The Construction of a New Evaluative GERD Questionnaire – Methods and State of the Art

David Armstrong a  Hubert Mönnikes b  Karna Dev Bardhan c  Vincenzo Stanghellini d

aMcMaster University, Hamilton, Canada; bCharité, Berlin, Germany; cDistrict General Hospital, Rotherham, UK; dS. Orsola-Malpighi Hospital, Bologna, Italy

Key Words
GERD · Questionnaire · Scale · ReQuest™ · Construction

Abstract
Gastroesophageal reflux disease (GERD) is one of the most prevalent diseases worldwide, and it is becoming increasingly important to monitor the effect of various interventions on GERD symptoms. There can be rapid temporal changes in the severity and frequency of patients’ symptoms as well as their health status and well-being, all of which could, theoretically, be monitored using diaries or questionnaires. However, current GERD monitoring instruments are not appropriate because they do not assess symptoms daily, they are not sufficiently responsive to short-term changes in health status or they are not adequately validated. To address these problems, the conceptual and psychometric requirements for a GERD symptom assessment questionnaire were identified. A dimension-based scale was designed to reduce the number of symptoms monitored on a daily basis, and the validation process was defined to produce parallel long and short forms of a scale for patients’ self-assessment of their GERD symptom response to therapy. These basic principles which underlie the successful development of a new, self-assessed symptomatic reflux questionnaire (ReQuest™) are also applicable to the development of validated questionnaires for daily symptom self-assessment in other disease areas.

Introduction
Gastroesophageal reflux disease (GERD) is a common, chronic condition which usually manifests symptomatically. In many patients, symptoms are accompanied by erosive esophagitis, diagnosed at endoscopy, and it is this physical manifestation which is most easily identified and monitored in response to treatment. However, a high proportion of GERD patients do not have erosive esophagitis and, even if they do, there is only limited correlation between the severity of endoscopic esophagitis and the severity or frequency of reflux symptoms [1, 2]. It is, therefore, extremely important to have a well-validated, reliable and sensitive instrument with which to document the burden of patients’ reflux symptoms, their quality of life and well-being, and their response to treatment. This is especially important for patients with endoscopy-negative reflux disease (enGERD), as their response to treatment cannot be monitored endoscopically [3]. The need for validated instruments to monitor symptoms and their
response to therapy is becoming increasingly evident in many areas of clinical research and practice; this is particularly important for conditions such as functional dyspepsia and the irritable bowel syndrome, which cause symptoms and impaired well-being in the absence of an objective structural abnormality.

Heartburn, acid regurgitation and odynophagia are considered to be highly specific for GERD and they are, therefore, often cited as the cardinal symptoms of GERD. However, the sensitivity of these symptoms for the diagnosis of GERD is low. The results of epidemiological and treatment studies indicate that heartburn is absent in 30–50% of patients with erosive esophagitis [4], and that esophageal erosions may be accompanied by other, less specific symptoms such as epigastric pain, pressure in the upper abdomen, a feeling of fullness, or flatulence [4]. These other symptoms are often disregarded, but this is inappropriate as they, too, are associated with impairment of patients’ quality of life and well-being. Furthermore, atypical symptoms respond, in many patients, to acid suppression therapy with proton pump inhibitors and, conversely, they persist in other patients, even if the typical symptoms have responded rapidly to proton pump inhibitor therapy.

There are several possible reasons for the dissociation between documented gastroesophageal reflux, documented esophageal injury and the presence of typical reflux symptoms. Symptoms may be absent if the esophagus is not sensitive to reflux; alternatively, reflux-related symptoms may be truly atypical or they may not be recognized by the patient as typical. Dissociation between reflux, lesions and symptoms may also occur because each is episodic or transient. For example, partial treatment may reduce reflux sufficiently to heal erosions without abolishing all symptoms. To date, there are very few data on the episodic or intermittent patterns of symptom variation in patients with erosive esophagitis and the extent of their response to therapy.

The measurement of symptom response is particularly important for patients with enGERD, defined by the Genval Workshop Report [5] as GERD in patients who have no endoscopic evidence of Barrett’s esophagus or esophageal mucosal breaks. Fass et al. [6] characterized enGERD as the presence of typical GERD symptoms caused by intrasophageal acid in the absence of visible esophageal mucosal injury at endoscopy. Thus, although enGERD is defined by the presence of symptoms, the various symptoms are difficult to monitor, and little is known about the timing and severity of their occurrence. Furthermore, although erosive esophagitis may be accompanied by atypical symptoms, it is not known whether the presence of these atypical symptoms, in the absence of esophageal erosions, can be considered indicative of enGERD.

To investigate GERD symptoms, their time of occurrence and their pattern, and to assess the effect of treatment, a number of different symptom scales have been developed. A comprehensive review [7], based on a determination of the critical components of an appropriate GERD symptom monitoring scale, concluded that there is, currently, no adequate scale for the evaluation of changes in GERD symptoms and quality of life. Most recently developed scales are primarily diagnostic, and those intended for evaluative purposes do not fulfill all of the requirements identified by Stanghellini et al. [7]. Specific deficiencies included incomplete assessment of symptoms during treatment, incomplete assessment of the full spectrum of GERD symptoms and incomplete fulfillment of the psychometric requirements for an instrument to be used as a primary outcome measure in clinical trials. In addition, some of these scales have not been validated whilst others have been validated at only one center [7].

The aim of this paper is to review the development of a sensitive, self-assessment symptom scale and describe how the standard approach may be adapted to produce a scale with the requisite attributes. The methodology is based on accepted psychometric theory and techniques for scale construction [e.g., ref. 8–10] and it will be described in detail, only insofar as it deviates from usual practice. The new self-assessment scale (ReQuest™) resulting from this process is reported elsewhere [11, 12], but the development strategy is presented in some detail, as the principles are applicable to the development and validation of self-assessed symptom questionnaires for other conditions.

**Scale Construction**

Scale construction is a complex, interdisciplinary process whereby a questionnaire is designed to provide a quantitative measurement of an abstract theoretical construct. Thus, to measure health status and treatment efficacy in GERD patients, the theoretical construct is composed of the various symptoms from which GERD patients suffer. In principle, this is simple, but in practice, there is considerable uncertainty about the extent to which individual symptoms are associated with GERD.
The process of scale construction is generally broken down into the following components which are designed to develop or identify: (1) the objects and purposes of measurement – what should be measured, based on the abstract theoretical model, (2) the method of questioning – how the information is to be obtained, (3) the graphical layout – how the questionnaire is presented, (4) the assessment standard, (5) the time point or points of assessment, (6) the psychometric theory on which the scale will be based, (7) the language or languages in which the scale will be presented, (8) a comprehensive item generation procedure (including interviews, patient focus groups, cognitive interviewing), (9) the initial exploratory raw scale, (10) the standard application, (11) the initial validation study to verify the scale empirically, (12) the item aggregation and item reduction process, (13) the construction and layout of the completed scale, (14) the psychometric calculations needed to score the scale, and (15) the interpretation and analysis of the scale scores.

**Objects and Purposes of Measurement**

In medical practice, the construction of a scale is determined by its purpose. For example, a diagnostic scale differs in construction from scales used to determine the stage of an illness at a given time point or changes in disease severity over time. In general, scales are constructed to achieve an optimal balance between specificity and sensitivity; however, most scales cannot be both highly specific and highly sensitive. Because it is intended to monitor symptoms in patients who have already been diagnosed with GERD, it was important that the new GERD symptom assessment scale is highly sensitive, whereas specificity was of secondary importance. Additionally, it was decided that the scale should be developed based on constructs applicable directly to GERD patients. For example, the Gastrointestinal Symptom Rating Scale is often used in GERD patients, despite the fact that it was originally developed for patients with the irritable bowel syndrome (IBS) or peptic ulcer disease (PUD). Subsequent validation studies have been performed in GERD patients to determine reliability and consistency, but it remains unclear precisely what the Gastrointestinal Symptom Rating Scale is measuring in GERD patients.

To be useful, an evaluative scale should record more than just the presence or absence of a given symptom; it should also record attributes such as frequency, intensity, duration and distress. Distress was not included in the ReQuest™ because it has been shown to be highly correlated with symptom frequency, intensity and severity. The remaining attributes, frequency and intensity, were assessed using a Likert scale and a visual analogue scale, respectively. The daily use of a visual analogue scale does obviously not allow patients to remember easily their status on the previous day, and therefore it is expected that this will enhance the sensitivity and test-retest reliability of the scale.

**Method of Questioning**

The formulation and presentation of the questions must be such that the respondent understands readily what information is required and how it can be provided, clearly and unambiguously. In general, patients should be given a restricted list of responses to each question; open-ended questions may provide greater opportunity to describe unique symptom features, but the answers are open to misinterpretation, and the subsequent analysis is very complex and time-consuming.

**Graphical Layout**

The layout and format of a questionnaire are crucial. The framework for presenting the questions and the linguistic structure contribute greatly to patients’ understanding of the task. If the structure is unclear, illogical or inconsistent, patients may not understand what is required of them, and this will compromise data quality. As a result, the ReQuest™ was tested in international patient focus groups during development to ensure that the layout was comprehensible and practicable [11].

**Assessment Standard**

Patients’ symptoms and their effect on well-being can be assessed either by the patients themselves or by an independent observer, such as the patient’s physician or a study investigator. From a theoretical standpoint, patients should report their own symptoms, particularly in the context of studies in which the response to therapy is being monitored. An independent observer’s judgment must still be based on the patient’s own assessment of a subjective phenomenon, and a second observer is likely to do no more than introduce another source of error or bias [13]. The increasing emphasis on patients’ self-reporting is gaining increasing support from regulatory authorities, with the expectation that self-assessed symptoms will become the primary outcome criteria in clinical studies of treatment for functional disorders. Patient self-assessment is also important because it is not feasible for an independent observer to assess symptoms on a daily basis, particularly when patients, with conditions such as GERD, are fully ambulatory in a community environment.
Time Points of Assessment

Gastroesophageal reflux is an intermittent event which can lead to frequent symptoms of variable severity and duration, although it does not always do so. A detailed assessment of GERD symptoms would require an hourly or even minute-by-minute record, but this is currently not feasible in clinical research or practice. On the other hand, less frequent assessment of symptoms, for example, every week or month, may not provide sufficient detail to evaluate the effects of different treatment options. As a compromise, it was decided that GERD symptoms should be assessed on a daily basis in the ReQuest™. A daily diary fits well into most patients’ daily schedules. A shorter interval would be too burdensome, whereas a longer interval might impair patients’ recall or recording of their symptoms. The decision to develop the scale in the form of a diary card is supported by the publication of Junghard et al. [14] in which they validated the use of a daily diary card, with a seven-graded scale, to measure the severity of pain or discomfort. However, for the ReQuest™ it was also necessary to establish content validity for all of the symptoms or items that were to be monitored using the daily diary card.

Psychometric Theory on Which the Scale Is to Be Based

The psychometric theory used most frequently for constructing scales in the medical field is the classical test theory (CTT). CTT is a method for analyzing one or more empirically identified personality characteristic(s) with the aim of deriving a quantitative measure of the level of each characteristic. According to CTT, the quantitative measure of the individual’s characteristic is an observed score that has two independent components: a ‘true score’ and an error. The ‘true score’ represents the real value for the level of that characteristic in the individual under investigation. This ‘true score’ is, therefore, independent of any changes attributable to the test environment or the use of parallel, alternate test procedures. The second component of the observed score is the error which may be affected by the test environment or format and the error is, therefore, unique to the individual measurement occasion.

The disadvantage of CTT is that the test outcome or scale score is predicated on the administration of a fixed-length test which uses the same set of questions or items on all occasions. Consequently, the scale cannot generally be shortened, lengthened or otherwise modified and the items cannot be weighted, without extensive revalidation [8]. Thus, the same questions must be administered in a standardized format to all patients.

In the framework of CTT, scales can be evaluated for objectivity, reliability, validity and responsiveness. Objectivity describes the extent to which a test result is independent of the person applying the test, the person analyzing the test and the person interpreting the test. Reliability describes the extent to which a test yields the same results after repeated measurements. Validity describes the accuracy with which a test measures the characteristic that it is intended to measure, and responsiveness describes the accuracy with which a test measures relevant changes in the target population.

Multilingual Strategy

The terms used to describe GERD symptoms vary considerably between countries and, within countries, between cultures. Symptom scales should, therefore, be developed and validated independently for different populations, even if the basic aims and design of the scale are applicable in all settings. Clinical trials are conducted increasingly as multinational, multicenter studies, and it is important to use the same, validated measurement tool for all patients. However, it is very difficult, logistically, to develop and validate a test simultaneously in many different languages, in part, because it requires a validated translation into the new languages, back-translation into the original language, cultural adaptation and, subsequently, testing in patient focus groups selected for different languages and cultures [11, 12]. It was, therefore, decided that the ReQuest™ should be constructed and verified, initially, in one language before being tested and validated in other languages [11].

Comprehensive Item Generation Procedure

An important step in developing a questionnaire is to generate a comprehensive item pool or list of symptoms. For this purpose, data on GERD-related symptoms are collected from a variety of sources, including GERD patients, physicians, experts and the relevant literature. This comprehensive, potentially over-inclusive list of items and symptoms is then reviewed to determine which items should be retained to fulfill the aims of the intended theoretical construct.

Initial Exploratory Raw Scale

A raw scale is the first iteration of an assessment scale; it uses those items which are retained after review of the comprehensive item pool and incorporates the form, structure, scoring and attributes of the scale, as determined previously in the construction process.
Standard Application
To ensure that a scale is used correctly and that it performs appropriately, everyone involved in the administration or completion of the scale must be fully informed about the scale and relevant aspects of its analysis, evaluation and interpretation. Development of the ReQuest™ was, therefore, accompanied by production of instruction manuals for the patient and the investigator, and an evaluation manual for the data analyst.

Initial Validation Study to Verify the Scale Empirically
Empirical verification requires that the performance of the raw scale is confirmed using empirical data, preferably in the context of a well-designed clinical trial, conducted according to the guidelines of Good Clinical Practice. The clinical trial should be designed and conducted such that patients' symptomatic response, measured using the raw scale, can be compared with their response measured using other comparable or complementary scales. This step was the basis for documenting the objectivity, reliability, validity and responsiveness of the ReQuest™.

Item Aggregation and Item Reduction
After empirical verification, some items may be observed to group together because they have one or more basic characteristics in common. Factor analysis will indicate whether some items can be aggregated into dimensions, and it may then be possible to reduce the initial item pool by excluding items that have a low frequency or are not definitely related to the content.

Construction and Layout of the Completed Scale
After completion of the previous steps, the scale can be revised and finalized.

Psychometric Calculations Needed to Score the Scale
After completion, one must confirm that the scale continues to fulfill the psychometric requirements of objectivity, reliability, validity and responsiveness. If the scale has been modified in crucial points, with respect to the context or layout, the psychometric features may change significantly. In this case, the Spearman-Brown formula for lengthening can be applied to estimate the change in test characteristics produced by these modifications [8, 15].

Problems and Possible Solutions

Problems
If there are still too many retained items or symptoms, despite the item aggregation and reduction procedures, completion of the scale will be too time-consuming for use as a daily assessment tool, and data quality will be compromised. The number of items may be reduced, still further, by expert review, before or after empirical verification [13]. However, it is important to decide who should conduct the expert review; arguably, it is the patients who assess their symptoms with the help of the scale and it is they, therefore, who are the real experts. Regardless of who conducts the review, there is a risk that expert review will lead to bias and the elimination of items before empirical verification may limit the validity of the final scale. Thus, if item elimination is necessary, it should be conducted after formal item reduction procedures and it should, if possible, address items that are answered or used by only a small number of patients. In the latter case, a minimal response incidence should be determined and applied to all items.

The number of GERD-related items poses a major problem for the ReQuest™ reflux scale. There are at least 50–100 items which merit inclusion in a GERD symptom scale and, although the number could be reduced [14], this would lead to an unacceptable reduction in scale validity. On the other hand, a scale that requires 50–100 questions to be answered daily is neither feasible nor acceptable. To be practicable, the ReQuest™ should comprise no more than 5–8 questions daily. In addition, the wording, description and language used to present these questions are absolutely critical. For example, Locke et al. [16] reported that the term ‘heartburn’ was understood differently by different patients. Thus, item reduction or elimination must ensure that the remaining terms are described clearly and unequivocally.

Solutions
To overcome these problems, it may be necessary to cluster GERD symptoms into dimensions, ‘dimension’ being a generic term for a collection of items or symptoms which can be grouped together on the basis that they share a common clinical or medical attribute. Once these symptoms or items are grouped together in a dimension, the attribution of a severity score to that dimension provides, to some extent, a global rating for the severity of all symptoms in that group. Based on an extensive review of the literature and discussions with experts, each item or symptom can be assigned to one of several dimensions,
and a provisional wording or title can be developed to describe each dimension [11, 12]. With this as basis, a raw scale of the ReQuest™ is constructed, using only the relevant dimensions. Thereafter, following the standard test construction procedure, the raw dimension scale is tested in GERD patients to determine whether the scale is understandable, practical and usable. The language and wording of the scales are revised to account for the fact that patients and physicians often use different phrases or terms to describe GERD symptoms, or that they understand the same term, such as ‘heartburn’, differently [17]. To resolve this, symptoms that are considered to be clinically meaningful in research or practice are translated or revised to ensure that they can be used interchangeably by physicians and patients.

The development of symptom dimensions then led to the construction of two scale versions – a short and a long version – which were validated to determine whether they measured the same symptom outcome [11, 12]. For the long version, comprising items and dimensions, the standard test construction procedure is applied at an item level. After testing, the initial assignment of the items to specific dimensions is reviewed, and items are reassigned to a new dimension if Spearman or Pearson rank correlation procedures indicate that they correlate better with other items in a different dimension. This procedure is also used to assess content and construct validity for the long version and, after validation, it is possible to ascertain which items (symptoms) are associated with each of the respective dimensions.

Following validation of the long scale version, one can assess the validity of the short version, which comprises only dimensions, by showing that it produces comparable or equivalent results to the long version at the dimension level. For this purpose, two approaches are available. One approach is to show that the long version is a simply lengthened version of the short version using the formula of Spearman and Brown [8, 15]. Another approach is to prove that the short and long versions are ‘parallel forms’ of the same test, at the dimension level. However, to demonstrate parallelism, it is necessary that both versions are completed by the same patients on separate occasions to confirm that they are measuring the same thing. Other than this, it is important that the test conditions are similar for the measurements and, in particular, that the patients receive no treatment that might introduce bias.

Based on these considerations, the standard scale construction procedure described above was modified (italicized sections) as follows for the development of the ReQuest™: (1) the objects and purposes of measurement – what should be measured, based on the abstract theoretical model, (2) the method of questioning – how the information is to be obtained, (3) the graphical layout – how the questionnaire is presented, (4) the assessment standard, (5) the time point or points of assessment, (6) the psychometric theory on which the scale is to be based, (7a) a multilingual strategy, (7b) a dimension selection. (8a) a comprehensive item generation procedure (including interviews, patient focus groups, cognitive interviewing), (8b) the assignment of items to the dimensions, (9) an initial, exploratory raw scale, (10) the standard application, (11) an initial validation study to verify the scale empirically, (12a) an item aggregation and item reduction process, (12b) prove parallelism (if two versions of one test were developed), (13) the construction and layout of the completed scale, (14) the psychometric calculations needed to score the scale, and (15) the interpretation and analysis of the scale scores.

This procedure led to two versions of the final scale, each with 7 dimensions. Both versions are designed to be completed daily by patients, and the 7 dimensions, encompassing all specific and non-specific GERD symptoms including sleep problems, are identical. Confirmation of the scale’s content and construct validity, its psychometric characteristics and responsiveness are presented in detail elsewhere [11, 12].

Discussion

The purpose of this paper is to present the challenges which arise when designing a valid, reliable and practicable scale for the daily, evaluative assessment of a plethora of GERD symptoms in the context of standard procedures for scale development in medical research and practice. An analysis of the problems inherent in the design of an evaluative scale led to the elaboration of a sophisticated, dimension-based scale, the ReQuest™, to fulfill the content and psychometric objectives, and to a subsequent modification of standard scale construction procedures to ensure appropriate validation. This approach was taken to account for the complex, multidimensional nature of GERD-related symptoms, but it should be applicable equally to the development of evaluative symptom assessment scales for other functional disorders such as the irritable bowel syndrome and non-ulcer dyspepsia.

The ReQuest™ permits daily monitoring of symptoms for GERD patients and it can, therefore, be used to monitor patients’ health status and well-being as well as treatment efficacy in clinical trials. In addition to its properties
as a dimension-oriented scale, the new instrument also functions as a daily diary. Daily diaries are used, increasingly, in clinical trials to record frequent records of patients’ symptoms and health status. However, unlike most diaries which are used simply to document points of special interest in clinical studies, the ReQuest™ meets the requirements of a fully validated questionnaire. Furthermore, this is the first report of a validated, dimension-oriented scale for GERD that can be used daily to monitor the rapid and episodic changes in symptom character and severity which are known to occur in GERD [18]. This is especially important because it is difficult to compare diaries and validated questionnaires [19] other than by evaluating face validity and, as a result, diaries have not generally been validated. One reason for the limited use of daily diaries in clinical research and practice is the practical difficulty inherent in carrying and completing a paper-based record. Because it is difficult to ensure that patients carry the diary with them at all times and complete it on a regular schedule, it has been proposed that electronic diaries or phone systems may provide a better medium for recording frequent events in a reliable manner. Electronic or phone-based systems do offer the promise of improved compliance compared with the use of a paper diary [20] and they can also be used to remind patients when a response is required and to correct or verify any data input. However, it should also be noted that the scales used for electronic diaries have not, generally, been validated any more extensively than current paper-based diaries. For this reason, further research is required to clarify whether an electronic diary-based questionnaire can be validated using CTT.

Although there has been much debate about who should assess the patients’ symptoms, it is increasingly clear that patient self-assessment is preferable since it is the patient, not the physician, who experiences and provides the initial assessment of any symptoms, and there is, subsequently, only a limited correlation between the physicians’ judgments and the patients’ self-reported health status. In general, physicians’ tendency to underestimate the severity of patients’ symptoms [21] may be responsible for the lower sensitivity and higher specificity of physicians’ assessments. This underlies the argument that physicians’ judgments should be used for diagnostic and disease severity staging assessments, even if they are not appropriate for an evaluative instrument such as that presented above. It has been argued that the burden of documentation may impair patients’ ability to assess their symptoms on a daily basis and that the quality of data gained will, therefore, be reduced. It was for this reason that the development of the ReQuest™ was supported, during the validation procedure, by a phone-monitoring program to remind the patients about completing the questionnaire and to answer possible questions in an effort to ensure high data quality [11, 12].

The idea of using dimensions in scales or questionnaires is not new. Aggregation of items after empirical verification usually leads to the identification of contributory factors, which are also known as dimensions. Factors, identified by factor analysis, are groupings of items that are considered to belong together because they share common characteristics. For example, factor analysis indicates that symptoms such as heartburn, acid belching and acid taste can be grouped or aggregated because they have the same attributes [22]. Similarly, Shaw et al. [22] reported that factor analysis of gastrointestinal symptoms identified 5 factors: IBS-diarrhea, IBS-constipation, reflux, dysmotility and pain. In both cases, the use of dimensions rather than individual items is appropriate, since daily assessment of multiple individual symptoms would be completely impracticable. The use of a small number of dimensions, therefore, permits frequent monitoring of the many constituent symptoms in an easy and concise fashion, using a questionnaire that is neither too complex nor too time-consuming for widespread use.

Acknowledgement

This work was supported by Altana Pharma AG, Konstanz, Germany.
References