urn:nbn:de:gbv:ilm1-2011imeko-039:5

FROM BREAST-Q © TO Q-SCORE ©: USING RASCH MEASUREMENT TO BETTER CAPTURE BREAST SURGERY OUTCOMES

Stefan J Cano [1], *Anne F Klassen* [2] *and Andrea L Pusic* [3]

[1] Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Plymouth, UK
[2] Department of Pediatrics, McMaster University, Hamilton, Canada
[3] Plastic and Reconstructive Surgery, Memorial Sloan-Kettering Cancer Center, New York, USA

Abstract – The two aims of this study were: 1) to bring together Rasch measurement methods (RMMs) with substantive clinically-grounded hypotheses to develop measures of health-related quality of life and patient satisfaction for reconstructive and cosmetic breast surgery; and 2) develop an accessible scoring program to provide automated clinically interpretable scores, based on calibrated item locations. We constructed a new patient reported outcome (PRO) instrument (BREAST-Q ©) from patient interviews (n=48), focus groups (n=18), and expert opinion. It was then field-tested in samples of breast surgery patients (n=1950 & n=817). Item generation led to three separate modules for different types of breast surgery, each with a pre- and post-operative version: 1) augmentation; 2) reconstruction; and 3) reduction. RMMs supported the summing of items to form a total score for all subscales, in each module. Based on these analyses the Q-Score © scoring algorithm program was developed, tested and finalised. The BREAST-Q © is an advance on the way PROs are currently measured and can provide essential information about the impact and effectiveness of breast surgery from the patients' perspective. The Q-Score © program enables the BREAST-Q © to be widely used and interpreted correctly, underpinned by an equivalent frame of reference across different clinical settings.

Keywords: patient reported outcome instruments; Rasch measurement; construct theory

1. INTRODUCTION

Traditionally the discussion of outcomes in plastic surgery has centred on the provider's perspective, focusing on measuring complications and considering photographic analyses. However, such data alone are no longer sufficient to support the progress being made in the field [1]. As the specialty of plastic surgery continues to develop, more sophisticated ways of examining outcomes are required. Recently, expressions such as "quality of life" and the "patient's voice" have caught the attention of consumers, the public, health care payers, and policymakers [2, 3].

Breast surgery is typical of the wider picture in plastic surgery. As such, rapidly advancing techniques, increasing involvement of patients in their own surgical decision-making and concern over escalating healthcare expenditures has resulted in growing scrutiny of surgical outcomes and cost [4-6]. More recently, this emphasis on evidence-based practice [7] has been coupled with a new focus on key indicators such as health-related quality of life and patient satisfaction (HR-QL) [1]. Thus, there is now more demand for high quality specially designed questionnaires, known as patient-reported outcome (PRO) instruments, in cosmetic and reconstructive breast surgery research, trials and practice [8].

Despite the growing demand, our systematic review found that only seven of 223 PRO instruments used in breast surgery studies had psychometric evidence to support their use in a breast surgery population [8]. Furthermore, only one instrument (developed for breast reduction patients) was reported to have psychometric properties in line with current proposed psychometric criteria [9]. However, this measure was limited in the range of outcomes it captured (i.e., did not address key areas such as aesthetics and body image) [8].

Therefore, we identified a need for a new clinically meaningful, scientifically sound PRO instrument that measures the perceptions of reconstructive and cosmetic breast surgery patients. Such an instrument could facilitate comparisons of different surgical techniques from a patient perspective, and provide a reference point for comparisons between studies and surgical populations. Therefore, the central objective of this project was to develop pre- and post-surgical measures of HR-QL and patient satisfaction for reconstructive and cosmetic breast surgery. As part of the study design process, we drew up a check list of three key areas to address, which we expand upon below. These ensured the development of an appropriate, practical, clinically meaningful and scientifically rigorous new instrument. As such, the new instrument should:

- appropriately capture the patient perspective and include clinically relevant and meaningful domains
- be underpinned by an appropriate measurement model
- be applicable for research and clinical settings and be easy to administer and score

1.1. The patient perspective

Guidelines for developing PRO instruments [9], including the current widely quoted US Food and Drug Administration's (FDA) scientific requirements for PROs in clinical trials [2, 10], highlight the importance of establishing validity. In particular, the FDA emphasises appropriate conceptual frameworks and definitions as being fundamental. These are best achieved using detailed qualitative assessments which should include: evaluating the extent to which a scale's items represent the construct to be measured; establishing the most appropriate item phrasing, structuring and context; and cognitive debriefing to ensure consistency in meaning.

In developing the new PRO instrument for cosmetic and reconstructive breast surgery, we selected a range of qualitative methods including in-depth patient and clinician interviews, literature review, panel meetings, and cognitive debriefing [11, 12].

However in addition to these methods, we also strove to develop explicit descriptions of each subscale, in order to maximise their utility as clinically interpretable tools. As such, the new PRO instrument was developed 'bottom-up' (from a construct definition), rather than 'top-down' (from a method of grouping items) to ensure that substantive clinical grounded hypotheses determined subscale content. This involved several rounds of iterative qualitative enquiry utilising the methods described above to establish clinical validity. This approach is keeping with Rasch paradigm [13, 14] (which we revisit below), and provides the optimal foundations to fully understand the measurement performance of each of the new subscales [15, 16].

Over the last 25 years one group outside of health measurement has developed the understanding of construct definitions and construct theories to an advanced level [17-19]. This group, led by Jack Stenner, has argued for a change in focus of assessing validity from studying the people to the items [18] and in particular the relationships between item characteristics and item scores. This forms the building blocks of the theory of the construct. Stenner et al use the following analogy to describe a construct theory: "The story we tell about what it means to move up and down the scale for a variable of interest (e.g. temperature, reading, ability, short term memory). Why is it, for example, that items are ordered as they are on the item map? [This] story evolves as knowledge increases regarding the construct" (p308) [17].

It is extremely rare to find PRO instruments associated with explicit construct theories in the health measurement literature. This may be due to historical reasons [20]. However, as a key part of our goal for this project was to appropriately capture the patient perspective and include clinically relevant and meaningful domains, we looked ahead to what could be achieved by following Stenner, et al's example, and therefore, focussed on developing as advanced as possible construct definitions (and thus begin to build construct theories) for each subscale.

1.2. An appropriate measurement model

In health measurement research, there are three main psychometric approaches broadly based on three types of measurement model: Classical Test Theory, Rasch Measurement Theory, and Item Response Theory [20]. Below, we briefly examine each of these approaches in order to present a justification for the approach that we chose.

1.2.1 Classical Test Theory

The dominant psychometric paradigm in the development and testing of PRO instruments, which is used in the current guidelines for developing PRO instruments, such as FDA document described above [2, 10] is classical (or traditional) test theory (CTT). Charles Spearman laid down the foundations of CTT in 1904 [21], in which he introduced the following equation:

$$O = T + E \tag{1}$$

where the observed score (O) is the person's manifest score on a scale. The true score (T) is the person's 'real' score. This is unobservable (a theoretical value) because of the associated measurement error – the error score (E). CTT then postulates that the observed score (O) is the sum of the True score (T) and the error score (E). Thus, it assumes that the relationship between the true score and the error score is additive rather than anything else (e.g. multiplicative).

Over the next 50 years, the role of CTT analyses grew with the accumulation of statistical evidence to establish the scientific robustness of measures (e.g. Kuder-Richardson's coefficients for internal inconsistency, Cronbach's alpha, correlations between replicated measurements) [22].

Importantly, CTT is grounded in the definition of measurement as proposed by Stanley Smith Stevens (ie 'the assignment of numerals to objects or events according to some rule') [23]. This definition differs in important respects from the more classical definition of measurement adopted throughout the physical sciences, which is that measurement is the numerical estimation and expression of the magnitude of one quantity relative to another [24]. CTT is based upon analyses of raw scores that are used to test the assumptions underlying a given measurement model; that the items can be summed (without weighting or standardization) to produce a score. The key traditional psychometric properties commonly associated with CTT are: data quality, scaling assumptions, targeting, reliability, validity, and responsiveness. We and others describe these assessments in more detail elsewhere [2, 25].

CTT was the cornerstone for psychometric evaluations during the last century in health measurement

[20]. The wide range of scale evaluations now associated with this approach provides relatively crude, but broadly useful examinations of the measurement performance of PRO instruments. However, there are some significant drawbacks to the approach. In brief, there are four main limitations of CTT. First, the measures generated are ordinal rather than interval (invariance is not hypothesized or experimentally tested) [26]. Second, scores for persons and samples are scale dependent, as they lack a stochastic frame of reference, resulting in item parameters that must be regarded as fixed [27]. Third, scale properties, such as reliability and validity, are sample dependent. As such, the marginal probabilities of measures (ie the probability distribution of scale scores) vary across population subgroups, as these subgroups may vary in the rate of the construct being measured [28]. Fourth, the data support group-level inferences but are not suitable for individual patient measurement [29].

1.2.2 Rasch Measurement Theory

Georg Rasch, a Danish mathematician, was principally concerned with the measurement of individuals rather than distribution of levels of a trait in a population. He argued that the core requirement of social measurement should be the same as that in physical measurement (ie 'invariant comparison'). With this in mind, he developed the simple logistic model (now known as the Rasch model). Through applications in education and psychology, he was able to demonstrate that his approach met the stringent criteria for measurement used in the physical sciences[30]. The formula for his model for scales including dichotomous items is:

$$Pr\{x_{ni}|\beta_n, \delta_i\} = \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \tag{2}$$

where $x_{ni} \in [0,1]$; β_n and $\delta_i$are the measurements of person *n* and item *i*, respectively, upon the same latent trait, and *e* is the natural logarithm constant (2.718).

Out of Rasch Measurement Theory are born Rasch measurement methods (RMMs), which use the Rasch model to evaluate the legitimacy of summing items to generate measurements, and their reliability and validity. The model articulates the set of requirements that must be met for rating scale data to generate internally valid, equal-interval measurements that are stable (invariant) across items and people [26]. The central tenet to RMMs is that they examine the extent to which observed data (patients' actual responses to scale items) accord with ("fit") predictions of those responses from a mathematical (Rasch) model. Thus, the difference between what should happen (expected) and what does happen (observed) indicates the extent to which measurement is achieved.

Statistical and graphical tests are used to evaluate the correspondence of data with the model. Certain tests are global, while others focus on specific observations, items or persons. There are seven key measurement properties that should be considered: thresholds for item response options; item fit statistics; item locations; differential item functioning (DIF); correlations between standardised residuals; person separation index (PSI), individual person change statistics. We describe these in more detail elsewhere [25].

Direct comparisons of CTT and Rasch psychometric methods in the health measurement literature are sparse, and at best superficial [31, 32]. In part, this may be due to the fact that the two approaches cannot be compared easily, as they use different methods, produce different information, and apply different criteria for success and failure.

However, importantly RMMs address each of the four limitations of CTT described above. First, the approach offers the ability to construct linear measurements from ordinal-level data, thereby addressing a major concern of using PRO instruments as outcome measures [33, 34]. Second, Rasch measurement methods provide item estimates that are free from the sample distribution and person estimates that are free from the scale distribution, thus allowing for greater flexibility in situations where different samples or test forms are used [35]. Third, the methods allow for the use of subsets of items from each scale rather than all items from the scale, without compromising the comparability of measures made using different sets of items. This is the foundation for item banking and computerised adaptive testing [36]. Fourth, RMMs enable estimates suitable for individual person analyses rather than only for group comparison studies.[37, 38]

1.2.3 Item Response Theory

Item Response Theory (IRT) is another body of psychometric methods that provides a foundation for statistical estimation of parameters that represent the locations of persons and items on a latent continuum [39]. In particular, IRT scale evaluations are used to ascertain the degree to which a given model and parameter estimates can account for the structure of and statistical patterns within a response dataset [13, 39].

There are three main models under the general banner of IRT. The one parameter (1P) model is essentially identical in structure to the Rasch model (Equation 2). Mathematical models relating the probability of a response to an item, to the person's location, the item's difficulty and the item's discrimination are known as two parameter (2P) models [40]. The addition of a third parameter (a person guessing parameter [41]) to the basic 2P model results in the 3P model [42]. Thus, the basic 2P model for rating scales including dichotomously scored items is:

$$Pr\{x_{ni}|\beta_n, \delta_i, \alpha_i\} = \frac{e^{x_{ni}[\alpha_i(\beta_n - \delta_i)]}}{1 + e^{[\alpha_i(\beta_n - \delta_i)]}} \tag{3}$$

which is equivalent to Equation (2), with the addition of α_i, which represents the slope of item *i* (discrimination). Thus, the basic 3P model is:

$$Pr\{x_{ni}|\beta_n, \delta_i, \alpha_i, c_i\} = c_i + (1 + c_i) + \frac{e^{x_{ni}[\alpha_i(\beta_n - \delta_i)]}}{1 + e^{[\alpha_i(\beta_n - \delta_i)]}} \quad (4)$$

In this instance, the additional c_i represents a constant describing the lower asymptote due to guessing for item *i*.

The general approach in IRT focuses on mathematical models that explain the observed data. Essentially, models are postulated and examined relative to data. When the observed data are not adequately explained by the mathematical model, that is when the data do not fit the chosen model closely enough, another model is tried. Thus, the justification for model selection is empirical evidence of its suitability [22]. The choice of one model over another hinges on whether it accounts better for the data [43]. The data are considered given in the sense of being validated on the basis of item content, expert opinion or other processes external to the data. Finally, the four shortcomings of CTT, described above, are overcome by IRT only haphazardly and indirectly, depending on whether the data in hand are found to fit a 1P model that, importantly, has not been conceived from the standpoint of specifying the requirements for objective inference.

1.2.4 Justification for model choice

As outlined above, modern psychometric methods, such as RMMs, have substantial advantages over CTT for developing new PRO instruments. However, given the apparent similarity between Rasch measurement theory and IRT, does it matter which approach is used? Rasch measurement theory and IRT are mathematically similar, so they are often considered as members of the same family of statistical techniques [13, 44]. This is inaccurate because advocates of Rasch measurement theory and IRT have different research agendas [13, 44, 45].

The distinction between Rasch measurement theory and IRT is subtle but important. IRT models are statistical models used to explain data, and as such the aim of an IRT analysis is to find the statistical model that best explains the observed data [13, 44]. When the observed data do not fit the chosen IRT model we seek another model to better explain the data. In contrast, Rasch measurement theory provides a mathematical model for guiding the construction of stable linear measures from rating scale (e.g. PRO instrument) data [30]. Therefore, the aim of RMMs are to determine the extent to which observed rating scale data satisfy the measurement model. When the data do not fit the model, we examine the data carefully to try and explain the misfit, but ultimately we choose data that satisfies the model's requirements. It is the central tenet of the Rasch Model that distinguishes it from IRT models. Specifically, its defining property is its mathematical embodiment of the principle of invariant comparison. These central tenets distinguish the Rasch measurement theory *diagnostic* paradigm from the IRT *modelling* paradigm [13].

Therefore, in developing the new PRO instrument for cosmetic and reconstructive breast surgery, we selected Rasch measurement methods (RMMs). In particular we used the Rasch model for ordered response categories, which was developed for scales or tests containing polytomous items [15, 46]:

$$\Pr\{X_{ni} = x\} = \frac{[exp(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ki})]}{\sum_{x=0}^{m_i}[exp(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ki})]} \quad (5)$$

Where $x \in [0,1,2,\ldots,m_i]$ is the integer response variable for person *n* with the ability β_n responding to item *i* with the difficulty δ_i and

$$\tau_{1i}, \tau_{2i,\ldots}, \tau_{mi}, \sum_{x=0}^{m} \tau_{xi} = 0$$

are thresholds between m_i +1 ordered categories where m_i is the maximum score of item i, $\tau_0 \equiv 0$ [46]. This implies a single dimension with values β, δ and τ located additively on the same scale. Thus, the set of positive integers *x* can contain person's response to summated rating scales. For example, in the new PRO instrument described in this paper, the satisfaction related items include response options containing the integers '1', '2', '3', and '4' which represent the semantic categories 'Very dissatisfied', 'Somewhat dissatisfied', 'Somewhat satisfied', and 'Very satisfied'.

Importantly, the equation in the bottom half of the Equation 5 is the 'normalising factor', which specifies the probabilities of exceeding all thresholds for all categories preceding *k*, so that the probability of person *n* scoring in category *k* depends upon all the locations of all the thresholds. This is important as it ensures that responses to items are constrained to a Guttman pattern, whose success is reflected by ordered thresholds. Thus, correctly ordered thresholds become an essential element of the validity of items [46]; an evaluation supported by readily available software [47] and theory [48, 49].

1.3. Applicable for research and clinical settings

RMMs for testing and evaluating PRO instruments are becoming increasingly used in health measurement research [2]. However, for new PRO instruments to be appropriately used and widely accepted in different clinical scenarios, clinicians require well targeted, reliable, and valid instruments that can also be easily scored. To achieve this requires both a psychometrically robust PRO instrument and a method of auto-

matically scoring its data, based on items that are appropriately calibrated within a specifically defined, clinically meaningful, frame of reference. This is an area which has received less attention in the health arena [20].

Therefore, the final step in the development process was to produce a standalone executable software application to allow data entry, automatic scoring, and export based on the most applicable scoring algorithms. We selected the item estimates and calibration algorithms housed in the RUMM 2030 software program [47]. This is because these algorithms are directly referable to work of George Rasch [30] and David Andrich [15, 38, 46, 48, 49].

2. METHODS

The key steps involved in developing the new PRO instrument included: development of a conceptual framework; item generation; scale formation; and psychometric evaluation. In addition, as described above, we aimed to build, from the ground up, clinical hypotheses for each subscale that could be used to postulate a construct theory for each, and then test each scale using RMMs. Local institutional ethics review board approval was obtained for participating centres.

2.1. Conceptual Framework Formation

To develop a conceptual framework of HR-QL and patient satisfaction in breast surgery, we conducted semi-structured interviews with breast reconstruction, augmentation, and reduction patients [11]. The interviews were used to collect rich, detailed data about the personal experiences of breast surgery patients. Interviews took place in Vancouver (Canada) between October and December of 2004. A maximum variation sample was chosen to ensure that a broad spectrum of age, ethnicity, and surgery types were represented. Patient interviews were tape recorded, transcribed, and analysed.

2.2. Item Generation, Preliminary Scale Formation, and Pretesting

Item generation involved developing an exhaustive list of potential items for each domain within our conceptual framework. As specific issues varied in importance by surgical group, separate modules were developed for breast augmentation, reduction, and reconstruction patients. Items for each module were developed using information generated only from interviews with patients who had undergone that particular type of breast surgery. We also examined existing published measures of HR-QL and patient satisfaction in breast surgery patients and added items not discussed by our interviewees. Finally, we had plastic surgeons, oncologist breast surgeons, nurses, and psychologists working at the University of British Columbia (Vancouver), Memorial Sloan-Kettering Cancer Center (New York), and University College London (London) nominate items that were missing from their perspective.

Our conceptual framework and draft subscales were then presented for feedback to two separate focus groups that included women from the initial qualitative interviews and new participants. These sessions were used to examine the degree to which our conceptual framework resonated with them and covered all relevant issues. In addition, experts at four academic medical centres in the United States and Canada were asked to review the framework and subscales. This led to finalizing draft versions of the subscales. Using cognitive debriefing techniques, we asked women to review these draft versions to determine their understanding of each item, to point out any unclear or ambiguous items, and to comment on the response options and recall periods. Finally, readability of the draft subscales was assessed to ensure that all content was targeted to a sixth-grade reading level, and revisions were made as necessary [50].

2.3. Preliminary Field-Testing, Scale Construction, and Psychometric Evaluation Postal Survey

Field-test versions of the three procedure-specific PRO instruments were mailed to breast surgery patients recruited from five centres in the United States and Canada (n=2715). Eligible participants included preoperative and postoperative patients who were able to read English and were aged 18 or older. To ensure a high response rate, we used personalised letters, standardised instructions, and up to two reminders as necessary [51, 52].

2.4. Further Validation Field-Testing, Psychometric Evaluation Postal Survey

Field-test versions of the three procedure-specific PRO instruments were mailed to breast surgery patients recruited from three centres in the United States and Canada (n=1244). Administration methods were the same as 2.3.

2.5. Psychometric Analysis

Essentially, RMMs are used to examine the extent to which the observed scale data 'fit' with predictions of those ratings from the Rasch model (which defines how a set of items should perform to generate reliable and valid measurements) [30]. Effectively, the difference between expected (as predicted by the model) and observed scores indicates the degree to which valid measurement is achieved [25, 38, 53]. In this project we examined four key tests for reliable and valid measurement using RUMM2030 software (fit, targeting, dependency, reliability) [47].

Fit. The items of each of the new PRO instrument's subscales must work together (fit) as a conformable set both clinically and statistically. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of each target construct. When items do not work together (misfit) in this way, the validity of a subscale

is questioned. In brief, evidence for item fit was based on four different indicators. These included: ordering of item response options (ordering of item thresholds [54]), two statistical indicators (fit residual; chi square), and one graphical indicator (item characteristic curve; ICC[35]).

Targeting. Scale-to-sample targeting concerns the match between the range of the target construct measured by each of the subscales and the range of target construct in the sample of women in each dataset. In brief, this was achieved by an examination of the spread of person and item locations in these two relative distributions. This analysis informs us as to how suitable the sample is for evaluating the subscales of the new PRO instrument and how suitable the subscales are for measuring their respective samples. Better targeting equates to a better ability to interpret the psychometric data with confidence [35].

Dependency. The responses to any of the items in each of the subscales should not directly influence the response to any another in the same subscale [55]. If this happens, measurement estimates may be biased and reliability may be artificially elevated. RMMs determine this effect by examining residual correlations.

Reliability. This was assessed using the Person Separation Index (PSI) [48], which is comparable to Cronbach's alpha (α) [56]. As such, both indices are estimates of the proportion of the variance of the distribution of person estimates. However, a key difference between the PSI and α is that when there is mistargeting between item and person locations, so that there is a skewed distribution with extreme raw scores, α remains more constant than the PSI. This is because α is based on raw scores while the PSI involves a non-linear transformation of these raw scores. The error variance for persons increases as the scores become more extreme, so with scores close to the extreme, the error variance increases in the PSI while this is not taken into account in α [48].

2.6. Final Cognitive Debriefing Interviews

After the completion of the field test, final modules and subscales were mailed to a small sample of patients in each of the three procedure groups. These patients participated in cognitive debriefing interviews by phone and were asked to discuss their understanding of the items and to identify unclear or ambiguous items. Acceptability and completion time were also examined.

2.7. Software development

Following finalising the items (and item responses) for each subscale in each module, further item analyses based on the RUMM paradigm were conducted to provide a precise set of item calibrations to base the construction of patient measures. As such, the item calibrations, in conjunction with the item statements, define what "more than" or "less than" means in terms of the patient location on the final calibrated scale, a fundamental criterion behind all scientific measurement. These item calibrations in conjunction with the RUMM2030 algorithms were used to develop the software engine for a new scoring program together with a Graphical User Interface to allow for ease of use. Together, these produce an application that would be able to provide automated clinically interpretable scores, based on fixed calibrated item locations.

3. RESULTS

3.1. Conceptual Framework Formation

Forty-eight breast surgery patients were interviewed, generating a total of 2749 statements about HR-QL and patient satisfaction. Based on patient interviews, research literature, and expert opinion, the following six key themes formed our conceptual framework of HR-QL and patient satisfaction in breast surgery (Fig. 1):

Fig. 1. BREAST-Q © conceptual framework

3.2. Item Generation, Preliminary Scale Formation, and Pretesting

The process of item generation led to three initial pools of items from augmentation (n = 145 items), reconstruction (n = 240 items), and reduction (n = 163 items) patients. Items within each pool were grouped into domains based on their conceptual meaning to represent coherent clinically meaningful constructs. These formed the domains of the conceptual framework for each type of surgery. The research team then iteratively and interactively examined each of the item lists in each of the domains to identify and retain those potential items that best represented aspects of the continuum of impact for each domain and to form the best potential subscale. Preoperative items were repeated in the postoperative subscales along with additional "postoperative only" questions (e.g., items related to scarring).

This process resulted in the core domains of the conceptual framework and led to the development of preliminary subscales for each of the three modules. Four additional preliminary subscales to address issues specific to single procedure groups were also created (i.e., *Reconstruction module*: satisfaction with abdominal appearance, physical well-being trunk and abdomen, and satisfaction with nipple-areola reconstruction; *Reduction module*: satisfaction with nipple-areola).

Instructions for completing each subscale asked patients to comment on their HR-QL or satisfaction aspects of during the previous 2 weeks. This recall period was determined to be acceptable to patients, clinically relevant, and best conveyed as sense of "current status" for the patient groups. An exception to this rule was made for the sexual well-being subscale as "2 weeks" was not felt to be an acceptable interval given that sexuality in the weeks preceding surgery was not necessarily indicative of a patient's usual status. Cognitive debriefing interviews included 12 reduction, 11 augmentation, and 23 reconstruction changes in wording.

3.2. Field Testing, Scale Formation, and Psychometric Evaluation Postal Survey

3.2.1 Sample

Questionnaire booklets were sent to 2715 women, and 1950 were returned completed (response rate, 72 percent; Table 1).

Table 1: Patient Characteristics: Field Testing

	Presurgery	Postsurgery
n	908	1807
Age		
Mean (SD)	43 (14)	47 (12)
Range	18–84	18–84
Surgery type		
Augmentation	222 (33%)	179 (14%)
Reduction	148 (22%)	316 (25%)
Reconstruction	295 (45%)	790 (61%)
Site		
New York (Memorial Sloan-Kettering Cancer Center)	80 (13%)	355 (28%)
British Columbia (University of British Columbia)	320 (48%)	478 (37%)
Michigan (University of Michigan)	146 (22%)	245 (19%)
New Hampshire	21 (3%)	207 (16%)
Utah	98 (14%)	—
Marital status		
Married/common law	449 (68%)	999 (74%)
Other	203 (32%)	349 (26%)
Ethnicity		
Caucasian	441 (69%)	1156 (86%)
Other	198 (31%)	182 (14%)
Education		
Less than high school	17 (3%)	41 (3%)
High school diploma	96 (15%)	169 (13%)
University diploma	521 (82%)	1122 (84%)
Employment status		
Employed	369 (61%)	847 (66%)
Retired	71 (12%)	145 (11%)
Unable to work	24 (4%)	45 (4%)
Other	142 (23%)	240 (19%)
Income (U.S. $)		
<$40,000	163 (27%)	257 (20%)
$40,000 to $100,000	304 (50%)	672 (52%)
>$100,000	143 (23%)	360 (28%)

3.2.2 Scale Formation

In each subscale, items with the best psychometric properties, while appropriately representing each of the target constructs, were retained in the reconstruction, reduction, and augmentation modules. We named this new PRO instrument the BREAST-Q ©. Further information is available from the authors.

3.2.3 Psychometric Evaluation

Rasch analysis supported the summing of items to form a total score for each subscale of each of the modules. Scale reliability was supported by satisfactory Person separation indices (>0.76), and validity was supported by three findings. First, item response option thresholds were ordered correctly for all items, indicating that the proposed response options were working well. Second, the item locations in each subscale were spread out (range, 0.9 to 4.4), indicating that each subscale defined a continuum. Third, the vast majority of residual correlations for each subscale were less than 0.30, supporting local independence among items (data available from authors). Fit to the Rasch model was good, as all of the retained items in each subscale had acceptable fit residuals and the majority of chi-square values were non-significant.

The minority of items that had fit statistics slightly larger than expected were examined and retained on the basis of appraisals of overall psychometric performance and clinical relevance. Further details about item calibrations, standard errors, fit residuals, and chi-square statistics are available from the authors.

To illustrate, we provide an example of one of the subscales below. As such, Fig. 2 shows the subscale structure and layout of the BREAST-Q © Reconstruction Module: Satisfaction with Breasts subscale. This subscale address issues surrounding patient satisfaction relating to their perceptions of the result of reconstruction surgery. This is almost always following unilateral or bilateral mastectomy following a diagnosis of breast cancer.

BREAST-Q™
RECONSTRUCTION MODULE (POST OPERATIVE) 1.0

SATISFACTION WITH BREAST(S)

With your breasts in mind, in the past 2 weeks, how satisfied or dissatisfied have you been with:

		Very Dissatisfied	Somewhat Dissatisfied	Somewhat Satisfied	Very Satisfied
a.	How you look in the mirror clothed?	1	2	3	4
b.	The shape of your reconstructed breast(s) when you are wearing a bra?	1	2	3	4
c.	How normal you feel in your clothes?	1	2	3	4
d.	The size of your reconstructed breast(s)?	1	2	3	4
e.	Being able to wear clothing that is more fitted?	1	2	3	4
f.	How your breasts are lined up in relation to each other?	1	2	3	4
g.	How comfortably your bras fit?	1	2	3	4
h.	The softness of your reconstructed breast(s)?	1	2	3	4
i.	How equal in size your breasts are to each other?	1	2	3	4
j.	How natural your reconstructed breast(s) looks?	1	2	3	4
k.	How naturally your reconstructed breast(s) sits/hangs?	1	2	3	4
l.	How your reconstructed breast(s) feels to touch?	1	2	3	4
m.	How much your reconstructed breast(s) feels like a natural part of your body?	1	2	3	4
n.	How closely matched your breasts are to each other?	1	2	3	4
o.	How your reconstructed breast(s) look now compared to before you had any breast surgery?	1	2	3	4
p.	How you look in the mirror unclothed?	1	2	3	4

Fig. 2. BREAST-Q © Reconstruction Module: Satisfaction with Breasts subscale.

The item content and ordering of this subscale reflects the hypothesised clinical hierarchy of the construct. The items are ordered from the items which reflect lowest through to highest patient satisfaction. Thus, a patient endorsing degrees of satisfaction with the first item "How you look in mirror clothed", but dissatisfaction with all other items indicates low satisfaction for this woman. Alternatively, a patient scoring "Very satisfied" to the last question ("How you look in mirror unclothed") indicates the highest level of satisfaction. This hierarchy is supported in the RMMs analyses, and is demonstrated in the item threshold map and correctly ordered thresholds (Spearman's Rho=0.97 between hypothesised and Rasch measurement derived subscale hierarchy; Fig. 3). This is further elaborated upon in Fig. 4 which shows category probability curves (CPCs) from item i) 'How equal in size your breasts are to each other'.

Item fit to the Rasch model was good overall. Three items just marginally fell outside (<0.4) of the fit residual guidelines of -2.5 to +2.5, but all three items demonstrated good psychometric properties in all other tests. Fig. 5 shows an example of one of the graphics associated with tests of fit for item n) 'How closely are your breasts matched to each other'.

The final illustration is presented in Fig. 6, which shows the targeting of the subscale to the sample. In this figure, the upper histogram represents the sample distribution of total BREAST-Q © Reconstruction Module: Satisfaction with Breasts subscale person measures. The lower histogram striped blocks represent the sample distribution of the item thresholds of the 15-items of the same subscale. The line shows the information function. The graph shows that the distributions of item thresholds and person measures are well matched. There is also some potential to extend the range of measurement by adding items to the extremes of the subscale range. The peak of the information plot is around 0.9 logits of the continuum which indicates the scale's best point of measurement. The PSI=0.94.

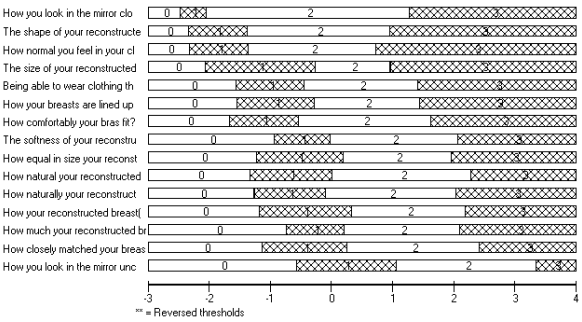

Fig. 3: Threshold map for all items in the BREAST-Q © Reconstruction Module: Satisfaction with Breasts subscale. The x-axis represents the construct (satisfaction with breasts after reconstruction surgery), with patient satisfaction increasing to the right. The y-axis shows each of the items response category 'Very Unsatisfied' labelled 0; Response category 'Somewhat dissatisfied' labelled 1 (black block); Response category 'Somewhat satisfied' labelled 2; Response category 'Very satisfied' labelled 3

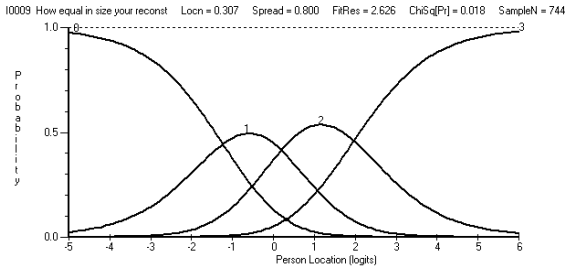

Fig. 4: Category probability curves for item i) 'How equal in size your breasts are to each other'. The x-axis represents the construct (satisfaction with breasts after reconstruction surgery), with patient satisfaction increasing to the right. The y-axis shows the probability of endorsing the response categories, reading left to right: 0 (first curve) 'Very Unsatisfied', 1 (second curve) 'Somewhat dissatisfied', 2 (third curve) 'Somewhat satisfied', and 3 (fourth curve) 'Very satisfied'. Key: Locn = location; FitRes = Fit residual; Pr = probability

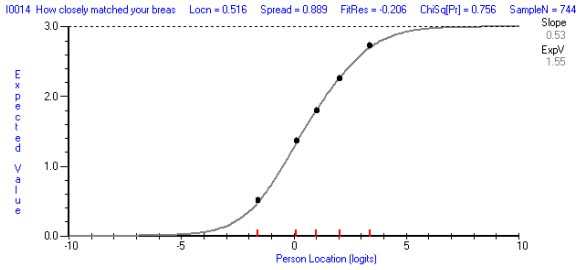

Fig. 5: Item characteristic curve for n) 'How closely are you breast matched to each other'. The x-axis represents the construct (satisfaction with breasts after reconstruction surgery), with patient satisfaction increasing to the right. The y-axis shows the expected value as predicted by the Rasch model. The black dots, which represent class intervals, are very close to the line indicating a close association between observed and expected scores. Key: Locn = location; FitRes = Fit residual; Pr = probability

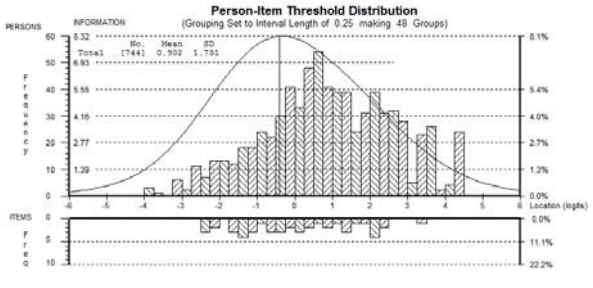

Fig. 5: Person-Item thresholds Distribution. The x-axis represents the construct (satisfaction with breasts after reconstruction surgery), with patient satisfaction increasing to the right. The y-axis shows the frequency of person measure locations (top histogram) and item locations (bottom histogram).

3.3. Further Validation Field Testing, and Psychometric Evaluation Postal Survey

3.3.1 Sample

Questionnaire booklets were sent to 1244 women, and 817 were returned completed (corrected response rate 66%; Table 2).

3.3.1 Psychometric Evaluation

Rasch analysis supported the summing of items to form a total score for all subscales of each module. Validity was supported by three findings. First, item response option thresholds were ordered correctly for all items in all subscales, indicating that the proposed response options worked well. Second, the item locations in each subscale were spread out (range of logit span 0.7-6.6) indicating that each subscale defined a continuum. Third, fit to the Rasch model was good as the vast majority of items in all subscales of each module had acceptable fit residuals, and Chi-square values that were non-significant. The minority of items falling outside recommended criteria had fit statistics marginally larger than expected. Scale reliability was supported by satisfactory Person Separation Indices (>/= 0.73) with the exception of Physical Wellbeing. Further information is available from the authors.

Table 2: Patient Characteristic: Further Validation Field Testing (post-surgery)

N	*817*
Age in years	
Mean (SD)	49 (12)
Range	20-82
Surgery type	
Augmentation	158 (19%)
Reduction	301 (37%)
Reconstruction	358 (44%)
Recruitment site	
New York (MSKCC)	158 (19%)
Vancouver (UBC)	301 (37%)
Dartmouth-Hitchcock (D-H)	358 (44%)
Marital status	
Married/Common law	596 (73%)
Other	221 (27%)
Ethnicity	
Caucasian	686 (84%)
Other	131 (16%)
Highest level of education	
Less than high school	20 (2%)
High school diploma	97 (12%)
University/diploma	700 (86%)
Employment status	
Employed	517 (63%)
Retired	70 (9%)
Unable to work	23 (3%)
Other	207 (25%)
Household income (Canadian/USD)	
<40K	135 (17%)
40-100K	343 (33%)
>100K	279 (34%)
Not reported	60 (7%)

MSKCC=Memorial Sloan Kettering Cancer Center; UBC = University of British Columbia

3.4. Final Cognitive Debriefing Interviews

Thirty patients (10 from each procedure group) reviewed the final modules and subscales. They reported completion time to be 10 to 14 minutes for the reconstruction, 10 to 12 minutes for reduction, and 8 to 10 minutes for augmentation modules. They found the subscales to be acceptable, comprehensive, and clear.

3.5. Software development

The Q-Score © application (Fig. 7.) was developed to transform each BREAST-Q © subscale score in each module. It provides the ability to read patient subscale response data into the program, score the set of responses to each subscale attempted, and write the complete set of transformed scores for all subscales attempted to an electronic file. Once the set of responses are accepted, the program immediately scores these data and estimates a Rasch-based person measure, ranging from 0 to 100. This measure is based on the calibration of each set of items in each subscale. All item response data, scoring and measures can then be exported into text file for further analyses.

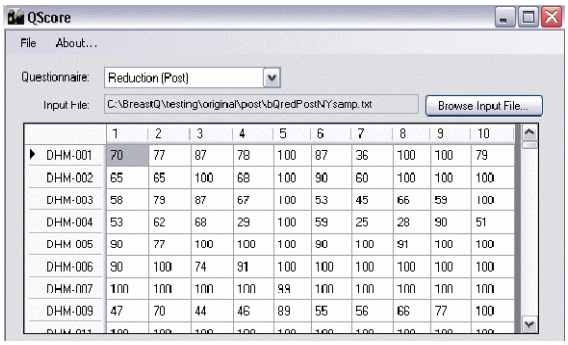

Fig. 7. Q-Score user interface.

4. DISCUSSION

Patients provide a unique and vital perspective on the success of breast surgery procedures. To fully capture and quantify their perceptions, appropriately constructed and validated PRO instruments are needed. The new instrument (the BREAST-Q©) consists of three procedure-specific modules (augmentation, reduction, and reconstruction) with independent subscales that examine those issues most important to women who have undergone each procedure. By closely targeting the subscales and items to the specific surgical group, each module has the potential to be more sensitive to patient perceptions and responsive to change following surgery.

4.1. The "Patient Voice"

Overall, patient input proved to be the most important element of the development process. In developing the conceptual framework for the BREAST-Q, we sought to create a model that would reflect the entirety of the patient experience. Patients expressed both a sense of "appraisal" of the results of surgery (e.g., satisfaction or dissatisfaction), as well as an awareness of the impact of the procedure on their health-related quality of life. Our conceptual framework thus consists of both HR-QL subscales (psychosocial, physical, and sexual well-being) and satisfaction subscales (satisfaction with breasts, satisfaction with overall outcome, and satisfaction with the process of care).

Patients in our interviews and focus groups repeatedly reflected on their relationship with the surgeon,

the information that they received, and the care provided by the office staff. Process of care is measured by separate subscales that examine satisfaction with preoperative information and the care provided by the plastic surgeon, the office staff, and other members of the medical team. These process measures may ultimately prove to be useful for quality improvement efforts. As an example, a plastic surgeon may use these subscales to obtain useful metrics for individual practice improvement.

4.1. Benefits of Rasch Measurement Methods

The use of RMMs to develop and test the subscales of the BREAST-Q © means that we have a good understanding of the *empirical* item order across each subscale. Thus, we know which items are associated with each and every possible subscale score. For example, using the BREAST-Q © Reconstruction Satisfaction with Breasts subscale (described above), a recent multicenter, cross-sectional study of 672 post mastectomy women, conducted by our group, found that women's satisfaction with their breasts was significantly greater among those who received silicone implants compared with those who received saline implants [57]. We can add the words used in the items in the range of the subscale scores for the silicone group (mean score 64) and the saline group (mean score 57; Fig. 8).

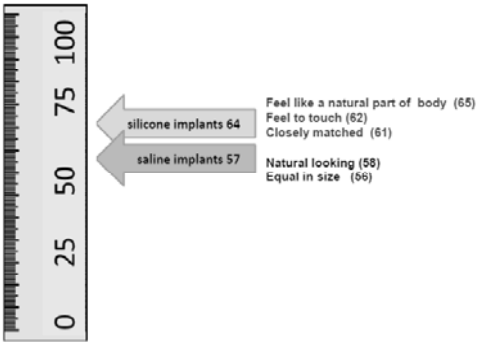

Fig. 8 Illustration of comparison of saline and silicone group means scores on the "ruler" of BREAST-Q © Reconstruction Satisfaction with Breasts subscale

These translate into the following: women in the silicone group scoring higher up the subscale and, therefore, typically satisfied with their look and feel of their reconstructed breasts, whereas women in the saline group category are towards the middle of the subscale and are satisfied with size and look of their breasts, but not how well they match or feel natural. The ability to provide qualitative statements to each group for each subscale score begins to make concrete the meaning of subscale scores and thus provides a clear base for the clinical interpretation of the BREAST-Q © [16].

4.3. A practical solution to scoring

Traditionally, PRO instrument users have been required to handle data manually, producing total scores via scoring syntax in statistical software packages, or other similar tools. These complications can be a barrier to the use of PRO instruments, especially for clinicians in busy practices or clinical researchers who are using these instruments as part of a larger study. Overcoming these barriers is especially relevant in the context of the more complex algorithms and software associated with RMMs. The new Q-Score © program offers a ready solution to this issue. It provides an easy, automatic, convenient and tangible 0-100 score transformation that maintains a common frame of reference and metric comparability across different clinical samples; linearised (not ordinal) measures satisfying statistical assumptions of unit additivity; and it supports both group and individual-patient level comparisons.

4.4. Next steps

Based on the development process and the preliminary validation data, we would argue that the BREAST-Q © is a promising PRO instrument that provides a scientifically rigorous and clinically valid means to examine the impact of breast surgery from the patient perspective. As increasing numbers of researchers and surgeons incorporate the BREAST-Q © into their studies and surgical practices, we envision a rapidly expanding knowledge base that will inform further clinical interpretation of BREAST-Q © data. Thus, as BREAST-Q © data grows from different clinical scenarios, the interpretation and, therefore, the clinical meaning, of its subscale scores will become increasingly clarified. As we seek to optimally manage surgical patients in an increasingly cost-restricted environment, the BREAST-Q © can be expected to provide meaningful data to guide determination of comparative effectiveness and patient advocacy.

We would recommend that further work is carried out on the BREAST-Q ©. First, our response rate (66%) in the validation filed test was lower than achieved in our first field-test (72%) [12]. This difference is probably due to the increased respondent burden of the substantially longer questionnaire booklet used in the validation study. And, in fact, this suggestion is supported by other studies. For example, the recent UK National Mastectomy and Breast Reconstruction Outcomes Audit [58] reported an 84% response rate (n=6882) for the BREAST-Q ©. But while it is unlikely that the lower response rate achieved in our study reflects problems with patient acceptability of the BREAST-Q ©, which in fact exceeds 60% (the average response rate [59] and proposed minimum [60] for clinical research), further work is required to establish specific response rates.

Second, further psychometric examination remains to be done, including: comparisons with objective clinical data; comparisons of test-retest administration techniques (combined versus individual mailings);

group- and individual-patient level clinical change (responsiveness); and formal cross-cultural validations of translated versions of the BREAST-Q ©.

Finally, developing the BREAST-Q © using RMMs means that we can legitimately continue to refine and improve the measurement performance of its subscales, while the current version is being used [13]. In addition, the scores generated from future versions of the BREAST-Q© will be directly comparable to the present version to retain continuity. Thus, further work will include further development of the construct theories underpinning each subscale, examination of differential item functioning, residual correlations, and the potential of including new or modified items to improve upon and/or expand the measurement continuum of each subscale, if and where necessary.

ACKNOWLEDGEMENTS

This study was funded by grants from the Plastic Surgery Educational Foundation. The authors thank Dr Barry Sheridan, RUMMLab, Australia and Professor David Andrich, University of Western Australia, for design, analysis and programming support in the development Q-Score. We would also like to thank Dr William Fisher for his advice and input in relation to this paper. Finally, we would like to thank the following researchers and surgeons for their invaluable assistance with research support and with the recruitment of patients and countless hours spent as expert reviewers: Vancouver, B.C., Canada: Drs. Patricia Clugston, Peter Lennox, Nicholas Carr, Nancy Van Laeken, and Robert Thompson; Hamilton, O.N., Canada: Dr. Jennifer Klok Dartmouth, N.H.: Drs. Carolyn Kerrigan and E. Dale Collins; London, U.K.: Mr. Ash Mosahebi and Mr. James Frame; New York, N.Y.: Miss Amie Scott Drs. Colleen McCarthy, Peter Cordeiro, Babak Mehrara, Joseph Disa, and David Hidalgo; and Ann Arbor, Mich.: Drs. Amy Alderman and Edwin Wilkins.

The BREAST-Q © is provided free of charge for unfunded academic research and individual clinical practice. There is a small charge for use in grant funded academic research. The scoring software, Q-Score ©, is also offered free to charge to all BREAST-Q © users. The BREAST-Q © is available at www.BREAST-Q.org and is jointly owned by Memorial Sloan-Kettering Cancer Center and the University of British Columbia. Drs. Cano, Klassen, and Pusic are co-developers of the BREAST-Q and, as such, receive a share of any license revenues based on the inventor sharing policies of these two institutions.

REFERENCES

[1] S. Cano, J. Browne, and D. Lamping, "Patient-based measures of outcome in plastic surgery: Current approaches and future directions" *British Journal of Plastic Surgery*, 57, 1-11, 2004.

[2] Food and Drug Administration, *Patient reported outcome measures: use in medical product development to support labelling claims.* Food and Drug Administration, Silver Springs, MD 2009.

[3] UK Department of Health, *Equity and excellence: Liberating the NHS.* 2010, Her Majesty's Stationery Office: London.

[4] R. Fitzpatrick, et al., "Methods of assessing health-related quality of life and outcome for plastic surgery" *British Journal of Plastic Surgery*, 52, 251-255, 1999.

[5] A. Pusic, et al., Clinical research in breast surgery: reduction and postmastectomy reconstruction. *Clinics in Plastic Surgery*, 35, 215-226, 2008.

[6] S. Cano, et al., "Health Outcome and Economic Measurement in Breast Cancer Surgery: Challenges and Opportunities" *Expert Review of Pharmacoeconomics & Outcomes Research*, 10 583-594, 2011

[7] Institute of Medicine and National Research Council, *Crossing the Quality Chasm: The IOM Health Care Quality Initiative.* The National Academies Press, Washington, DC, 2001.

[8] A. Pusic, et al., "Measuring Quality of Life in Cosmetic and Reconstructive Breast Surgery: A Systematic Review of Patient-Reported Outcomes Instruments" *Plastic and Reconstructive Surgery*, 120, 823-837, 2007.

[9] Scientific Advisory Committee of the Medical Outcomes Trust, "Assessing health status and quality of life instruments: attributes and review criteria" *Quality of life Research*, 11, 193-205, 2002.

[10] D. Revicki, "FDA draft guidance and health-outcomes research" *Lancet*, 2007, 369, 540-542, 2007.

[11] A. Klassen, et al., "Satisfaction and quality of life in women who undergo breast surgery: A qualitative study" *BMC Women's Health*, 2009. 9, 11-18, 2009.

[12] A. Pusic, et al., "Development of a New Patient Reported Outcome Measure for Breast Surgery: The BREAST-Q" *Plastic and Reconstructive Surgery*, 124, 345-353, 2009.

[13] D. Andrich, "Controversy and the Rasch model: a characteristic of incompatible paradigms?" *Medical Care*, 42, I7-I16, 2004.

[14] M. Wilson, *Constructing measures: an item response modelling approach*. Lawreence Erlbaum Associates, Mahwah, NJ, 2005.

[15] D. Andrich, J. de Jong, B. Sheridan, "Diagnostic opportunities with the Rasch model for ordered response categories" In: J. Rost and R. Langeheine (eds.) *Applications of latent trait and latent class models in the social sciences*, Waxmann Verlag, Munster 59-70, 1997.

[16] W. Fisher and A. Stenner, "Integrating qualitative and quantitative research approaches via the phenomenological method" *International Journal of Multiple Research Approaches*, 5, 89-103, 2011.

[17] A. Stenner, et al., "How accurate are lexile text measures?" *Journal of Applied Measurement*, 7, 307-322, 2006.

[18] A. Stenner and M. Smith, "Testing Construct theories" *Perceptual and Motor Skills*, 55, 415-46, 1982.

[19] A. Stenner, M. Smith, and D. Burdick, "Towards a theory of construct definition" *Journal of Educational Measurement*, 20, 305-316, 1983.

[20] S. Cano, S. and J. Hobart, "The problem with health measurement" *Patient Preference and Adherence*, 5, 279-290, 2011.

[21] C. Spearman, "The proof and measurement of association between two things" *American Journal of Psychology*, 15, 72-101, 1904.
[22] M. Novick, "The axioms and principal results of classical test theory" *Journal of Mathematical Psychology*, 3, 1966.
[23] S. Stevens, "On the theory of scales of measurement" *Science*, 103, 677-680, 1946.
[24] J. Michell, "Measurement scales and statistics: A clash of paradigms" *Psychological Bulletin*, 100, 398-407, 1986.
[25] J. Hobart and S. Cano, "Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods" *Monograph for the UK Health Technology Assessment Programme*, 13 1-200, 2009.
[26] B. Wright and J. Linacre, "Observations are always ordinal: measurements, however must be interval" *Archives of Physical Medicine and Rehabilitation*, 70, 857-860, 1989.
[27] S. Embretson and S. Hershberger, eds. *The new rules of measurement.* Lawrence Erlbaum Associates: Mahwah, NJ, 1999.
[28] S. Cano, et al., "The ADAS-cog in Alzheimer's Disease clinical trials: Psychometric evaluation of the sum and its parts" *Journal of Neurology Neurosurgery and Psychiatry*, 81, 1363-1368, 2010.
[29] C. McHorney and A. Tarlov, "Individual-patient monitoring in clinical practice: are available health status surveys adequate?" *Quality of Life Research*, 4, 293-307, 1995.
[30] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Education Research, Copenhagen, 1960.
[31] C. McHorney, S. Haley, and J. Ware, "Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. comparison of relative precision using Likert and Rasch scoring methods." *Journal of Clinical Epidemiology*, 50, 451-461, 1997.
[32] L. Prieto, J. Alonso, and R. Lamarca, "Classical test theory versus Rasch analysis for quality of life questionnaire reduction" *Health and Quality of Life Outcomes*, 1, 27, 2003.
[33] J. Whitaker, et al., "Outcomes assessment in multiple sclerosis trials: a critical analysis" *Multiple Sclerosis*, 1, 37-47, 1995.
[34] T. Platz, et al., "Clinical scales for the assessment of spasticity, associated phenomena, and function: a systematic review of the literature" *Disability and Rehabilitation*, 27, 7-18, 2005.
[35] B. Wright and G. Masters, *Rating scale analysis: Rasch measurement*, MESA, Chicago, 1982.
[36] J. Linacre, "Computer-adaptive testing: a methodology whose time has come" In: S. Chae, et al., (Eds) *Development of computerised middle school achievement tests*, Komesa Press, Seoul, 2000.
[37] B. Wright, "Solving measurement problems with the Rasch model" *Journal of Educational Measurement*, 14, 97-116, 1977.
[38] D. Andrich, *Rasch models for measurement*. Sage, 1988, Beverley Hills, CA.
[39] F. Lord and M. Novick, "Statistical theories of mental test scores. Behavioural science: quantitative methods" Addison-Wesley, Reading, MA,1968.
[40] J. Lumsden, "Person reliability" *Applied Psychological Measurement*, 1, 477-482, 1977.
[41] M. Waller, *Estimating parameters in the Rasch model: removing the effects of random guessing*, Educational Testing service: Princeton, NJ, 1976.
[42] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability" In: F. Lord (Ed), *Statistical Thoeries of mental test scores*, Addison-Wesley: Reading, MA, 1968.
[43] D. Thissen and L. Steinberg, "A taxonomy of item response models" *Psychometrika*, 51, 567-577, 1986.
[44] R. Massof, "The measurement of vision disability" *Optometry and Vision Science*, 79, 516-552, 2002.
[45] J. Hobart, et al., "Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations" *Lancet Neurology*, 6, 1094-105, 2007.
[46] D. Andrich, "A rating formulation for ordered response categories" *Psychometrika*, 43, 561-573, 1978.
[47] D. Andrich and B. Sheridan, *RUMM 2030*, RUMM Laboratory Pty Ltd: Perth, WA, 1997-2011.
[48] D. Andrich, "An index of person separation in latent trait theory, the traditional KR20 index, and the Guttman scale response pattern" *Education Research Perspectives*, 9, 95-104, 1982.
[49] D. Andrich, "An elaboration of Guttman scaling with Rasch models for measurement" In: N. Tuma (Ed) *Social Methodology*, Jossey-Bass: San Francisco, CA, 1985.
[50] R. Flesch, "A new readability yardstick" *Journal of Applied Psychology*, 32, 221–233, 1948.
[51] D. Dillman, *Mail and telephone surveys: the total design method*, Wiley, New York, NY, 1978.
[52] D. Dillman, J. Smyth, and L. Christian, *Internet, Mail, and Mixed-mode Surveys: The Tailored Design Method*, John Wiley & Sons, NJ, 2009.
[53] R. Massof, "Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires" *Ophthalmic Epidemiologgy*, 18, 1-19, 2011.
[54] C. Hagquist and D. Andrich, "Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling" *Personality and Individual Differences*, 36, 955-968, 2004.
[55] I Marais and D. Andrich, "Formalising dimension and response violations of local independence in the unidimensional Rasch model" *Journal of Applied Measurement*, 9, 200-215, 2008.
[56] L Cronbach, "Coefficient alpha and the internal structure of tests" *Psychometrika*, 16, 297-334, 1951.
[57] C. McCarthy, et al., "Patient satisfaction with postmastectomy breast reconstruction: A comparison of saline and silicone implants" *Cancer*, 2010. 116, 5584-5591, 2010.
[58] UK NHS Information Centre, National Mastectomy and Breast Reconstruction Outcomes Audit. NHS Information Centre, Leeds 2011.
[59] D. Asch, M. Jedrziewski, and N. Christakis, "Response rates to mail surveys published in medical journals" Journal of Clinical Epidemiology, 50, 1129-36, 1997.
[60] F. Badger and J. Werrett, "Room for improvement? Reporting response rates and recruitment in nursing research in the past decade" *Journal of Advanced Nursing*, 51, 502–510, 2005.

Author(s):

Stefan J Cano, Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Room N13 (ITTC Building 1), Tamar Science Park, Davy Road, Plymouth, PL6 8BX, UK Tel: +44 1752 315245; Fax: +44 1752 315254; e-mail: stefan.cano@pcmd.ac.uk

Anne F Klassen, Department of Pediatrics, McMaster University, IAHS Building, Room 408D, 1400 Main Street West, Hamilton, ON L8S 1C7, Canada, e-mail: aklass@mcmaster.ca

Andrea L Pusic, Plastic and Reconstructive Surgery, Memorial Sloan-Kettering Cancer Center, Room MRI-1007, 1275 York Avenue, New York, NY 10065, USA, e-mail: pusica@mskcc.org