsequent analyses.[16] Only the extreme responses were labeled (0, never; 6, always), to minimize the effects of differences in adjective interpretation.

Ten versions of the OCI were produced, all of which started with item 1 (comfort) but varied in the ordering of negative item pairs. This was done to establish the influence of question order and was achieved using randomly generated number sequences. Block randomization, with blocks of 10 digits, was used to determine the version that each subject received for all subparts of the study.

Subjects

The total pool of subjects numbered 452 (median age, 34 years; range, 18–75 years; 154 men and 298 women). The different arms of the study used subgroups of this sample, as described subsequently and summarized in Table 2. Subjects were recruited from students and staff at Cardiff University and the University Hospital of Wales by postings on notice boards and an electronic mail circulated through the organizations' mailing lists and from patients attending for optometric eye examinations in private practice through literature in the waiting area and clinician invitation. All questionnaires were self-completed and returned via drop in boxes or by mail. Major exclusion criteria were language barriers to understanding, overt cognitive impairment, and visual acuity worse than 6/9 in either eye. Informed consent was obtained, and all procedures adhered to the tenets of the Declaration of Helsinki. Ethical approval was obtained from the Cardiff School of Optometry and Vision Science Ethical Committee.

Psychometric Properties of the OCI

An exploratory analysis on the initial completion of the OCI by all subjects to all items was undertaken to detect items and individuals that did not fit the Rasch model. It is generally accepted that regardless of their descriptive merits, such questions should be removed, because the Rasch paradigm is the only item–response method that conforms to the tenets of axiomatic measurement theory, and so items must conform to its expectations, rather than the converse.[13] The appropriate treatment of nonfitting persons is more controversial. Some authorities argue that it is not possible to assess everyone equally well and that those with idiosyncratic response patterns should be ignored during calibration, to uphold the legitimacy of Rasch-derived measures for the majority,[17] but to others the omission of data is never appropriate. In this work, such individuals were removed for instrument calibration purposes, although the influence of their exclusion on item difficulty estimates was gauged by a comparison analysis with them included.

The issue of empiric concordance with the idealizations of the Rasch model for both items and persons is assisted by fit statistics that reflect differences between observed and expected responses, or response residuals. Goodness of fit is evaluated with weighted mean squared residual errors across persons for each item and across items for each person. Two different weighting schemes are used. In the first, for each person–item encounter the squared residual error is normalized to the expected variance; this normalized squared residual is called the "outfit" statistic because it is sensitive to outlying errors. In the second, the mean squared residual is normalized to the average expected variance; this normalized mean squared residual is called the "infit" because it is most representative of inlying errors. Outfit and infit statistics can be expressed in a mean square (MNSQ) ratio-scale form with expectation 1 and range 0 to $+\infty$, $1 + \delta$ indicates $100 \times \delta$ percent more variation between the observed and model-predicted response patterns than would be expected if the data and the model corresponded perfectly; or as z-scores in SD units (ZSTD), expectation 0, and range $-\infty$ to $+\infty$.[18–20]

The first 100 subjects recruited, excluding those from private practice, were asked to repeat the questionnaire in 14 ± 7 days, to allow assessment of its test–retest repeatability and reliability; five of these subjects were lost to follow-up.

Construct Hypotheses

Three construct hypotheses were tested to corroborate that the OCI was measuring ocular surface discomfort as intended. Several hypotheses were used because there is no single gold-standard comparative. First, the OCI was tested for concordance with the current best available instrument for assessing symptom severity, the OSDI, in all subjects other than those recruited from private practice ($n = 337$). These questionnaires were supplied together with the topmost instrument alternated. Second, the relationship between the OCI and the median of three recordings TBUT in one randomly selected eye in the portion of these 337 subjects who were agreeable ($n = 102$) was

TABLE 1. Fit Statistics for All 15 Items in the Preliminary Analysis

Item	Infit		Outfit	
	MNSQ	ZSTD	MNSQ	ZSTD
1. In the last week, did your eyes feel comfortable?	0.76	(−4.0)	0.82	(−2.5)
2. In the last week, how often did your eyes feel dry?	1.07	(1.1)	0.98	(−0.3)
3. When your eyes felt dry, typically, how intense was the dryness?	0.90	(−1.6)	0.82	(−2.4)
4. In the last week, how often did your eyes feel gritty?	1.13	(1.9)	1.02	(0.3)
5. When your eyes felt gritty, typically, how intense was the grittiness?	0.92	(−1.1)	0.83	(−2.1)
6. In the last week, how often did your eyes feel stingy?	1.02	(0.3)	0.93	(−0.8)
7. When your eyes stung, typically, how intense was the stinging?	0.88	(−1.8)	0.79	(−2.4)
8. In the last week, how often did your eyes feel tired?	1.13	(1.8)	1.19	(2.5)
9. When your eyes felt tired, typically, how intense was the tiredness?	0.97	(−0.5)	0.99	(−0.1)
10. In the last week, how often did your eyes feel painful?	0.93	(−0.9)	0.83	(−1.8)
11. When your eyes felt painful, typically, how intense was the pain?	0.95	(−0.6)	0.90	(−1.0)
12. In the last week, how often did your eyes itch?	1.12	(1.8)	1.15	(1.9)
13. When your eyes itched, typically, how intense was the itching?	1.05	(0.8)	1.10	(1.2)
14. In the last week, how often did your vision change between clear and blurred?	1.37	(5.0)	1.34	(3.8)
15. When your vision was changeable, how bothersome was it?	1.15	(2.2)	1.05	(0.6)

TABLE 2. Subject Information for the Various Arms of the Study

Study Arm	Number	Median Age (y)	Proportion Female	Median OCI Score (0–100 Scale)	Comparison of Median OCI Score with Total
Repeated OCI	95	33.5	(70/95) 74%	35	$P = 0.92$
OCI versus OSDI	337	29	(223/114) 66%	33	$P = 0.22$
OCI versus TBUT	102	29	(42/102) 59%	44	$P < 0.001$
Ocular lubricants	65	38	(39/65) 60%	49	$P < 0.001$
Total	452	34	(154/452) 66%	35	—

The column on the far right shows the results of a comparison between differences in the average OCI score of the subpopulation versus that for the overall sample, using the Mann-Whitney test due to skewed data. Predictably, considering their recruitment criteria, subjects in the ocular lubricant study arm tended to have more severe symptoms than those in other subsets. This also occurred in the TBUT comparison group, revealing some self-selection bias, with the subjects experiencing worse symptoms more likely to volunteer for additional tests.

quantified.[21] Third, the change in the OCI score after the use of ocular lubricants for 28 ± 3 days, either 0.3% carbomer 934 (Lacryvisc; Alcon, Hünenberg, Switzerland) or 0.18% sodium hyaluronate (Vismed; TRB Chemedica AG, Haar/Munich, Germany), in subjects with moderate DES ($n = 65$) was investigated. These subjects were recruited from the 150 subjects with the highest OCI scores, including those from private practice, who met the following criteria: TBUT <10 seconds and staining of the cornea with fluorescein and bulbar conjunctiva with lissamine green between grades 1 the 3, inclusively, with the Oxford scheme.[22] The latter two hypotheses presuppose that the severity of symptoms is inversely proportional to TBUT and that the use of these ocular lubricants reduces symptoms, respectively; assumptions that are in line with accepted thinking.[23–26]

Rasch Analysis

Questionnaires were processed with Rasch analysis software (Winsteps, ver. 3.58.1; Winsteps, Chicago, IL).[27,28] Rasch methods are based on the natural logarithm of odds ratios and so measure both item difficulties and person abilities in log-odd units (logits). Each logit step increases the odds of observing the event specified in the measurement model by a factor of 2.72. This scaling can later be linearly transformed to a more user-friendly 0-to-100 scale that avoids the use of negative values.

Statistics

Factor analysis supplemented the Rasch fit statistics to verify that the items were unidimensional.[29] Using this technique, the linear combination of the items that best accounted for variation in the response matrix variables was found. Correlations of each item to this principle factor were calculated.

Repeatability is the ability of an instrument to obtain the same score given identical conditions. It was calculated as confidence intervals (CIs) for the variance of replicate measures about their mean to give the 95% repeatability coefficient (R_c).[30]

The reliability of person measures indicates how much an observed score reflects the true value relative to measurement error. The true variability of a sample can be estimated with replicate testing using the intraclass correlation coefficient (ICC), which is based on the separation of sources of variance with the repeated-measures analysis of variance (ANOVA) method. The ICC gives the correlation of the instrument's scale with a hypothetical one that truly measures what it is supposed to and $\sqrt{[ICC/(1 - ICC)]}$ is equivalent to the signal-to-noise ratio.[31] In addition, the true variability of the sample was estimated from the initial questionnaire completed by all subjects using the mean squared standard error of the misfit of persons from the Rasch model. When calculated this way, reliability was reported as the ratio of the estimated true standard deviation of the sample to the measurement error to give a separation index that is directly equivalent to the signal-to-noise ratio.[18] A similar methodology was used to compute the reliability of item difficulty estimates.

Correlation coefficients between variables were calculated with either the Pearson method (r) when both variables were on interval scales, or with the Spearman ranked method (r_s) when this was not the case. The sampling distribution of these statistics is not normal, and so a Fisher transformation was used to facilitate the calculation of CIs.[9]

RESULTS

Preliminary Analysis

The fit statistics of the item asking about visual stability, referred to hereafter as blur, was very high (Table 1; Fig. 1), which indicates inconsistent responses about this question relative to other items, and therefore that it probed a different latent trait or was often misunderstood. Conversely, the fit statistic of the item inquiring about comfort was very low, indicating that the sample were responding to this question in an overly predictive way, intimating redundancy and poor sensitivity to modifying variables.

Idiosyncratic response patterns were not clustered at either end of the range of person measures of ocular discomfort generated by the 15-item instrument (Fig. 2); (22/452) 4.9% of

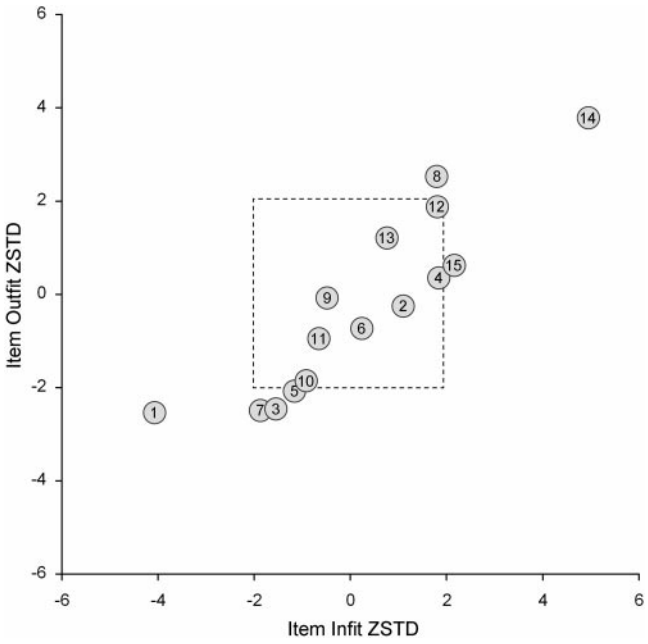

FIGURE 1. Infit versus outfit MNSQs expressed as *z*-scores in standard deviation units for all 15 items in the preliminary analysis. The *box* encloses items that are within ± 2 ZSTD of the value expected by the Rasch model of this data set; items outside have less than a 5% probability of being compatible. Items located a long way from this box are therefore more likely to corrupt, rather than to contribute, to measurement.

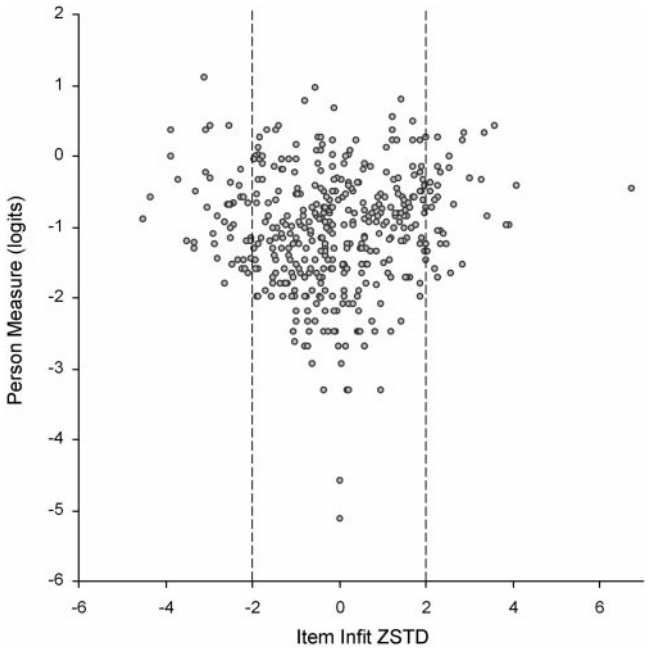

FIGURE 2. Person infit, also termed information-weighted fit (nonoutlier sensitive), expressed as *z*-scores versus person measures in logits for all 452 subjects in the preliminary analysis. *Dashed vertical lines* enclose the region that is ± 2 ZSTD from the value expected by the Rasch model. Persons located outside this zone have increasingly idiosyncratic response patterns, compared with most of the respondents.

the subjects had infit MNSQs that differed by more than three standard deviations from the expected value (compared with <0.14% for a normal distribution). The questionnaires of these anomalous subjects were reviewed for transcription errors, but no explanatory cause was established.

Consequently, items 1 (comfort), 14 and 15 (frequency and intensity of blur), and subjects with either an infit or outfit ZSTD that exceeded ±3 SD were removed to clean the data for instrument optimization.

Category Structure

Rating category step calibrations and empiric average category measures advanced in an ordered manner, indicating that people could discriminate reliably between them (Table 3). However, the advances were relatively small between categories 1, 2, 3, and 4; seen visually as narrow probability curves in Figure 3. This is preferable to very large differences, where less informative dead zones appear between categories, but indicates considerable overlap and alludes to functional redundancy. Accordingly, the effect of merging categories was assessed. Numerous single or combinations of category mergings were tried: 0–1, 1–2, 2–3, 3–4, 1–2 and 3–4, and 1–2–3. All these were adjudged to be inferior to the original category structure because they reduced person separation and tended to worsen item fit.

Item Calibration

After the removal of items 1, 14, and 15, the fit statistics of the remaining 12 questions improved so as to meet accepted guidelines (Table 4).[32,33] The unidimensional nature of the reduced number of questions was supported by unrotated factor analysis, which found a principle factor with correlations with the individual items that ranged from 0.63 to 0.79. A comparison of item difficulty estimates with the 22 misfitting persons included gave similar values that differed by <0.07 logits.

Person and Item Estimates

Inspection of the person ability/item difficulty map in Figure 4, where both subjects and items appear along the same scale, a linear transformation of the Rasch logit scale running from 0 to 100 units (8.92 × logit value + 45.17), indicates that items targeted the upper end of discomfort in the sample. In questionnaire design, average item difficulty is often manipulated to coincide with average person ability by selectively removing items that target relatively few people, but this was not done here because it is envisioned that the OCI will be used in settings where extreme degrees of ocular irritation are encountered more commonly than in the study's sample. Item separation was 11.12, and person separation was 2.66, which indicates stable item difficulty estimates and good instrument ability to differentiate between persons.

Influence of Question Order

The mean OCI measure of the 10 versions of the 12-item OCI did not significantly differ ($P = 0.41$; one-way ANOVA), suggesting that the order of questions does not influence person response patterns.

Reliability and Repeatability

The 95% CI for the two-way random-effects ICC of the OCI for the test sample was 0.81 to 0.91, and the instrument's 95% R_c when transformed to a 0-to-100 scale was 13.1 units.

Construct Hypotheses

The OCI exhibited reasonable concordance with the OSDI: 95% CI for r_s was 0.68 to 0.78 ($P < 0.0001$). OCI measures tended to be greater than OSDI scores in subjects with mild degrees of discomfort but lower for those with high levels (Fig.

TABLE 3. Category Diagnostics for the Secondary Analysis after the Removal of Grossly Misfitting Items and Persons

Response	Frequency	Infit MNSQ	Outfit MNSQ	μ_x	τ_x	ω_{min}	ω_{max}
0	1625	0.95	0.97	−2.10	—	−∞	−2.12
1	911	1.00	0.89	−1.45	−1.19	−2.12	−1.17
2	743	1.05	0.89	−0.97	−1.05	−1.17	−0.60
3	658	0.95	0.90	−0.52	−0.67	−0.60	0.03
4	682	1.01	1.03	−0.14	−0.39	0.03	1.04
5	297	1.18	1.21	0.30	0.95	1.04	2.77
6	53	1.38	1.20	0.52	2.35	2.77	+∞

μ_x, the mean ability of the people observed in each category (x); τ_x, the rating category step calibration, defined as the functional ability for which the probability of response x equals the probability of response $x - 1$ when other responses are restricted—the bottom category has no prior transition and so has no value; ω_{min} and ω_{max}, the relative logit values of the category boundaries—that is, the functional ability at which persons are expected to change the response.

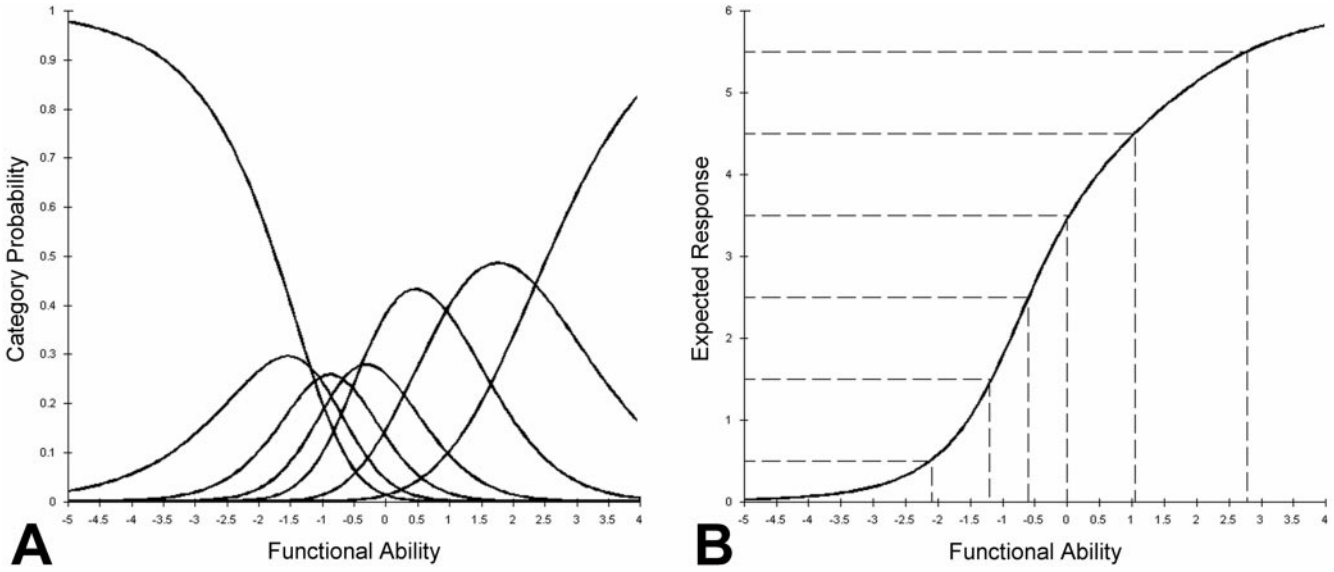

FIGURE 3. Estimates of response probabilities with seven categories (**A**) and the expected response (**B**) as a function of functional reserve, based on the response matrix after the removal of grossly misfitting items and persons. The category probabilities are unimodal and sequentially ordered, and each is, for some functional ability, the most likely to be chosen. The intersections of the *dashed vertical lines* with the abscissa in the expected-response graph correspond with the category boundaries (ω_{min} and ω_{max}) listed in Table 3.

5). Extreme floor response patterns (all zeros) were included in these calculations: 3% (9/337) and 8% (26/337) with the OCI and OSDI, respectively. There were no ceiling response patterns (all maximum responses). In addition, 31% (104/337) of respondents had not experienced one or more environmental triggers of the OSDI in the past week.

The logarithm of median TBUTs, transformed to reduce the positive skew of their distribution, correlated inversely with OCI score (Fig. 6): 95% CI for r was -0.23 to -0.56 ($P < 0.0001$).

The OCI was able to detect the improvement in symptoms of individuals with DES before and after treatment with ocular lubricants: 95% CI of the treatment difference was -5.5 to -8.0 units ($P < 0.0001$; paired t-test).

DISCUSSION

Symptoms may be present most of the time but mild and vice versa, and so items asking about the frequency and intensity of symptoms could have probed different latent traits in violation of the requirement of the Rasch model for unidimensionality. It was therefore reassuring that the fit statistics of all questions were within suggested guidelines.[32,33] Indeed, the perfect pairing of the difficulties of items that asked about the same symptom suggests that most individuals did not differentiate between frequency and intensity, which is consistent with the reports of others.[34]

The range of average item difficulties of the 12-item OCI was relatively narrow (-1.14 to 0.74 logits). This result threatens to limit the range of persons for whom the instrument performs well, because the information yield of each item is inversely proportional to the disparity between its difficulty and the ability of the person taking that item. However, the instrument's range of applicability is broadened by its polytomous responses that differ more in their difficulties (-2.72 to 3.60 logits) than the items that, with perhaps the exception of tiredness and pain, were essentially synonymous. The similarity of item difficulties may account for the variation in relative frequency of various symptoms in dry eye populations reported in the literature. Toda et al.[35] found, similar to this study, that ocular fatigue is the most common complaint of these patients in Japan, above dryness and pain; whereas,

TABLE 4. Estimates of Item Difficulties and Fit Statistics from the Secondary Rasch Analysis after the Removal of Grossly Misfitting Items and Persons

Item	Difficulty (logits)	SE	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
11. Pain (int.)	0.74	0.05	0.91	(−1.1)	0.85	(0.61)
10. Pain (freq.)	0.66	0.05	0.96	(−0.5)	0.87	(0.62)
7. Sting (int.)	0.36	0.05	0.86	(−2.0)	0.75	(0.68)
6. Sting (freq.)	0.26	0.05	0.99	(−0.2)	0.89	(0.67)
5. Gritty (int.)	0.25	0.05	0.98	(−0.2)	0.90	(0.66)
4. Gritty (freq.)	0.12	0.05	1.16	(2.2)	1.04	(0.66)
13. Itch (int.)	0.09	0.05	1.09	(1.2)	1.15	(0.63)
12. Itch (freq.)	−0.04	0.04	1.17	(2.4)	1.19	(0.64)
3. Dryness (int.)	−0.14	0.04	0.89	(−1.7)	0.81	(0.72)
2. Dryness (freq.)	−0.33	0.04	1.08	(1.2)	1.00	(0.73)
9. Tiredness (int.)	−0.82	0.04	1.00	(0.0)	1.03	(0.69)
8. Tiredness (freq.)	−1.14	0.04	1.12	(1.6)	1.20	(0.69)

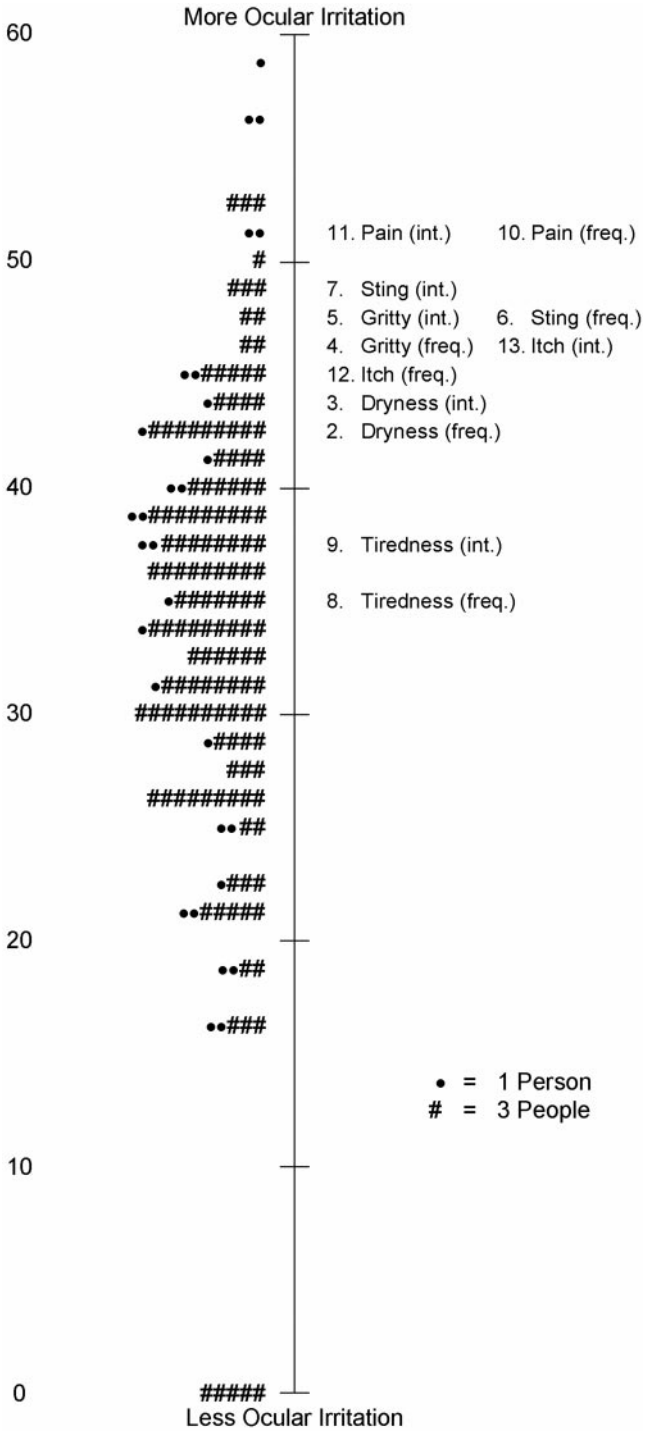

FIGURE 4. Person ability/item difficulty map for the 12-item OCI. Persons and items are located to the left and right of the vertical line, respectively; both appear along the same abstract scale, which is a linear transformation of the Rasch logit scale that runs from 0 to 100 units. More able people (those with more ocular surface discomfort) and harder items (those more likely to beget lower responses) are at the top of the scale, and less able people and easier items at the bottom. Each item is located at its average difficulty and could be exploded into its seven categories.

Begley et al.[34] and Nichols et al.[36] independently reported that dryness was more common than tiredness in North America. Alternative explanations for these discrepancies include geographic variability in the interpretation of adjectives used to describe symptoms or differences in the wording of questions between studies.

Item order did not influence response patterns. This result was anticipated, because the influence of contextual factors is generally limited to when questions ask about attitudes or are emotionally weighted.[37]

The repeatability and reliability of the OCI were acceptable, particularly considering that the symptoms of ocular surface disease are known to exhibit considerable variability and so observed differences between replicate testing would have embodied both measurement error and real person variation.[34]

The OCI exhibited a moderate positive correlation with the OSDI and a moderate negative correlation with TBUT, as predicted. That the strength of these correlations was not greater is not necessarily a cause for concern, because the OSDI differs from the OCI in that it probes several, albeit related, dimensions; and a low TBUT is just one of many causes of ocular surface irritation. Further authentication of the premise that the OCI evaluates ocular discomfort was that its score improved in subjects with DES after treatment.

Floor response patterns may result from the complete absence of symptoms or poor instrument sensitivity. The OCI elicited such responses less often than did the OSDI; indeed, the developers of the OSDI reported that an even greater proportion of their sample (12.2%) responded this way.[4] The discrepancy suggests that the OCI is better able to measure milder degrees of discomfort than is the OSDI. Another concern for the OSDI was the high proportion of subjects who responded "not applicable" to one or more of its environmental trigger items, reducing the precision of its estimate of ocular discomfort in these cases and, because it is based on raw data counts, altering test difficulty in unknown ways.

A drawback of Rasch analysis is that it cannot estimate person measures for extreme floor/ceiling raw scores. The complete absence of symptoms or the notion that symptoms could not be worse is at odds with its philosophy, yet the rejection of any data is undesirable in clinical trials. Several methods have been proposed to generate definite measures for such response patterns that assume that an extreme score implies a measure only slightly out of the range of the test.[38] The software used in this work assigned 0 scores a value of 0.3 score points and subtracted 0.3 score points from maximum scores to allow the estimation of person measures.[39] This was considered when the OCI was linearly rescaled so that extreme raw scores correspond to its measurement scale bounding values of 0 and 100.

In the calibration of the OCI, approximately 5% of subjects were excluded because, based on statistical considerations, their response patterns were deemed incompatible with the Rasch model for the whole data set. These subjects were excluded to ensure that the instruments measures were valid in terms of measurement theory for most of the respondents. However, it is likely that if the OCI is used in clinical trials some subjects will respond in abnormal ways, as identified by their fit statistics. The OCI does generate measures of discomfort for these persons, although of relatively low precision, and so they can be included in any analysis. It would, however, be prudent to check such data for transcription errors and to investigate whether these subjects are unusual in any other regard. Also, subsequent analysis can be repeated with and without these persons. The inclusion of misfitting persons is unlikely to have a significant effect on results unless they constitute a relatively large proportion of the study sample.

Another issue for those using the OCI is what level of significance to ascribe to the units of its scale. As an interval measure, it can be surmised that an increase from 5 to 10 units denotes the same increase as from 15 to 20 units, although it cannot be assumed that this represents a doubling of symptom

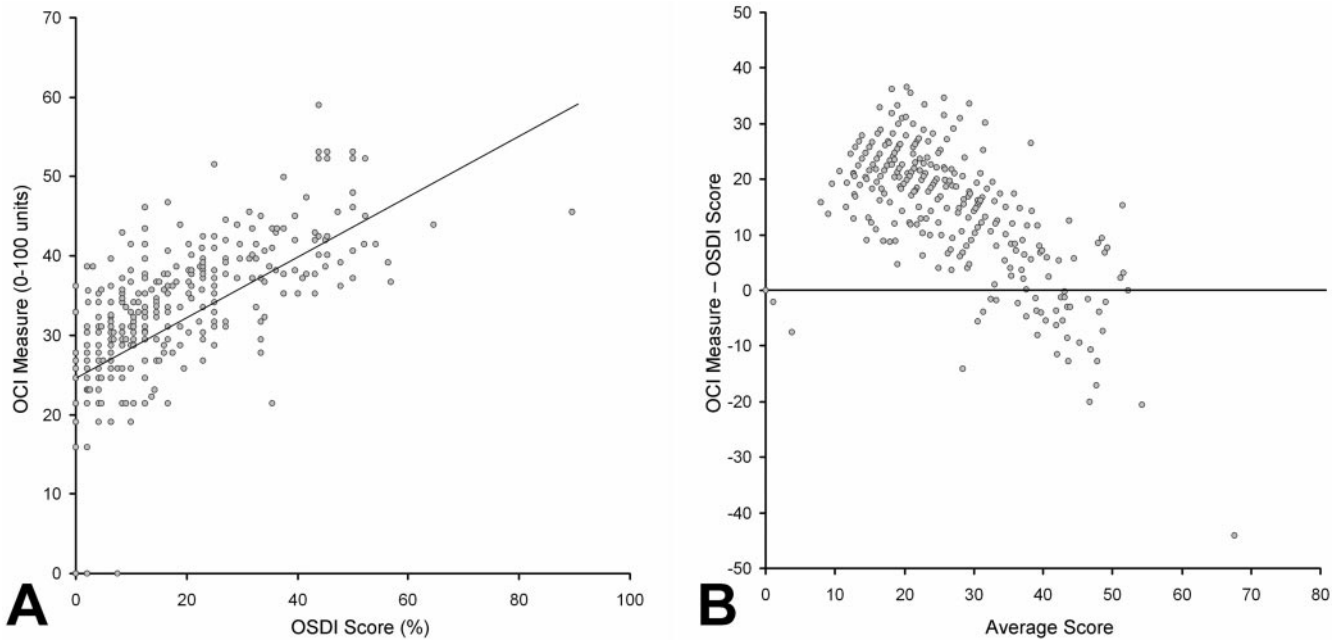

FIGURE 5. OCI measures versus OSDI scores with the best-fitting linear regression (**A**) and a difference-average plot of the same data (**B**). The OCI score tends to exceed the OSDI score for low levels of ocular discomfort; this situation is reversed for higher levels of ocular discomfort.

severity as it would if it were a ratio scale. However, these changes have no intrinsic clinical significance. Clinical significance must be ascertained by future work that compares changes in instrument scores with minimally important changes defined on external criteria.[40] Of note in this regard, in this study the use of ocular lubricants in subjects with DES was moderately well appreciated and typically reduced the OCI score by more than six units. Based on these data, it seems reasonable to suggest that that changes of three or more units are likely to be noticed by patients and therefore that this step can be regarded as an estimate of a minimally important treatment difference.

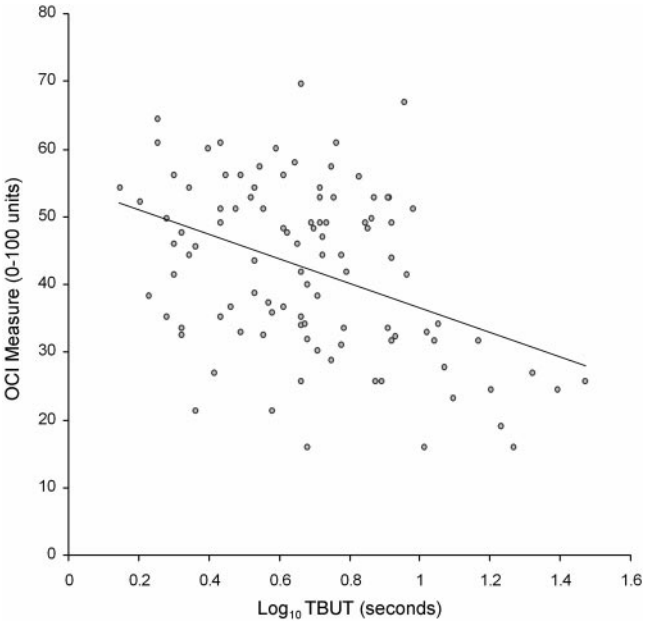

FIGURE 6. A scatterplot of OCI measures versus the logarithm of the median TBUT. The fitted linear regression illustrates the inverse relationship between these variables.

Summary

The OCI produces valid measures of ocular surface irritation when scored with maximum-likelihood iterative procedures. Good results can be achieved by using Rasch software with item difficulties and category structure anchored to the values reported in this paper, or with a computer program written in commercial software (Excel; Microsoft; Redmond, WA) available freely from the corresponding author (OCI Calculator, online at http://www.iovs.org/cgi/content/full/48/10/4451/DC1), as are copies of the questionnaire (OCI Questionnaire, http://www.iovs.org/cgi/content/full/48/10/4451/DC1).

The OCI is suitable for use in clinical trials to assess the impact of ocular surface disease on patients' well-being and the effectiveness of therapeutic strategies. Its major benefits over existing instruments are that, through Rasch analysis, it produces estimates on a linear interval scale rather than ordinal ranks and so is better able to quantify change and, through statistical methods, to account more satisfactorily for missing data. However, the clinical significance of its units requires empiric determination and, as with all questionnaires that employ Rasch methods, it struggles to deal with extreme raw scores.

Acknowledgments

The authors thank Helen Court for helpful discussions regarding Rasch analysis and Mark Gosnold for assistance in data collection.

References

1. McMonnies CW. Key questions in a dry eye history. *J Am Optom Assoc.* 1986;57:512–517.
2. McMonnies C, Ho A, Wakefield D. Optimum dry eye classification using questionnaire responses. *Adv Exp Med Biol.* 1998;438:835–838.
3. Nichols KK, Nichols JJ, Mitchell GL. The reliability and validity of McMonnies Dry Eye Index. *Cornea.* 2004;23:365–371.
4. Schiffman RM, Christianson MD, Jacobsen G, Hirsch JD, Reis BL. Reliability and validity of the Ocular Surface Disease Index. *Arch Ophthalmol.* 2000;118:615–621.

5. Stevenson D, Tauber J, Reis BL. Efficacy and safety of cyclosporine A ophthalmic emulsion in the treatment of moderate-to-severe dry eye disease: a dose-ranging randomized trial. The Cyclosporin A Phase 2 Study Group. *Ophthalmology*. 2000;107:967–974.
6. Sall K, Stevenson OD, Mundorf TK, Reis BL. Two multicenter, randomized studies of the efficacy and safety of cyclosporine ophthalmic emulsion in moderate to severe dry eye disease. CsA Phase 3 Study Group. *Ophthalmology*. 2000;107:631–639.
7. Michell J. Quantitative science and the definition of measurement in Psychology. *Br J Psychol*. 1997;88:355–383.
8. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989; 70:857–860.
9. Altman DG. *Practical Statistics for Medical Research.* London: Chapman and Hall/CRC; 1999.
10. Krantz DH, Luce RD, Suppes P, Tversky A. *Foundations of Measurement.* Vol 1. New York: Academic Press; 1971.
11. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut; 1960.
12. Massof RW. A systems model for low vision rehabilitation. II. Measurement of vision disabilities. *Optom Vis Sci*. 1998;75:349–373.
13. Massof RW. The measurement of vision disability. *Optom Vis Sci*. 2002;7:516–552.
14. Michell J. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.
15. Cox EP. The optimal number of response alternatives for a scale: a review. *J Market Res*. 1980;17:407–422.
16. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev*. 1956;63:81–97.
17. Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas*. 2002;3:85–106.
18. Wright BD, Masters GN. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press; 1982.
19. Smith RM. The distributional properties of Rasch item fit statistics. *Educ Psychol Measurement*. 1991;51:541–565.
20. Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. *J Outcome Meas*. 1998;2:66–78.
21. Johnson ME, Murphy PJ. The effect of instilled fluorescein solution volume on the values and repeatability of TBUT measurements. *Cornea*. 2005;24:811–817.
22. Bron AJ, Evans VE, Smith JA. Grading of corneal and conjunctival staining in the context of other dry eye tests. *Cornea*. 2003;22: 640–650.
23. Lemp MA. Report of the National Eye Institute/Industry workshop on clinical trials in dry eyes. *CLAO J*. 1995;21:221–232.
24. Nichols KK, Smith JA. Association of clinical diagnostic tests and dry eye surveys: the NEI-VFQ-25 and the OSDI. *Adv Exp Med Biol*. 2002;506:1177–1181.
25. Begley CG, Chalmers RL, Abetz L, et al. The relationship between habitual patient-reported symptoms and clinical signs among patients with dry eye of varying severity. *Invest Ophthalmol Vis Sci*. 2003;44:4753–4761.
26. Brignole F, Pisella PJ, Dupas B, Baeyens V, Baudouin C. Efficacy and safety of 0.18% sodium hyaluronate in patients with moderate dry eye syndrome and superficial keratitis. *Graefes Arch Clin Exp Ophthalmol*. 2005;243:531–538.
27. Linacre JM. *WINSTEPS Rasch Measurement Computer Program.* Chicago: Winsteps.com; 2005.
28. Andrich D. A rating scale formulation for ordered response categories. *Psychometrika*. 1978;43:561–573.
29. Kline P. *An Easy Guide to Factor Analysis*. London: Routledge; 1994.
30. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135–160.
31. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
32. Pesudovs K, Garamendi E, Keeves JP, Elliot DB. The Activities of Daily Vision Scale for cataract surgery outcomes: re-evaluating validity with Rasch analysis. *Invest Ophthalmol Vis Sci*. 2003;44: 2892–2899.
33. Wright BD, Linacre JM. Reasonable mean-square fit values. *Rasch Meas Trans*. 1994;8:370.
34. Begley CG, Chalmers RL, Mitchell GL, et al. Characterization of ocular surface symptoms from optometric practices in North America. *Cornea*. 2001;20:610–618.
35. Toda I, Fujishima H, Tsubota K. Ocular fatigue is the major symptom of dry eye. *Acta Ophthalmol* (*Copenh*). 1993;71:347–352.
36. Nichols KK, Mitchell GL, Zadnik K. Performance and repeatability of the NEI-VFQ-25 in patients with dry eye. *Cornea*. 2002;21:578–583.
37. Rimal RB, Real K. Assessing the perceived importance of skin cancer: how question-order effects are influenced by issue involvement. *Health Educ Behav*. 2005;32:398–342.
38. Wright BD. Estimating measures for extreme scores. *Rasch Meas Trans*. 1998;12:632–633.
39. Linacre JM. Estimating measures with known polytomous item difficulties. *Rasch Measure Trans*. 1998;12:638.
40. de Vet HC, Terwee CB, Ostelo RW, Beckerman H Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes*. 2006;4:54.