Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2009;61:37–48
DOI: 10.1159/000200767

Published online: February 10, 2009

Item Reduction of the Voice Handicap Index Based on the Original Version and on European Translations

T. Nawka[a] I.M. Verdonck-de Leeuw[b] M. De Bodt[c] I. Guimaraes[d] E.B. Holmberg[e] C.A. Rosen[f] A. Schindler[g] V. Woisard[h] R. Whurr[i] U. Konerding[j]

[a]Department of Phoniatrics and Paedaudiology, University of Greifswald, Greifswald, Germany; [b]Department of Otolaryngology-Head and Neck Surgery, VU University Medical Center, Amsterdam, The Netherlands; [c]Department of Otolaryngology and Communication Disorders, University of Antwerp, Edegem, Belgium; [d]Speech Therapy Department, Escola Superior de Saude do Alcoitao, Lisbon, Portugal; [e]Karolinska Institute, Department of Logopedics and Phoniatrics, Karolinska University Hospital, Huddinge, Sweden; [f]Department of Otolaryngology-Head and Neck Surgery, University of Pittsburgh Voice Center, Pittsburgh, Pa., USA; [g]Department of Clinical Sciences 'L. Sacco', University of Milan, Milan, Italy; [h]Department of Otorhinolaryngology, Head and Neck Surgery, Larrey University Hospital of Toulouse, Toulouse, France; [i]Department of Neuro-otology, National Hospital for Neurology and Neurosurgery, London, UK; [j]Department of Community Medicine, University of Greifswald, Greifswald, Germany

Key Words

Voice Handicap Index · Factor analysis · Item reduction · International short-scale

Abstract

Objective: Constructing an internationally applicable short-scale of the Voice Handicap Index (VHI). ***Methods:*** Subjects were 1,052 patients with 5 different types of voice disorder groups from Belgium, France, Sweden, Germany, Italy, The Netherlands, Portugal, and the USA. Different 9- and 12-item subsets were selected from the 30 VHI items using (1) the first factor of an unrotated factor analysis (narrow range subsets) and (2) the first three factors after promax rotation (broad range subsets). Country-specific subsets were selected to test deviations from the international subsets. For all subsets, reliability was investigated using Cronbach's alphas and correlations with the total VHI. Validity was investigated using regression on voice disorder groups. All analyses were performed for the total and for all country-specific subject samples. ***Results:*** Reliability was high for all item subsets. It was lower for the international compared to the country-specific subsets and for the broad range compared to the narrow range subsets. Validity was best for the broad range subsets. Validity was better for the international than for the country-specific subsets. For all statistics the 12-item subsets were not essentially better than the 9-item subsets. ***Conclusion:*** The international broad range 9-item subset forms a scale which approximates well the total VHI.

Copyright © 2009 S. Karger AG, Basel

The study was presented at the IALP Voice Committee Symposium during the 27th World Congress of the IALP in Copenhagen, Denmark, in 2007. It reports on the results of a European-wide project that was initiated at the PEVOC5 Conference in Graz; Guest Editor: Jan Svec, Groningen.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
1021–7762/09/0611–0037$26.00/0

Accessible online at:
www.karger.com/fpl

Univ.-Prof. Dr. med. Tadeus Nawka
ENT, Phoniatrics and Paedaudiology
Walther-Rathenau-Strasse 43–45, DE–17475 Greifswald (Germany)
Tel. +49 3834 866 215, Fax +49 3834 866 201
E-Mail nawka@uni-greifswald.de

Introduction

Systematic evaluations of voice-related medical interventions benefit from using a validated and meaningful questionnaire on voice problems [1]. This questionnaire should be part of a multidimensional voice analysis protocol, as proposed by the European Laryngological Society [2]. Currently, the Voice Handicap Index (VHI) is most widely used for this purpose [3]. The VHI consists of 30 items with 5 response levels, scored 0–4. Accordingly, the VHI sum score ranges from 0 to 120. A higher score corresponds to a more severe subjectively experienced vocal handicap. The total VHI can be divided into 3 subscales with 10 items each: the Physical (P-items), Functional (F-items), and Emotional (E-items) subscale. The physical subscale addresses the patient's perceptions during sound production of the voice, the functional subscale addresses the ability to communicate in various settings, and the emotional subscale addresses emotional aspects of the voice problem.

The VHI was developed and first validated in the USA. It has since been translated and validated in several languages, including Dutch, Flemish Dutch, French, German, Italian, Portuguese, and Swedish. The standard procedures for the translation process for the non-US-English language versions include a number of parallel translations, review of the translations leading to a consensus version, backward translations, comparison with the original American version and pilot testing of the translation. Instead of literal translation, some European versions were adapted to a culturally relevant form [4]. The widespread use of the VHI and its translations into European languages has made it familiar to many clinicians [e.g. 5–11]. In a previous study [4], the validity of translations of the VHI was demonstrated for a dataset of the European workgroup on VHI. Using confirmatory factor analysis, the study revealed that an oblique three-factor measurement model best fits the data, i.e. the total item variance of the VHI can best be explained by three latent dimensions which are correlated with each other.

One problem which may arise with the use of the VHI is due to its length. In routine diagnostics, voice patients may need to undergo several further measurements. Therefore, the 30 items of the VHI might require too much time (about 10–15 min). For this reason, two shortened versions of the VHI have been proposed: the VHI-10 [12] and the VHI-12 [13] (see Appendix 1). The VHI-10 has been constructed by selecting those items which discriminate best between patients and a control group as well as between pre- and post-treatment [12]. The VHI-12 is based on factor analysis [13] with test-retest validation [14]. Empirical studies concerning the psychometric properties of the VHI-10 rendered satisfying results [15, 16]. A study concerning the VHI-12 revealed that the sum scores of this instrument can be transformed into the sum scores of the total VHI (VHI-30) by multiplication with 2.5 [17]. Both scales have already been applied in clinical studies [17–20].

The studies just cited certainly provide important insights on a national level. However, each of these short-scales is only constructed on the basis of data from one single country and one specific language version. To be specific, the VHI-10 has been developed using a subject sample from the United States and the American English version, the VHI-12 using a subject sample from Germany and the German version of the VHI. Due to specific cultural terms and conditions, in these countries different items of the original VHI may have been selected for a shortened version. Therefore, it is questionable whether the results of the studies performed with the VHI-10 and the VHI-12 can be generalized to other countries. In order to guarantee international comparability, a shortened version of the VHI is desirable which has been shown to form a reliable and valid measurement instrument in all countries in which it is to be applied. In this study, such a shortened international version of the VHI is constructed using data from Belgium, France, Sweden, Germany, Italy, The Netherlands, Portugal, and the USA.

Methods

Patients

The data analyzed here are the same which have already been applied by Verdonck-de Leeuw et al. [4]. These data stem from the 8 countries mentioned above and belong to 1,052 patients (table 1) who sought therapeutic advice because of voice complaints. The median age was 45 years (range 12–86); there were 360 males (34%) and 692 (66%) females. Patients were classified as suffering from (1) vocal dysfunction without organic vocal fold changes, (2) vocal fold nodules, (3) structural lesions of the epithelium and lamina propria, (4) unilateral paresis of the vocal fold, or (5) laryngitis. The distribution of subjects in the diagnostic categories differs among countries.

Procedure

VHI questionnaires were collected from patients before logopedic, surgical or medical voice treatment. The de-identified data were sent to the central database manager. Only age, gender, and voice lesion category were provided together with the scores on all VHI items.

Table 1. Patient cohort (n = 1,052)

Country	Voice disorder					n	Male	Age
	dysfunction	nodules	structural	paresis	laryngitis		%	median (range)
Total	234	131	448	137	102	1,052	34	45 (12–86)
Belgium	39	16	81	4	8	169	47	45 (15–73)
France	–	58	–	–	–	58	31	40 (15–72)
Germany	90	15	121	31	18	275	29	45 (12–76)
Italy	26	2	34	13	–	75	25	43 (17–78)
Netherlands	–	–	99	43	55	197	42	46 (13–85)
Portugal	30	27	52	4	8	121	29	43 (18–70)
Sweden	36	7	33	10	2	88	28	49 (29–81)
USA	13	6	28	11	11	69	30	45 (15–86)

Statistical Analyses

The statistical analyses presented in the following aim at determining a subset of the VHI items which consists of about only 10 items and which constitutes a reliable and valid measurement instrument of the patient-based perception of vocal handicap, although it is much shorter than the original instrument. To explicate the purpose of the different analyses, some general remarks concerning the conceptions of reliability and validity as well as the general logic of the approach applied here are required. Therefore, the description of the statistical analyses is divided into three parts. The first part is concerned with the concepts of reliability and validity, the second with the general logic of the analyses presented here, and the third with the specific statistical operations.

Concepts of Reliability and Validity

Reliability

Reliability means that the measurement instrument produces the same value measured under the same conditions. There are three different approaches to investigate reliability: the first consists in comparing the values produced by the same measurement instrument applied to the same persons at two different points of measurement (test-retest reliability); the second consists in comparing the values produced by two different measurement instruments which are supposed to measure the same construct and which have been applied to the same persons at the same time (parallel-test reliability); the third consists in comparing the values produced by different parts of the same measurement applied to the same persons at the same time. There are at least two versions of the latter approach: one consists in investigating the relations between two halves of the measurement instrument (split-half reliability), the other in investigating the relations between all single items (internal consistency) [21]. The analyses presented here focus on internal consistency and, in a somewhat broader sense, on parallel-test reliability.

Validity

Validity means that the measurement instrument measures what is intended to be measured. There are two sources of information which can be used for checking the validity: the first source is based on the features of the single items constituting the measurement instrument and, especially, on the principles for selecting these items; the second source consists of the instrument's relations to external variables. Validity based on the first source is usually referred to as content validity, validity based on the second source as criterion validity. Criterion validity can be further distinguished into concurrent and predictive validity. In concurrent validity, the criterion is assessed at the same time as the measurement instrument in question; in predictive validity, the criterion is assessed later [22]. In the literature, still the concept of construct validity is applied. The problem of establishing construct validity arises when the quantity to be measured is not assumed to reveal in one single empirical phenomenon but, with a smaller or larger probability, in very different empirical phenomena. Construct validity is given when the test relates to these different empirical phenomena in the theoretically predicted manner. Investigating construct validity requires the integration of content validity and several aspects of criterion validity [22]. The analyses presented here focus on content and on concurrent criterion validity.

Basic Logic of Statistical Analyses

Reliability in the sense of internal consistency and content validity cannot usually be both maximized at the same time. Maximal internal consistency is achieved when items are selected which are as similar as possible. In contrast, content validity requires that the whole realm of phenomena is covered in which the construct, i.e., in this case, the subjectively experienced voice handicap, manifests. Therefore, two statistical approaches were applied here parallel to each other: one to achieve maximal reliability and the other to achieve maximal validity.

The statistical approach which aims at maximizing content validity is based upon the idea that the VHI consists of three dimensions as described in the Introduction. An optimal representation of these three dimensions requires the same number of items to be selected for each dimension. This, in turn, implies that the number of selected items must be dividable by 3. The numbers dividable by 3 which are nearest to the originally envisaged number of 10 are 9 and 12. Merely theoretical considerations do not imply a clear decision for 9 or 12 items. A 9-item scale is shorter and thereby more economical than a 12-item scale. A 12-item scale, however, will most probably have better psychometric char-

acteristics. Therefore, accuracy and economy must be weighed against each other. To provide the basis for this weighing, 9- and 12-item subsets were alternatively tried as possible scales.

Different subsets might be optimal for different countries. This implies that applying one and the same subset for all countries might be associated with a loss of accuracy in comparison with the country-specific subsets. To investigate this loss of accuracy, statistical analyses both for the total subject sample and for the subject samples of each country were performed.

Specific Statistical Operations

There are 14,307,150 different ways of selecting 9 out of 30 items and even 86,493,225 ways of selecting 12 out of 30 items. Therefore, the specific statistical operations presented in the following are separated into two parts. The first part aims at selecting those subsets of items which promise most to constitute a reliable and valid short-scale; the second part aims at investigating the reliability and validity of the selected subsets.

Selection of Subsets

To select item subsets which are likely to form a highly reliable scale, principal component factor analyses with one fixed factor were performed. A factor analysis of this kind yields an unobservable variable which correlates maximally with the different items contained in the whole questionnaire and which represents the phenomena that are mainly assessed by this questionnaire. The correlations between this variable and the single items (in more technical terms: the item loadings on the first factor) were applied for item selection. The 9 items with the highest loadings were selected as candidate for a 9-item scale. The 12-item subset was constructed by adding the 3 items with the next highest loadings. To select item subsets for subjects from all countries (international subsets), this approach was performed for the total subject sample. To construct country-specific item subsets the whole approach was performed for the subject samples of each country separately. The item subsets produced by a factor analysis with one fixed factor can be expected to cover a rather narrow range of the phenomena described by the original VHI-30 items. Therefore, in the following, these subsets are referred to as narrow range subsets.

To select item subsets which are likely to form a highly valid scale, a principal component factor analysis with three factors and subsequent promax rotation was performed. A three-factor analysis with promax rotation aims at finding three unobservable variables (i.e. the factors) by which the analyzed items are predicted as accurately as possible via linear regression equations. The specific feature of promax rotation consists in rotating the three factors in such a way that they fulfill two conditions: (1) each item correlates as much as possible with one of the factors and as little as possible with the remaining two factors and (2) the three factors are allowed to correlate with each other, i.e. oblique angles between the factors are permitted. A three-factor model was chosen because the original VHI was meant to be composed of three dimensions. Moreover, Verdonck-de Leeuw et al. [4] found that a rotated three-factor solution with oblique angles fits best to data. The fact that Verdonck-de Leeuw et al. found oblique angles was the reason for choosing a rotation which allows for oblique angles in this study, too.

The 9-item subsets were formed by selecting for each of the 3 factors the 3 items with the highest correlations with the respective factor. Correspondingly, the 12-item subsets were formed by selecting for each factor the 4 items with the highest correlations. The same number of items was chosen from each factor in order to guarantee that the resulting scale represents these three factors to the same extent. Just like in the analyses for the narrow range subsets, the factor analyses with three factors were conducted both for the total subject sample and for the subject samples of each country separately. Again, international item subsets were constructed using the data of the total subject sample, whereas country-specific item subsets were constructed using only the data of the subject samples from the respective country. The subsets produced by the three-factor analyses can be expected to cover a quite broad range of the phenomena described by the original VHI-30 items. Therefore, in the following, these subsets are referred to as broad range subsets.

Investigation of Reliability and Validity

For all selected subsets internal consistency was determined using Cronbach's alpha [23]. For all subsets these analyses were performed with the country-specific subject samples. For the original VHI-30 and the 4 different international item subsets, i.e. the narrow range 9-item, the narrow range 12-item, the broad range 9-item and the broad range 12-item subset, these analyses were also performed for the total subject sample. In addition to internal consistency, also a kind of parallel-test reliability was examined. The VHI-30 was considered as a parallel version of the scales formed by these subsets. For all subsets the correlations of the sum score with the VHI-30 were computed with the country-specific subject samples. For the 4 international item subsets these correlations were also determined for the total subject sample.

To check concurrent criterion validity, the relations of the different subsets' sum scores to the different categories of voice disorder grouping were investigated. Of course, the same type of laryngeal pathology can affect people in different ways. However, systematic differences in the central tendencies of index values in different diagnostic categories can be expected [12]. Multivariate linear regression analyses were computed to investigate these relations. The dummy-coded lesion categories were used as independent and the different sum scores as dependent variables. The multiple correlations resulting from these analyses reflect the extent to which the voice disorder categories influence the subjectively experienced voice handicap assessed by the respective sum score. As far as the construct of subjectively experienced voice handicap can be assumed to be related to the objective disease, these correlations are indicators of validity. For the original VHI-30 as well as for the 4 different international item subsets, these analyses were performed both for the total subject sample and, except for France, for the subject samples of each country separately. Except for France, the analyses were also conducted for each country-specific subset using only the data of the corresponding country-specific subject sample. For France, no country-specific analyses were possible because all subjects of the French sample belonged to the same lesion category of vocal fold nodules.

To illustrate the relation between the voice disorder categories and the sum scores, also the scores' means and standard deviations were computed for each category. To make sum scores of the 9-item subsets, the 12-item subsets and the VHI-30 comparable all scores were normalized to a scale range of 0–100.

All analyses were performed with SPSS 15.0 (SPSS Inc., Chicago, Ill., USA).

Table 2. Items[a] for subsets[b]

Narrow range[c]	
International	F16**, E25*, P20, E7, P14*, E24**, F11, F5**, F1**, (E29*, F12*, E28)
Belgium	E25**, E7, E23**, E24**, F5**, P20**, F6*, F16, F8**, (F12, P10, F11*)
France	F11**, F8**, F6*, F3, P20**, E28**, E27, P14**, F16, (E7*, P10*, F1)
Germany	P14**, F19**, E25, F8**, P20**, E7, F11**, F12, E9, (F16, F5*, F1)
Italy	P20**, E7, F8**, P17, F3**, F11, P14**, F12, E27, (E25*, F19, F16)
Netherlands	E25**, F16**, E28, E24*, E29, F11, E30**, E27, E9, (F19, E7, F5*)
Portugal	P20**, E7, F16, F19**, E15, E23**, P14, F1, P26, (P10, F8*, F11*)
Sweden	F19**, F16*, E24, F12, F11**, P20**, F8**, P14**, E25, (E7, E15*, F1*)
USA	E9**, E27, P21, E28**, F6**, E7**, P18, F8, P2, (F1*, E25*, F3)

	Factor 1	Factor 2	Factor 3
Broad range			
International	E24, F16, E29, (E25)	P4, P17, P21, (P14)	F1, F3, F5, (F12)
Belgium	P20, P14, F5, (F1)	F19, F11, F8, (E28)	E25, E23, E24, (F6)
France	F8, F12, F11, (F6)	P20, P14, P10, (P2)	E24, E30, E28, (E7)
Germany	P14, P20, P17, (P4)	F11, F8, F19, (F5)	E30, E28, E29, (E27)
Italy	P20, P14, P18, (P26)	F8, F3, F5, (F1)	F16, E25, E29, (F22)
Netherlands	E25, F16, E30, (E24)	F1, F5, F12, (F3)	P4, P20, P14, (P21)
Portugal	P20, E23, P17, (P4)	E30, E29, E28, (E25)	F8, F11, F19, (E24)
Sweden	P14, P20, P13, (F1)	F8, F19, E28, (F16)	F11, F6, E15, (P2)
USA	E7, E9, F6, (F1)	P26, E28, E29, (E25)	F16, E15, P14, (F12)

[a] The items are numbered like in the original publication [3].
[b] Additional items for the 12-item subsets in parentheses.
[c] Items which are also contained in the corresponding broad range scale are marked with one asterisk for the 12-item and with two asterisks for the 9-item scales.

Results

Identification of Subsets

The different factor analyses yielded different results and, therefore, item selections (table 2, and see Appendix 2). For the total subject sample the one-factor analysis explained 39.5% of the variance of the VHI, the three-factor analysis 51.4%.

Investigation of Reliability and Validity

Reliability

Internal Consistency

In the total subject sample, Cronbach's alpha of the VHI-30 is 0.95; Cronbach's alphas of the international subsets range from 0.85 to 0.91 (table 3). In the country-specific subject samples, Cronbach's alphas range from 0.91 to 0.97 for the VHI-30 (table 3, column 2) and from 0.74 to 0.97 for the different item subsets (table 3, columns 3–10). The minimum is the value for the international broad range 9-item subset in France, the maximum the value for the country-specific narrow range 12-item short subset in the USA. Except for both 12-item subsets in the German subject sample, the country-specific subset always yielded a higher Cronbach's alpha than the corresponding international subset. In a similar way, Cronbach's alphas for the 9-item subsets are always smaller than Cronbach's alphas of the corresponding 12-item subset. Except for the subjects from USA, each narrow range subset has a higher Cronbach's alpha than the corresponding broad range subset. For the subjects from the USA, Cronbach's alpha of the broad range subset is higher.

Parallel-Test Reliability

In the total subject sample, the correlations of the international item subsets with the VHI-30 range from 0.94 to 0.96; in the country-specific subject samples, the correlations of the international and the country-specific item subsets range from 0.90 to 0.99 (table 4). The mini-

Table 3. Cronbach's alphas for the total VHI-30 and the different short scales[a]

	Cronbach's alpha							
	VHI-30	narrow range subsets				broad range subsets		
		country-specific		international		country-specific		international	
		12 items	9 items	12 items	9 items	12 items	9 items	12 items	9 items
Total sample									
International	0.95	–	–	0.91	0.89	–	–	0.89	0.85
Country-specific subsamples									
Belgium	0.95	0.93	0.91	0.91	0.90	0.91	0.89	0.89	0.85
France	0.91	0.90	0.88	0.86	0.83	0.87	0.84	0.81	0.74
Germany	0.96	0.92	0.90	0.91	0.90	0.90	0.88	0.91	0.87
Italy	0.93	0.90	0.89	0.88	0.86	0.87	0.85	0.85	0.80
Netherlands	0.92	0.91	0.90	0.90	0.87	0.87	0.84	0.85	0.78
Portugal	0.95	0.92	0.91	0.91	0.89	0.90	0.87	0.89	0.86
Sweden	0.95	0.94	0.93	0.93	0.92	0.92	0.90	0.91	0.87
USA	0.97	0.97	0.96	0.92	0.90	0.94	0.93	0.92	0.91

[a] For sample sizes see table 1.

Table 4. Sum score correlations of subsets with VHI-30[a]

	Narrow range subsets				Broad range subsets			
	country-specific		international		country-specific		international	
	12 items	9 items	12 items	9 items	12 items	9 items	12 items	9 items
Total sample								
International	–	–	0.96	0.95	–	–	0.96	0.94
Country-specific subsamples								
Belgium	0.97	0.95	0.97	0.96	0.97	0.96	0.97	0.95
France	0.95	0.93	0.94	0.94	0.95	0.94	0.94	0.93
Germany	0.97	0.96	0.97	0.96	0.97	0.96	0.97	0.96
Italy	0.97	0.96	0.96	0.94	0.96	0.95	0.94	0.90
Netherlands	0.93	0.91	0.95	0.93	0.94	0.93	0.94	0.92
Portugal	0.98	0.96	0.97	0.96	0.96	0.95	0.98	0.96
Sweden	0.98	0.96	0.98	0.97	0.98	0.97	0.97	0.96
USA	0.98	0.97	0.98	0.97	0.99	0.98	0.98	0.98

[a] For sample sizes see table 1. All correlations differ significantly from zero with $p < 0.01$.

mum is for the international broad range 9-item subset in the Italian subject sample, the maximum for the country-specific broad range 12-item subset in the US subject sample. In all cases, the correlation for the 12-item subset is at least as high as the correlation for the corresponding 9-item subset. In most cases, the correlation for the country-specific subset is at least as high as the correlation for the corresponding international subset. Exceptions are all narrow range subsets in the Dutch subject sample, the narrow range 9-item subsets in the French, the Belgian and the Swedish subject sample and all broad range subsets in the Portuguese subject sample. There is no clear ordinal relation between the narrow range and the broad range subsets.

Table 5. Values[a] for voice disorder groups as obtained by total VHI and international subsets

	Dysfunction	Nodules	Laryngitis	Structural	Paresis
Total VHI	32.9 ± 22.32 26.3	35.6 ± 18.40 29.8	39.1 ± 22.52 32.3	40.1 ± 23.01 32.7	48.7 ± 24.49 39.6
12-item broad range subset	15.4 ± 9.64 31.0	16.6 ± 7.81 34.7	18.2 ± 10.37 38.1	18.5 ± 10.18 38.0	23.4 ± 10.55 48.0
9-item broad range subset	11.5 ± 7.31 31.0	12.6 ± 5.93 35.2	14.0 ± 8.18 39.7	14.0 ± 7.63 38.6	17.6 ± 7.80 48.3
12-item narrow range subset	12.4 ± 9.92 24.2	12.8 ± 8.14 26.4	14.3 ± 10.49 28.5	15.0 ± 10.43 30.0	20.0 ± 11.17 40.4
9-item narrow range	10.3 ± 7.85 27.4	10.8 ± 6.67 29.9	11.6 ± 8.15 31.0	12.4 ± 8.27 33.2	16.5 ± 8.85 44.8

[a] First line: mean score values and standard deviations, second line: mean score values normalized to a scale range from 0 to 100.

Validity

In the total subject sample the VHI as well as the sum scores of all 4 international item subsets differ for the five voice disorder categories (table 5). The basic pattern of the mean values is the same for the VHI and the four sum scores. The highest mean values are found in the category of paresis, followed by those in structural problems, laryngitis, and nodules. Correspondingly, the lowest values are found for dysfunction (fig. 1).

In the total subject sample, the adjusted multiple R^2 is lowest for the VHI-30 and highest for both broad range subsets (table 6). This suggests that the sum scores of the two broad range subsets tend to be most sensitive to the effects of the voice disorders categories and the total VHI least sensitive. All five R^2 differ statistically significant from zero. In the analyses for the country-specific subject samples, the association between the sum scores and the voice disorders categories is statistically significant only for Belgium and Germany. For these two countries the pattern of R^2 differs from the pattern for the total sample. The R^2 for the VHI-30 are not smaller than all R^2 of the other international subsets. Altogether, the analyses of the country-specific subject samples render no consistent evidence for or against one of the subsets considered here.

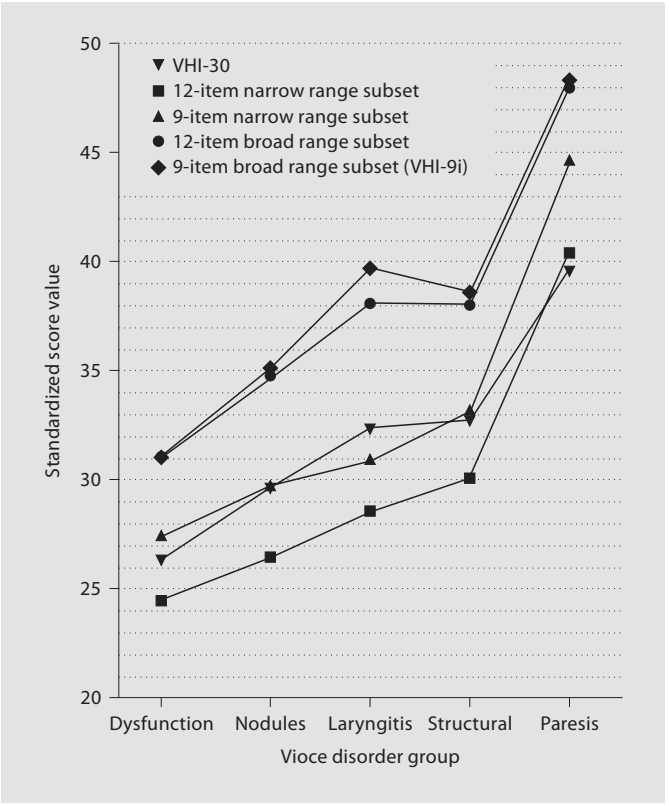

Fig. 1. Mean standardized score values for VHI, 9- and 12-item narrow and broad range international subsets.

Discussion

The goal of the present study was to construct a shortened international version of the VHI with optimal reliability and validity. For this purpose a two-step proce-

Table 6. Adjusted R^2 for regression of total VHI and subscales on voice disorder categories[a]

	VHI	Narrow range subsets				Broad range subsets			
		country-specific		international		country-specific		international	
		12 items	9 items	12 items	9 items	12 items	9 items	12 items	9 items
Total sample									
International	0.039***	–	–	0.046***	0.048***	–	–	0.052***	0.052***
Country-specific subsamples									
Belgium	0.07**	0.09**	0.08**	0.08**	0.10***	0.09**	0.10***	0.06**	0.05*
Germany	0.05**	0.06***	0.05**	0.05**	0.06**	0.04**	0.03**	0.04**	0.04**
Italy	0.00	0.02	0.00	0.04	0.05	0.04	0.03	0.07	0.09*
Netherlands	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.00
Portugal	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
Sweden	0.04	0.01	0.00	0.02	0.02	0.03	0.03	0.05	0.06
USA	0.01	0.01	0.00	0.02	0.02	0.03	0.02	0.04	0.04

[a] For lesion categories and sample sizes see table 1. Statistically significant deviation from zero is marked with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. For France, no adjusted R^2 could be determined because all French subjects belonged to the same lesion category.

dure was applied. Firstly, item subsets were selected using factor analyses from the VHI as candidates for the final short-scale. These item subsets were a 9- and a 12-item narrow range subset as well as a 9- and a 12-item broad range subset for the total subject sample as well as for each of the country-specific samples. Secondly, reliability and validity of the selected subsets were investigated. Reliability was examined by means of internal consistency and correlation with the total VHI, validity by means of multiple regression on lesion categories.

The analyses refer to data from seven Western European countries and from the USA. Correspondingly, the scope of the study is restricted. To produce a more valid international scale, data from other countries and in other languages should be included. Subjects from 12 to 86 years have been included. One might suspect that the youngest subjects might not respond to the VHI in a reliable way. Yet, presently, there is no age-specific version of the VHI. Hence, it is common practice to apply it to the age range considered here. There is a further restriction of the method applied here: it is not guaranteed that patients would respond in the same way to the 9 selected items when they are presented alone, as they respond to them when the items are presented as part of the total VHI. Whether they do so should be subject to further research.

The subsets selected for the different countries were quite diverse. There are at least 3 different explanations for these differences: (1) The differences might be an effect of statistical error produced by the sampling process, i.e., the true factor structures in the whole populations of the eight different countries might be the same, but the samples investigated here deviate from this true factor structure merely by chance. (2) The differences between the countries might be caused by the different distributions of voice disorder types in the country-specific subject samples. (3) The differences might be caused by the fact that subjects from different countries understand and answer the questions of the VHI in a different way. Yet, whatever the reason for the differences is, they are not important as long as the psychometric properties of the international subsets are not essentially worse than the psychometric properties of the different country-specific subsets.

As expected, the internal consistencies decrease with decreasing number of items, are lower in the broad range than in the narrow range subsets, and are lower in the international than in the country-specific subsets. The first result is trivial because Cronbach's alpha usually increases with number of items. The second result is also not surprising because, in contrast to the broad range subsets, the narrow range subsets were produced by selecting the items with the highest item-total correlation. This selection procedure also implies that country-specific narrow range subsets must have higher internal consistencies than the corresponding international subsets.

In contrast, country-specific broad range subsets must not necessarily be superior to the international ones. The fact that this superiority does exist might be explained by the fact that broad and narrow range subsets have between 5 and 8 items in common (table 2). All subsets, even the broad range 9-item subsets, have considerably high internal consistencies. This is a consequence of the extremely high internal consistency of the total VHI.

For all subsets, the lowest internal consistencies are found for France, The Netherlands, and Italy (table 3). This holds true for the total VHI, as well as for both country-specific and both international subsets. For France and The Netherlands, this result can be explained by the fact that at least one of the two extreme diagnostic categories, i.e. dysfunction and paresis, are not represented (table 1). Removing an extreme diagnostic category reduces the systematic variance without affecting the error variance. Cronbach's alpha, however, is an estimation of the ratio between systematic variance and the sum of systematic and error variance. Therefore, removing an extreme category will diminish Cronbach's alpha.

Cronbach's alphas are slightly higher for the narrow than for the broad range subsets. For interpreting this effect, still a further characteristic of Cronbach's alpha must be taken into consideration. Cronbach's alpha treats differences in the content of items as error variance. For this reason, Cronbach's alpha must be lower for heterogeneous than for homogeneous measurement instruments. This, however, does not necessarily mean that also the test-retest reliability and the parallel-test reliability are lower for heterogeneous than for homogeneous instruments. On the contrary, these reliabilities can even be higher for the heterogeneous test [23]. This, in turn, means that the slightly higher Cronbach's alphas are no sufficient evidence for a superiority of the narrow range subsets.

The pattern of the correlations with the total VHI is similar to the pattern of the internal consistencies. The correlations are lower for the 9- than for the 12-item subsets and lower for the international than for the country-specific subsets. The first of these two results is trivial because the 12-item subsets have more items in common with the total VHI than the 9-item subsets. The second of these results is trivial as far as narrow range subsets are concerned. The country-specific narrow range subsets consist of those items with the highest item-total correlation within the respective country-specific subject sample. The finding that also the country-specific broad range subsets are better predicted than their international counterparts might again be explained by the fact that broad and narrow range subsets have many items in common. The sum scores of the narrow range subsets do not correlate distinctly better with the total VHI-30 than the broad range sum scores. This contrasts to the throughout better internal consistencies of the narrow range subsets. These results suggest that the higher internal consistencies of the narrow range subsets are actually a methodological artifact.

Just like the internal consistencies, the correlations of the sum scores with the total VHI are considerably high. This holds true even for the international 9-item subsets and although the different subsets consist of different items. These results can again be attributed to the extremely high internal consistency of the total VHI. Except for the country-specific narrow range 9-item subset, the lowest correlations are found for France, The Netherlands and Italy. At least for France and The Netherlands this can, just like the corresponding result for the internal consistencies, be explained by the lower variance in the subject samples of these two countries.

The results for validity as assessed by the adjusted multiple R^2 for the regression on voice disorder categories are remarkably different from the results concerning reliability. The adjusted multiple R^2 are by far closer to their lower than to their upper limit. Moreover, they do not differ significantly from zero for most of the country-specific analyses. These results are not surprising either. The same type of laryngeal pathology can affect people in different ways. The subjectively experienced voice handicap is not only determined by the type of voice disorder but is subject to several other influences. Nevertheless, different mean index values can be expected for different diagnostic categories [12]. Such differences have been found in the data presented here. In the total sample they are strongest for the broad range subsets and weakest for the total VHI. This suggests that the subsets are even more valid than the total VHI and that the validity of the two broad range subsets is largest. The analyses for the country-specific subject samples do not yield such a clear picture. At most, the results suggest that the kind of validity considered here is neither attenuated by removing the greater part of the items from the total VHI, nor by replacing the country-specific with the international subsets.

Altogether, applying one of the subsets is associated with only a small loss of reliability and no loss of validity compared with the total VHI. There is even no substantial loss when a 9-item subset is used instead of a 12-item subset. The same holds true if an international item subset is applied instead of a country-specific item subset. These results suggest that all important aspects of the

voice-related complaints, which are covered by the total VHI, are retained in the international 9-item subsets. The analyses presented here do not clearly indicate whether a broad or a narrow range subset is to be preferred. There are, however, some clues. On the one hand, the Cronbach's alphas are slightly higher for the narrow range subsets compared with the broad range subsets; on the other hand, for the total subject sample the adjusted multiple R^2 for the regression on the voice disorder categories are slightly higher for the broad range subsets. Considering the specific characteristics of Cronbach's alpha, the first finding is not sufficient for inferring a higher reliability of the narrow range subsets. There is, however, an, admittedly weak, evidence for a better validity of the broad range subsets. Moreover, there is a good reason to put a greater weight on validity than on reliability. Reliability reflects the accuracy with which an instrument measures whatever quantity is measured, whereas validity reflects the extent to which the instrument in fact measures what is supposed to be measured. All these considerations are arguments for a broad range subset.

The considerations just presented imply that the international 9-item broad range subset constitutes a scale which approximates the total VHI very well. For further use this scale is labeled 'VHI-9 international (VHI-9i)'. In contrast to the previously developed VHI-10 and VHI-12 (see Introduction), the VHI-9i is based on data from more than one country. Thus, if there was a study being planned that involved, for example, a German and a French group of patients, if there is not enough time for using the total VHI and if the subjectively experienced voice handicap is not the main target of the investigation, then the VHI-9i would be the best choice.

Appendix 1

VHI-10 and VHI-12

Item[a]	Text
VHI-10	
F1	My voice makes it difficult for people to hear me.
F3	People have difficulty understanding me in a noisy room.
P10	People ask, 'What's wrong with your voice?'
P14	I feel as though I have to strain to produce voice.
F16	My voice difficulties restrict my personal and social life.
P17	The clarity of my voice is unpredictable.
F19	I feel left out of conversation because of my voice.
F22	My voice problem causes me to lose income.
E23	My voice problem upsets me.
E25	My voice makes me feel handicapped.
VHI-12	
F1	My voice makes it difficult for people to hear me.
F3	People have difficulty understanding me in a noisy room.
F5	My family has difficulty hearing me, when I call them throughout the house.
F8	I tend to avoid groups of people because of my voice.
P14	I feel as though I have to strain to produce voice.
P17	The clarity of my voice is unpredictable.
F19	I feel left out of conversation because of my voice.
P21	My voice is worse in the evening.
E24	I am less outgoing because of my voice problem.
E27	I feel annoyed when people ask me to repeat.
E28	I feel embarrassed when people ask me to repeat.
E30	I'm ashamed of my voice problem.

[a] Refers to the item number of the VHI-30.

Appendix 2

Items of the International 9-Item Broad Range Short-Scale, VHI-9i

Item[a]	Text
English	
F1	My voice makes it difficult for people to hear me.
F3	People have difficulty understanding me in a noisy room.
P4	The sound of my voice varies throughout the day.
F5	My family has difficulty hearing me, when I call them throughout the house.
F16	My voice difficulties restrict my personal and social life.
P17	The clarity of my voice is unpredictable.
P21	My voice is worse in the evening.
E24	I am less outgoing because of my voice problem.
E29	My voice makes me feel incompetent.
Dutch	
F1	Door mijn stem kan ik mij moeilijker verstaanbaar maken.
F3	Mensen verstaan me moeilijk in een lawaaierige omgeving.
P4	De klank van mijn stem varieert in de loop van de dag.
F5	Mijn familieleden horen me moeilijk als ik ze roep ergens in huis.
F16	Mijn stemproblemen beperken mijn persoonlijk en sociaal leven.
P17	De helderheid van mijn stem is onvoorspelbaar.
P21	Mijn stem is 's avonds slechter.
E24	Ik ben minder spontaan door mijn stemprobleem.
E29	Door mijn stem voel ik mij onbekwaam.
French	
F1	On m'entend difficilement à cause de ma voix.
F3	On me comprend difficilement dans un milieu bruyant.
P4	Le son de ma voix varie au cours de la journée.
F5	Les membres de la famille ont du mal à m'entendre quand je les appelle dans la maison.

Item[a]	Text
F16	Mes difficultés de voix limitent ma vie personnelle et sociale.
P17	La clarté est imprévisible.
P21	Ma voix est plus mauvaise le soir.
E24	Je suis moins sociable à cause de mon problème de voix.
E29	A cause de ma voix je me sens incompétent(e).
German	
F1	Man hört mich wegen meiner Stimme schlecht.
F3	Anderen fällt es schwer, mich in einem lauten Raum zu verstehen.
P4	Der Klang meiner Stimme ändert sich im Laufe des Tages.
F5	Meine Familie hört mich kaum, wenn ich zuhause nach ihnen rufe.
F16	Meine Stimmschwierigkeiten schränken mich in meinem Privatleben ein.
P17	Bevor ich spreche, weiss ich nicht, wie klar meine Stimme klingen wird.
P21	Abends ist meine Stimme schlechter.
E24	Ich bin weniger kontaktfreudig wegen meines Stimmproblems.
E29	Wegen meiner Stimme fühle ich mich unfähig.
Italian	
F1	La mia voce è udita con difficoltà dalla gente.
F3	La gente ha difficoltà a capirmi in una stanza rumorosa.
P4	Il suono della mia voce varia durante la giornata.
F5	La mia famiglia ha difficoltà ad udirmi quando li chiamo in casa.
F16	Le mie difficoltà di voce restringono la mia vita personale e sociale.
P17	La chiarezza della mia voce è imprevedibile.
P21	La mia voce è peggiore la sera.
E24	Esco di meno per i miei problemi di voce.
E29	La mia voce mi fa sentire un incapace.
Portuguese	
F1	A minha voz faz com que seja difícil os outros ouvirem-me.
F3	As pessoas têm dificuldade em me compreender num local ruidoso.
P4	O som da minha voz varia ao longo do dia.
F5	A minha família tem dificuldade em me ouvir quando os chamo dentro de casa.
F16	As minhas dificuldades com a voz limitam a minha vida pessoal e social.
P17	A clareza da minha voz é imprevisível.
P21	Tento modificar a minha voz de modo a soar diferente.
E24	Saio menos por causa do meu problema de voz.
E29	A minha voz faz-me sentir incompetente.
Swedish	
F1	Min röst gör det svårt för människor att höra mig.
F3	Andra människor har svårigheter med att förstå mig i en miljö med mycket ljud.
P4	Min röstkvalitet varierar under dagen.
F5	Min familj har svårigheter med att höra mig när jag ropar på dem från en annan del av huset.
F16	Mina röstproblem begränsar mitt liv, bade personligt och socialt.
P17	Det händer att det inte går att förutsäga om min röst kommer att låta klar eller inte.
P21	Min röst är sämre på kvällen.
E24	Jag är minder utåtriktad på grund av mitt röstproblem.
E29	Min röst får mig att känna mig oduglig.

[a] Refers to the item number of the VHI-30.

References

1 Behrman A: Common practices of voice therapists in the evaluation of patients. J Voice 2005;19:454–469.

2 Dejonckere PH, Bradley P, Clemente P, et al: A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 2001;258:77–82.

3 Jacobson BH, Johnson A, Grywalski C, et al: The Voice Handicap Index (VHI): development and validation. Am J Speech Lang Pathol 1997;6:66–70.

4 Verdonck-de Leeuw IM, Kuik DJ, De Bodt M, et al: Validation of the Voice Handicap Index by assessing equivalence of European translations. Folia Phoniatr Logop 2008;60:173–178.

5 Bouwers F, Dikkers FG: A retrospective study concerning the psychosocial impact of voice disorders: Voice Handicap Index change in patients with benign voice disorders after treatment (measured with the Dutch version of the VHI). J Voice 2007, epub ahead of print.

6 Glas K, Hoppe U, Eysholdt U, Rosanowski F: Smoking, carcinophobia and Voice Handicap Index. Folia Phoniatr Logop 2008;60:195–198.

7 Helidoni ME, Murry T, Moschandreas J, et al: Cross-cultural adaptation and validation of the Voice Handicap Index into Greek. J Voice, doi:10.1016/j.jvoice.2008.06.005.

8 Schindler A, Bottero A, Capaccio P, et al: Vocal improvement after voice therapy in unilateral vocal fold paralysis. J Voice 2008;22:113–118.

9 Siupsinskiene N, Vaitkus S, Grebliauskaite M, Engelmanaite L, Sumskiene J: Quality of life and voice in patients treated for early laryngeal cancer. Medicina (Kaunas) 2008;44:288–295.

10 Woisard V, Bodin S, Yardeni E, Puech M: The Voice Handicap Index: correlation between subjective patient response and quantitative assessment of voice. J Voice 2007;21:623–631.

11 Lee CF, Carding PN, Fletcher M: The nature and severity of voice disorders in lung cancer patients. Logoped Phoniatr Vocol 2008;33:93–103.

12 Rosen CA, Lee AS, Osborne J, Zullo T, Murry T: Development and validation of the Voice Handicap Index-10. Laryngoscope 2004;114:1549–1556.
13 Nawka T, Wiesmann U, Gonnermann U: Validierung des Voice Handicap Index (VHI) in der deutschen Fassung. [Validation of the German version of the Voice Handicap Index]. HNO 2003;51:921–930.
14 Nawka T, Gonnermann U, Wiesmann U: Stimmstörungsindex. http://wwwegmsde/en/meetings/dgpp2003/03dgpp034shtml 2003.
15 Günther S, Rasch T, Klotz M, et al: Bestimmung der subjektiven Beeinträchtigung durch Dysphonien. Ein Methodenvergleich. HNO 2005;53:895–900, 902–894.
16 Deary IJ, Webb A, Mackenzie K, Wilson JA, Carding PN: Short, self-report voice symptom scales: psychometric characteristics of the Voice Handicap Index-10 and the vocal performance questionnaire. Otolaryngol Head Neck Surg 2004;131:232–235.
17 Gugatschka M, Rechenmacher J, Chibidziura J, Friedrich G: Vergleichbarkeit und Umrechnung von Stimmstörungsindex (SSI) und Voice Handicap Index (VHI). Laryngorhinootologie 2007;86:785–788.
18 Amir O, Tavor Y, Leibovitzh T, et al: Evaluating the validity of the Voice Handicap Index-10 (VHI-10) among Hebrew speakers. Otolaryngol Head Neck Surg 2006;135:603–607.
19 Lam PK, Chan KM, Ho WK, et al: Cross-cultural adaptation and validation of the Chinese Voice Handicap Index-10. Laryngoscope 2006;116:1192–1198.
20 Nunez Batalla F, Caminero Cueva MJ, Senaris Gonzalez B, et al: Voice quality after endoscopic laser surgery and radiotherapy for early glottic cancer: objective measurements emphasizing the Voice Handicap Index. Eur Arch Otorhinolaryngol 2008;265:543–548.
21 Allen M, Yen W: Introduction to Measurement Theory. Monterey/CA, Brooks/Cole Publishing, 1979.
22 Cronbach L, Meehl P: Construct validity in psychological tests. Psychol Bull 1955;52:281–302.
23 Cronbach L: Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.