1040-5488/07/8408-0663/0 VOL. 84, NO. 8, PP. 663–674
OPTOMETRY AND VISION SCIENCE

FEATURE ARTICLE ON LINE

The Development, Assessment, and Selection of Questionnaires

KONRAD PESUDOVS, PhD, FAAO, JENNIFER M. BURR, MRCOphth, MSc(Epidemiology), CLARE HARLEY, PhD, and DAVID B. ELLIOTT, PhD, FAAO

NH&MRC Centre for Clinical Eye Research, Department of Ophthalmology, Flinders Medical Centre and Flinders University of South Australia, Bedford Park, South Australia, Australia (KP), Health Services Research Unit, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, United Kingdom (JMB), and Department of Optometry, University of Bradford, Richmond Road, Bradford, West Yorkshire, United Kingdom (CH, DBE)

ABSTRACT

Patient-reported outcome measurement has become accepted as an important component of comprehensive outcomes research. Researchers wishing to use a patient-reported measure must either develop their own questionnaire (called an instrument in the research literature) or choose from the myriad of instruments previously reported. This article summarizes how previously developed instruments are best assessed using a systematic process and we propose a system of quality assessment so that clinicians and researchers can determine whether there exists an appropriately developed and validated instrument that matches their particular needs. These quality assessment criteria may also be useful to guide new instrument development and refinement. We welcome debate over the appropriateness of these criteria as this will lead to the evolution of better quality assessment criteria and in turn better assessment of patient-reported outcomes. (Optom Vis Sci 2007;84:663–674)

Key Words: factor analysis, instrument, quality assessment, quality of life, questionnaire, Rasch analysis, reliability, responsiveness, validity, visual disability

The assessment of health-related quality of life (HR-QoL) has been an important expansion of the assessment of the impact of disease and its treatment beyond the traditional areas of symptoms, signs, morbidity, and mortality. It provides a more holistic assessment of the effects of disease on the person to include such dimensions as a patient's physical, social, and emotional wellbeing. Most funding organizations now insist on a patient-reported outcome for a clinical trial of any disease intervention/treatment or assistive device. Because of the breadth of content of HR-QoL and its patient-reported nature, it has been measured using questionnaires (called instruments in the research literature), which are efficient tools for gathering large amounts of data quickly. Given the large number of instruments that have been developed over the last few years, investigators may find it difficult to decide upon an appropriate instrument or decide whether a questionnaire needs to be specially developed for their study. Another problem is that originally (1920s to 1950s), the primary purpose of instrument development was to determine people's attitudes and how that range of attitudes was distributed in the population rather than to produce a score on a quasi continuous variable.[1] As the use of instruments extended beyond psychology, to medical fields, the format, and purpose of instruments changed. Unfortunately, the change in the design and application of instruments also meant that traditional methods of scoring and validation became outdated, but this was not recognized for many of the originally developed instruments.

Throughout this article, we highlight key quality criteria (summarized in Table 1) that build upon previous contributions to the field.[2,3] It is our aim to present a robust set of quality criteria to be used by researchers and practitioners in the selection of instruments, and we welcome comments and suggestions for their refinement or further development. These proposed quality assessment criteria (Table 1) provide a framework for a systematic review of instruments in the disease area under study to determine if any existing instruments are adequate for the intended use in the intended target population. In the absence of a sufficiently reliable, valid instrument with content appropriate to the intended use, the development of a new instrument for the intended purpose can be justified. The qual-

TABLE 1.
Quality assessment tool for evaluation of health status questionnaires

Property	Definition	Quality criteria[a]
Development of the instrument		
Prestudy hypothesis	The prestudy specification of the aim of the instrument and the intended population	√√ A clear description is provided of the aim of the instrument and the intended population √ Only one of the above x Neither reported
Intended population	The extent to which the instrument has been studied in the intended population	√√ Intended population studied √ Partly studied only or sample size was small (less than 50 patients) x Not studied in the intended population, only generic
Actual content area	The extent to which the content meets the prestudy hypothesis specifications	√√ Content as intended, and is relevant to the intended population √ Some of the intended content areas missing x Content area not relevant to intended population
Item identification	Selection of the items relevant to the target population for inclusion in the pilot instrument	√√ Comprehensive consulting with patients, (focus groups or in-depth interviews) and a literature review √ Minimal consultation with patients and experts opinion and literature review x No consultation with patients
Item selection	Determining the items included in the final instrument	√√ A pilot instrument was developed and tested with Rasch or factor analysis and statistical justification provided for removing items, plus items with floor and ceiling effects removed and the amount of missing data considered √ Only some of above techniques were used x No pilot instrument OR no statistical justification of items included in the final instrument
Unidimensionality	Demonstration that all items fit with a single underlying construct	√√ Rasch analysis using fit statistics (0.7–1.3) or item-trait interaction or Factor analysis on Rasch scores (1st factor loadings >0.4 for all items) √ Rasch fit statistics mostly within 0.7 to 1.3 range but some less well fitting items retained, or Cronbach's α >0.7, and <0.9 or factor analysis on raw scores (1st factor loadings >0.4 for all items) √ Rasch analysis does not support unidimensionality or Factor analysis does not support unidimensionality or Cronbach's α <0.7 or >0.9
Response scale	Categories used to rate the items	√√ Statistically justified scale without significant missing data, floor and ceiling effects, and a demonstration of ordered thresholds on Rasch analysis √ Some, but not all of above x Methods for determining response scale not justified statistically
Scoring	A description of how the instrument should be scored	√√ Rasch scoring of a statistically justified response scale √ Summary scoring of a statistically justified response scale x Scoring system not described or scoring of a statistically unjustified or faulty scale
Performance of instrument (validity and reliability)		
Validity		
Convergent validity	Amount of correlation with a related measure	√√ Tested against appropriate measure, correlates between 0.3 and 0.9 √ Debatable choice of measure, but correlation between 0.3 and 0.9 x Tested and correlates <0.3 or >0.9
Discriminant validity	The degree to which an instrument it not similar to (diverges from) other instruments that it should not be similar to	√√ Tested against appropriate measure, correlates <0.3 √ Debatable choice of measure, but correlation between <0.3 x Tested and correlates >0.3
Predictive validity	The extent to which the instrument can predict a future event	√√ Tested against appropriate measure, correlates >0.3, or significant difference between groups √ Debatable choice of measure, but correlation >0.3 or significant difference between groups x Tested and correlates <0.3 or nonsignificant difference between groups
Other evidence for construct validity	Any other hypothesis driven testing	√√ Hypothesis stated, tested and proven √ Construct validity claimed but debatable under scrutiny x Construct validity claimed but does not hold up to scrutiny
Test-retest (T-R) agreement	The extent to which the results are repeatable when taken by the same observer	√√ LOA appear tight and less than MID, or weighted Kappa or ICC ≥0.8 (T-R) or 0.70 (int.)
Interobserver agreement/intermode (int.) agreement	The extent to which the results are repeatable between observers/ The extent to which the results are repeatable between modes of administration	√ LOA broader but still close to MID, or weighted Kappa or ICC 0.60 to 0.79 (T-R) or 0.50 to 0.69 (int.) x LOA ≫ MID, weighted Kappa or ICC <0.60 (T-R) or 0.50 (int.) or incorrect statistical test or inadequate sample (<30 subjects),
Person and item separation reliability	A Rasch analysis indication of reliability—the proportion of true variance in the observed variance	√√ Reliability of ≥0.8 for both person and item separation or a G value or separation ratio >2 √ Only one of person or item separation of ≥0.8, or a G value or separation ratio >2 x Person or item separation of <0.8, or a G value or separation ratio <2 0 Not reported (not a Rasch scaled measure)
Interpretation	The extent to which score differences are meaningful	√√ Normative data (Mean scores and SD) and MID given for a representative target population, and test population demographic reported √ MID or normative data or demographic details of study populations, or ad hoc population x No normative data and no MID
Responsiveness	The extent to which the instrument can detect clinically important changes over time	√√ Score changes >MID for measures of progression over time or changes with intervention. Effect size or responsiveness statistic given √ Changes over time but relationship to MID not reported, small sample, and inadequate time frame x Score changes ≤MID

[a]If not reported, scored as "0"; √√, Positive rating; √, Minimal acceptable rating; x, Negative rating.

MID, minimally important difference; LOA, limits of agreement; ICC, intraclass coefficient; SD, standard deviation.

ity assessment criteria provided in Table 1 can also be used to guide new instrument development and refinement.

Overview

The organizational structure of this manuscript follows that of listing of the Quality Assessment Criteria in Table 1. We start with issues involved in the development of an instrument. These include defining the purpose of the instrument and its target population; the steps taken in defining the content of the instrument; and the steps involved in developing the rating scale and scoring system. The second half of the manuscript deals with the performance of an instrument. This includes the different types of validity, and reliability as well as responsiveness and interpretation of the results.

By way of example, the quality criteria assessment from Table 1 has been applied to four refractive error-related QoL instruments in Table 2. The Psychosocial Impact of Assistive Devices,[4–7] the Refractive Status Vision Profile (RSVP),[8–12] the National Eye Institute Refractive Quality of Life[12–16] and the Quality of Life Impact of Refractive Correction[17–19] were assessed on the premise they are to be used in a study comparing QoL among spectacle and contact lens wearers. All articles (three to four per instrument) contributing to the description, development, and validation of the instruments were included in the assessment.

Prestudy Hypothesis and Intended Population

Studies describing the development of an instrument should clearly state the specific construct the instrument is intended to measure and the intended population of study. If the instrument was not developed on a comparable population to your target population then relevant content is unlikely to have been included. For example the Impact of Visual Impairment questionnaire which was developed and validated using a low vision population[20] was shown to perform poorly in a clinical glaucoma population with respect to targeting of item difficulty to patient ability.[21] The same is true if important population subsets were omitted, because the breadth of population and extent of generalizability are important: for example, an instrument for assessing quality of life in the different modes of correction of refractive error,[17] should include items relevant to all modes of refractive correction (e.g., spectacles, contact lenses, and refractive surgery) to ensure the content is relevant to all subtypes. This is the basis of the results in Table 2 where the RSVP instrument only scores one tick for intended population as it was developed primarily of refractive surgery candidates (therefore "partly studied only" as per Table 1), whereas the other instruments scored two ticks (intended population studied).

Representation, Face, and Content Validity

Representation (or translation)[22] validity is an over-arching term relating to how well the construct under measurement is represented by an instrument.[23] This term exists to draw together face validity and content validity which both address the content of the instrument but in different ways.

Face validity is the basic idea that an instrument should appear to measure what it purports to measure and this can be assessed by inspection.[22] However, face validity is probably best considered to be the weakest form of validity. Demonstrating that an instrument has face validity is probably best confined to consideration of whether the concept seemingly being measured, and the rating scale used etc [e.g., is frequency (or difficulty) the right concept], appears to be sensible. The items should be phrased in simple, unambiguous language, kept brief, clear, and avoid over intellectualization, multiple concepts, and double negatives. As a guide, items should be written at a comprehension level suitable for a 12 years old.[24]

Unfortunately, face validity may be misused. The purpose of drawing together a group of items is to measure a latent trait represented by those items. If the instrument has face validity, then it should appear to measure what it intends to measure. In depth analysis of the items included, or objections to missing items, is not appropriate for face validity. After all, it is likely that various collections of items could measure the same underlying construct. Therefore heavy emphasis on the inclusion or exclusion of specific items is not appropriate for face validity.

Actual Content Area

The actual content area quality assessment criterion addresses face validity; the extent to which the content meets the intended concepts specified in the prestudy hypothesis. The assessment is somewhat subjective, but can be assisted by clear definitions from the developers of the instrument as to what the framework of the content is. Clarification of the content areas is especially important for instruments which measure multiple traits. The actual content area violates the intended content area when aspects of the intended content area are missing or content not relevant to the intended content are included.

Content validity is the extent to which the items in the instrument reflect the entirety of the concept being measured. Content validity cannot be formally assessed because it is difficult to prove conclusively that the items chosen were representative of all possible items.[25] However, instrument development methods such as item identification and item reduction are important for establishment of content validity (see quality criteria in Table 1).[26]

Breadth of content is critical to content validity. Many instruments purport to measure quality of life, but only measure a few dimensions; often vision-related activity limitation only (visual functioning or visual disability would be more appropriately called vision-related activity limitation to be in line with the World Health Organization International Classification of Functioning, Disability and Health[27]). However, QoL has many other dimensions e.g., emotional, spiritual, vocational, economical attributes etc. So to purport to measure QoL but to only or principally measure activity limitation means that any inferences one may draw about QoL impacts will be incorrect unless they are confined to activity limitation only. This problem is called construct under representation,[28] and is common in vision-related instruments including the popular NEI-VFQ.[29] So the name of an instrument is actually very important as it feeds into defining the concept that the instrument purports to measure. The title of the VF-14 (Visual Function Index 14) instrument and the research article that introduced it quite clearly indicates that it measures activity limitation

TABLE 2.
Quality assessment of 4 refractive error-related quality of life instruments: Psychosocial Impact of Assistive Devices (PIADS),[4–7] the Refractive Status Vision Profile (RSVP),[8–13] the National Eye Institute Refractive Quality of Life (NEI-RQL)[12,14–17] and the Quality of Life Impact of Refractive Correction (QIRC)[18–20]

	Hypothesis	Intended population	Actual content area	Item identification	Item reduction	Unidimensionality	Response scale	Scoring scale
PIADS	√√	√√	√√	√	√	x	√	√
RSVP[a]	√√	√	√√	√√	√	x	x	x
NEI-RQL	√√	√√	√√	√√	x	x	x	x
QIRC	√√	√√	√√	√√	√√	√√	√√	√√

[a]A Rasch-analyzed version of the RSVP (Garamendi et al., 2006) with a modified response scale and a reduced number of items has been shown to have greater responsiveness and test-retest reliability than the standard instrument. It also provides a unidimensional score, statistically justified response and scoring scales and good Rasch separation reliability.
[b]Conflicting reports of normative data levels and responsiveness of the RSVP are provided by Schein et al. (2001) and Nichols et al. (2001).

and does not claim to measure quality of life but it has been misinterpreted as assessing QoL.[30–32]

Item Identification

To ensure a good breadth of relevance, at least three approaches should have been taken for item generation. These include obtaining sample statements, experiences, and opinions directly from: individuals within the target population, through focus groups or one-to-one interviews; experts working in the area (not just clinicians, but individuals who have contact with patients and may develop expertise in understanding the impact of the condition on the person); and the published literature in the field. Patient interviews are useful for gathering a range of opinions on a topic and can help to draw views from particular minority groups. Focus groups are useful for eliciting mediated responses that are likely to be common to the majority of individuals in a given population and can also be more productive than in-depth patient interviews due to the enthusiasm and interaction created by the discussion process.[33,34] Expert knowledge is a valuable resource, but should not be used as the sole procedure for generating items because clinicians tend to focus on presenting complaints. There may also be issues that the patient does not present to a clinician, but which have an impact on their quality of life. For example, the RSVP is a clinician-developed instrument of QoL for refractive surgery and has been shown to include too many items related to symptoms and functional problems,[11] whereas patients are more concerned about issues such as convenience, cost, health concerns, and well being.[17]

Pilot Questionnaire

Item generation will typically produce a vast number of items. An item removal process is required to determine which items to retain for the final instrument. A pilot questionnaire is best used for this process (see quality criteria in Table 1). The pilot questionnaire indicates how well each item taps the underlying construct being measured, and allows poorly discriminating, unreliable or invalid items to be removed. The respondent population for the pilot data should have been broad and representative of the target population.

Unidimensionality and Item Reduction

Item reduction is performed to maximize item quality, measurement precision, and targeting of items to persons. Unidimensionality is the demonstration that all items included in an instrument fit with a single underlying construct (e.g., VR-QoL) and is a prerequisite to allow appropriate summation of any set of items[24,35] and an important asset if a meaningful measurement is to be obtained.[35,36] A number of statistical methodologies are used to justify item reduction and give insight into dimensionality:

- Conventional descriptive statistics
- Cronbach's alpha
- Factor analysis
- Rasch analysis

Statistical methods for item reduction serve to highlight the worst performing items, which are removed. The items are removed one at a time with the analyses performed iteratively to calculate the improvement in the instrument and to identify the next candidate item for removal. Traditionally, the following descriptive and statistical analyses have been used to determine candidate items for removal.[4,5,26]

- Missing data. Items that have large percentages (>50%) of missing data are likely to be ambiguous, or not applicable to many respondents.
- All items should approximate a normal distribution, as identified using histogram plots, nonsignificant results on tests of normality such as Kolmogorov-Smirnov or Shapiro-Wilk, or Skewness and Kurtosis values within −2.00 to +2.00. Although items at the ends of the scale will likely deviate from normal.[37]

Unidimensionality of the whole instrument must be considered when deciding which items to remove. Traditionally, Cronbach's alpha and factor analysis were used to assess unidimensionality. Cronbach's alpha determines the correlation of every item in the instrument with every other item. The nearer Cronbach's alpha to 1, the more internally consistent the scale is Cronbach's alpha can also be calculated on the items with each item in turn deleted from the analysis. If alpha increases (relative to the alpha of all items included) this indicates that the item removed was not contribut-

TABLE 2.
Continued

Discriminant validity	Convergent validity	"Other" validity	Test-retest reliability	Interobserver or intermode agreement	Rasch separation reliability	Interpretation	Responsiveness
0	√√	√√	√√	√√	0	√√	√√
0	√	√	√	0	0	√[b]	√[b]
0	√√	√√	√√	0	0	√√	√√
0	0	√√	√√	0	√√	√√	√√

ing to unidimensionality. Because Cronbach's alpha is essentially determined by the average of the correlation coefficients between items, exceptionally high values of Cronbach's alpha (>0.90) may be indicative of redundancy (e.g., in the RSVP, see Table 2). Although this does not contravene unidimensionality, redundancy is a problem if the process of creating the "overall score" for the instrument involves just adding all the item scores together. In such a case, the overall score overweighs the importance of the issue that is served by redundant items. Therefore, in our quality criteria, we downgrade those instruments with Cronbach's alpha >0.90 (Table 1). Similarly, as Cronbach's alpha is not independent of the number of items and may be elevated by including many items. For these reasons Cronbach's alpha should probably be considered to be more of a traditional indicator than a useful one.[38] Nevertheless we retain it in our quality criteria as it is such a commonly reported metric: values should be >0.70 and <0.90.

Factor analysis is a multivariate statistical method that is used to explain variance within a matrix of variables and reduces those variables to a number of factors. This method can be used to determine whether an instrument possesses unidimensionality.[26] The proportion of the variance described by the principal (most significant) factor indicates whether the instrument tests in one or more content areas. In addition, factor analysis can be "rotated" by various techniques such as Varimax or Oblimin to find items which can have high communality and thus form additional factors. This grouping of items into additional factors can be used to justify the creation of subscale indices as items that load onto the same factor are likely to sample the same content area specified by the factor to which they contribute. Subscales should be proposed hypothetically and justified with confirmatory factor analysis rather than simply being the product of exploratory factor analysis.[39] Once demonstrated to exist by factor analysis, subscales themselves should also be assessed for unidimensionality. Factor analysis can guide item reduction by indicating both failure to fit (items contributing to <0.40 of a particular factor) and redundancy (>0.80). Ideally, factor analysis should be performed on Rasch-scaled data, so that items do not group simply because of similar item difficulty.[40]

More recently developed instruments have used Rasch analysis to help guide item reduction.[41–45] Rasch analysis provides a more detailed view of dimensionality through both model and item fit statistics.[38] The item-trait interaction score, reported as a χ^2, reflects the property of invariance across the trait. A statistically non significant probability value ($p > 0.05$) indicates no substantial deviation from the model which implies unidimensionality.[20] The infit and outfit statistics also help to identify which items contribute most to the measurement of the latent trait. Infit and outfit means squares have an expected value of 1.00. Infit means (<0.8) represent items are too predictable (they have at least 20% less variation than expected). These overfitting items may be redundant or lack variance to contribute new information to the measure. Mean outfit values >1.20 represent misfit (at least 20% more variance than was expected) and suggests that the item measures something different to the overall scale. Acceptable values for item inclusion may be 0.80 to 1.20 for a strict definition (often used for infit) or 0.70 to 1.30 or even higher for lenient definition. Alternatively, fit residuals may be used, in which case values >2.5 or probability values below the Bonferroni adjusted alpha value (i.e., 0.05/number of items) are also used to indicate misfit to the model. Rasch analysis can also indicate the effect of removing an item on overall scale performance. If removal of an item considerably decreases person separation that item should be retained.[36] Person separation is an indicator of the ability (precision) of the instrument to differentiate between different people's quality of life. Person separation is expressed as the ratio of the adjusted standard deviation to the root mean square error and a person separation value of 2.0 or more is indicative that subjects are significantly different in QoL across the measurement distribution.[46]

Targeting of Items to Persons

Rasch analysis also provides insight into targeting of item difficulty to person ability and can therefore be used to remove items that less well target the population.[47] Figs. 1 and 2 show person-item maps for a group of cataract patients responding to the Activities of Daily Vision Scale (ADVS),[48] a visual activity limitation instrument. This analysis rank orders the items and participant responses. The means of the two distributions (person and item) are denoted as 'M'. If the items were well targeted to the subjects, the means of the two distributions should be close (e.g., 0.5 logits) to each other. In Fig. 1, the original conventionally validated ADVS is shown, and it can be seen that the means are far apart. Fig. 2 shows how item reduction of the ADVS, using Rasch analysis, provides better targeting of item difficulty to patient ability, with the 'M' values now closer together. This was achieved through

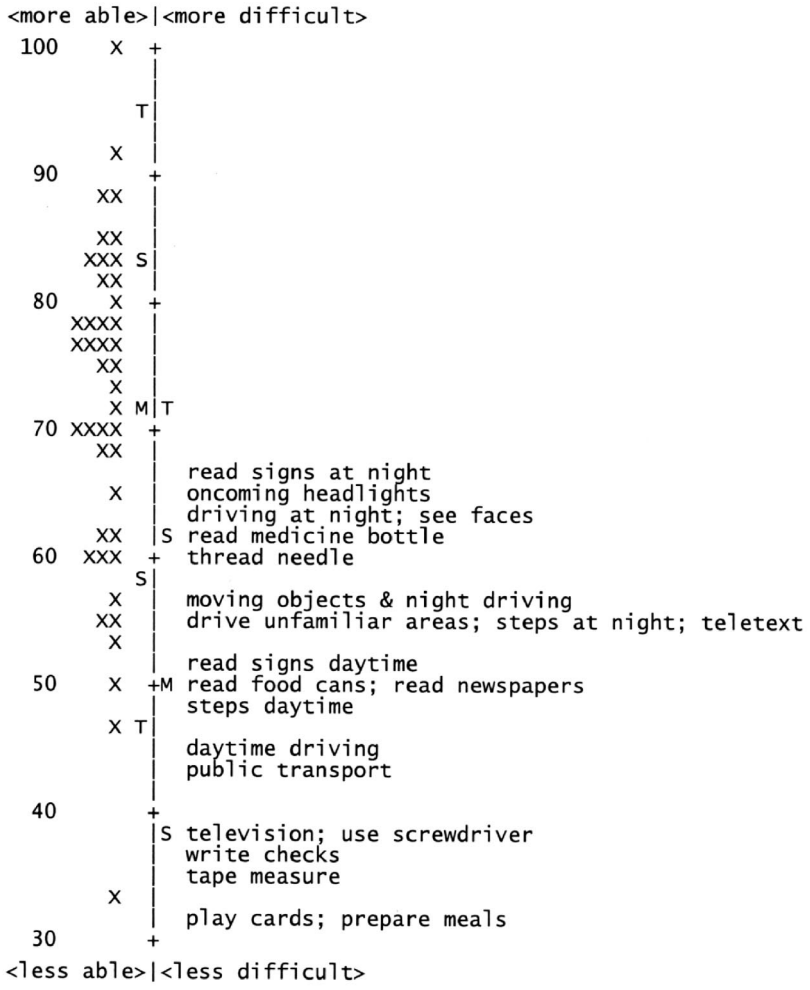

FIGURE 1.

Patient activity limitation/item difficulty map for the 22-item ADVS. On the left of the dashed line are the patients, represented by X. On the right are the cross-over points between each response scale (level of the scale where the answer category is most probable to be rated by a patient with that activity limitation). More able patients and more difficult items are near the bottom of the map; less able patients and less difficult items are near the top. The scale is in log units (0–100). M, mean; S, 1 SD from the mean; T, 2 SD from the mean.

removal of items that were too easy for patient ability. This item reduction approach can lead to a minimum item set, which has the optimum instrument efficiency and the advantage of shortening test time and reducing user and respondent burden.

Criteria to guide item removal that incorporate all of these statistical approaches have been proposed.[17,49] The suggested infit and outfit ranges are only guides and can depend largely on sample size.[50]

1. Infit mean square outside 0.7 to 1.30
2. Outfit mean square outside 0.70 to 1.30
3. Item with mean furthest from subject mean
4. High proportion of missing data (>50%)
5. Ceiling effect—a high proportion of responses in item end-response category (>50%)
6. Items with markedly different standard deviation of item scores to other items
7. Items that do not demonstrate a normal distribution, as identified using histogram plots, tests of normality such as Kolmogorov-Smirnov or Shapiro-Wilk, or Skewness and Kurtosis values outside −2.00 to +2.00.

Rating Scale

Unfortunately, many QoL instruments use traditional summary scoring where an overall score is derived through summative scoring of responses. Summary scoring is based on the hypotheses that all questions have equal importance and response categories are accordingly scaled to have equal value with uniform increments from category to category. In cases where the items in an instrument no longer have equal importance, the logic of averaging scores across all items becomes questionable. For example, in a summary scaled visual activity limitation instrument, the ADVS, "a little difficulty" scores 4, "extreme difficulty" is twice as bad and scores 2, and "unable to perform the activity due to vision" is again twice as bad with a score of 1. The ADVS ascribes the same response scale to a range of different items, such that "a little difficulty" "driving at night" receives the same numerical score as "a little difficulty" "driving during the day", despite the former being by far the more difficult and complex task. This rationale of "one size fits all" is flawed in this case, and Rasch analysis has been used to confirm

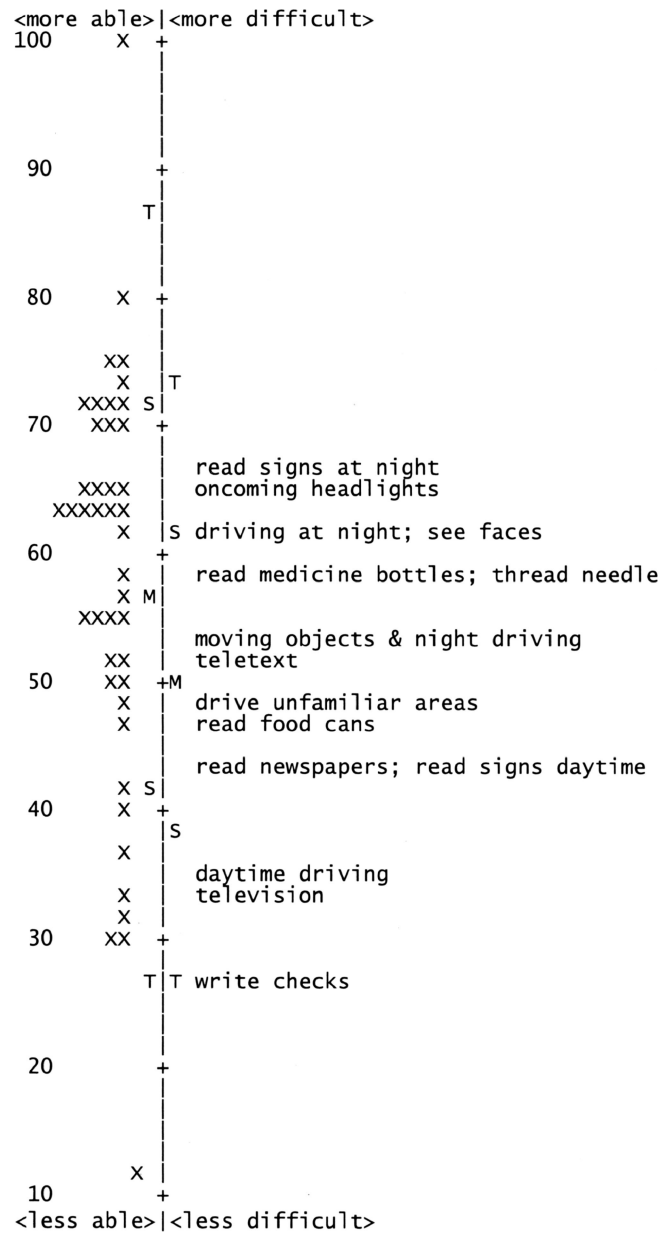

FIGURE 2.

Patient activity limitation/item difficulty map for the revised 15-item ADVS. The patient and item means are much closer together now that items that were too easy have been removed.

that differently calibrated response categories can help to provide a valid and contextual scale that truly represents QoL.[50]

By resolving inequities in a scale arising from differential item difficulty, Rasch analysis provides a self-evident benefit in terms of accuracy of scoring. This process also removes noise from the measurement which in turn improves sensitivity to change and correlations with other variables.[11,51] For example, the standard scoring of the Refractive Status and Vision Profile (RSVP) failed to show any difference in QoL between a group of spectacle and contact lenses wearers in optometric practice and a group of spectacle and contact lenses wearers about to undergo refractive surgery. When Rasch analysis was used to differentially calibrate each item, significant differences between the groups was found, with the prerefractive surgery group having a lower self-reported QoL than the control group, as might be expected.[11] This occurs through the reduction of noise in the original measurement which chiefly arises from considering all items to be of the same value. Note that conventionally developed instruments can also be reengineered using Rasch analysis[11,50,52] and it is possible to use the Rasch calibrations from these studies to convert summary-scaled data from these instruments.[20,53,54]

Rasch analysis provides the additional benefit that it can be used to determine the optimum number of response categories. It has been shown that people tend to only use four or five categories[55] and in some cases just three are used.[17] Using too many response options can also disrupt the expected order of categories. This disruption can be detected using Rasch analysis, which calibrates the responses for each category. If the analysis shows redundancy or disruption to category order, it may be necessary to combine adjacent response categories. Fig. 3 illustrates how Rasch analysis was applied to an instrument that determined the extent of pain from ocular surface disease.[56] The Faces Pain Scale originally used a seven-category response format (seven faces with different expressions of pain designed to be chosen to represent how the participants feels about their ocular pain) but Rasch analysis revealed that category 5 of the scale was underutilized and for no part of the scale was it the most likely to be selected; this category needed to be collapsed into an adjacent category. Rasch analysis determined that a 5-point scale would be more appropriate for this particular instrument. Visual analog scales are an extreme example of this problem. Users have the misconception that a 10-cm line scored by the millimeter results in a 101 category scale. However Rasch analysis shows that people tend to only use four or five categories.[55] When using Rasch analysis, investigation of category ordering and any repair of disordered thresholds should be undertaken before item reduction.

Response category design and function is also important when using the Rasch model. If all items have the same format and use the same categorical rating scale then a single Andrich rating scale can be used.[57] This means that all items use the same differences between response category values. If one prefers, one can use a partial credit model where response categories for all items are allowed to vary independently.[58] However, the use of a partial credit model introduces additional degrees of freedom and diminishes the value of item fit statistics as indicators of unidimensionality. For scales with several types of rating scales or question format, a different rating scale should be used for each type and the partial credit model is most appropriate.

Rasch analysis is also useful where there are missing data in patient or respondent answers, which is a common occurrence. With Rasch analysis, person estimates are made from valid data only, so missing data are effectively ignored, without adding noise to the measure. This is an important attribute of Rasch-scaled instruments as there are special cases where items with high rates of missing data may be important, such as driving in cataract populations.

Performance of the Instrument

Validity. *Construct validity* refers to whether an instrument measures the unobservable construct (such as "quality of life") that it purports to measure. Construct validity cannot be demonstrated by one simple test e.g., a correlation, because validation is an on-

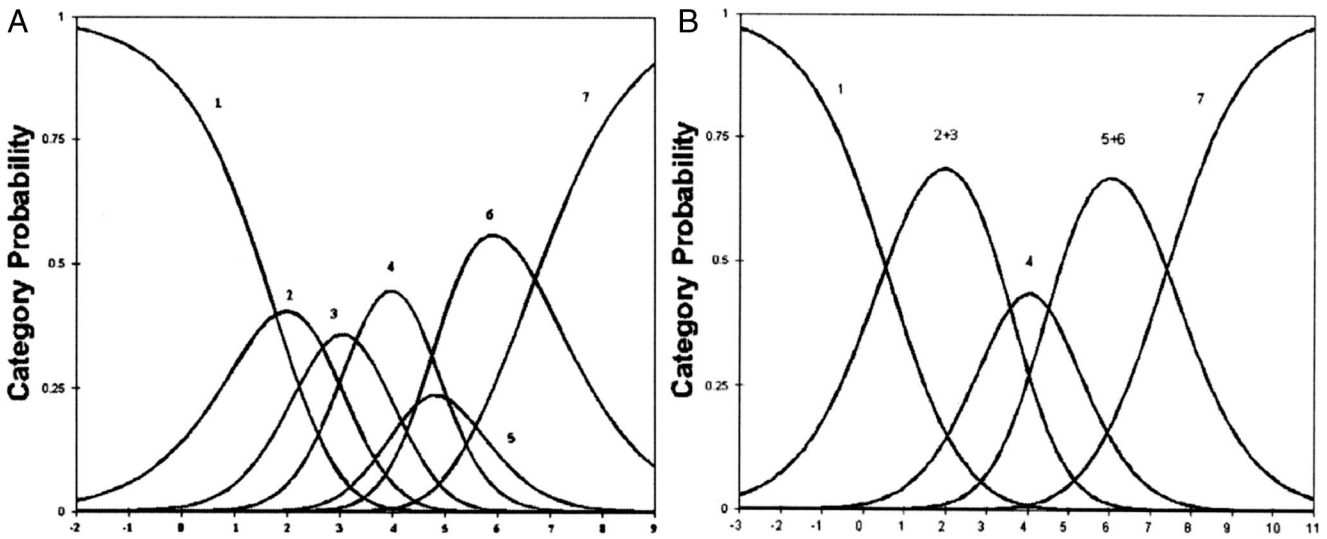

FIGURE 3.

(A) Rasch model category probability curves for the faces pain scale representing the likelihood that a subject with a particular pain severity will select a category. The x-axis represents pain. For any given point along this scale, the category most likely to be chosen by a subject is shown by the category curve with the highest probability. At no point is category 5 the most likely to be selected. This suggests there are too many categories and these are not used in order. (B) Rasch model category probability curves for the faces scale shortened to 5 categories by combining categories 2 and 3, and 5 and 6. This model gives excellent symmetry and the thresholds are now ordered. Both figures reproduced with permission from J Pain 2005;6:630–6.

going process requiring the statement of hypotheses and the testing thereof; if an instrument measures a trait, then it should correlate with another measure etc. There are specific types of validities that together contribute to construct validity e.g., concurrent, convergent, predictive, and discriminant validity. Although it is not possible to perform all of these tests, it is important that construct validity should be a hypothesis driven process. Sometimes the hypothesis will be simple and easily fall under the heading of e.g., convergent validity. Other times, complex hypothesis testing will not be readily subclassified but be critical to the establishment of construct validity. With the right set of hypotheses and tests, a persuasive picture of construct validity can be developed.

Criterion validity is a traditional definition of validity where an instrument is correlated with an existing "standard" or accepted measure which measures the same thing. However, criterion validity can be further subdivided so we use "criterion-related validity" as an umbrella term here.

Convergent validity is the classic form of criterion validity where a new instrument is correlated with something that measures a related construct. For visual activity limitation instruments, correlation with visual acuity (VA), or an existing validated visual activity limitation instrument (e.g., the VF-14[32]) is typically used to indicate convergent validity. Suitable statistical analyses are a Pearson correlation coefficient for continuous variables or, for dichotomous data, a chi squared analysis with a Phi coefficient as a measure of the correlation. Note that for convergent validity, a very high correlation (>0.90) is not advantageous as it suggest that the new instrument provides information so similar to a previously developed instrument or other measure that it provides no significant additional information. So, a moderate correlation may actually be better than a high one because it indicates that the two measures are related but the instrument is also providing different information. However, a low correlation implies that two measures which are hypothesized to be related are not very well related at all. A cutoff of 0.3 is probably appropriate as a minimum correlation between two measures which should be related. Therefore, the hypothesis is critical for convergent validity.

Discriminant validity is the degree to which an instrument diverges from other instruments that should be dissimilar. This is probably the validity test performed least often; no results in Table 2. For refractive-error related QoL instruments, it might be simple to show a poor correlation to an instrument designed for measuring visual activity limitation, because disability is not typically a component of the former. The statistical test required is again a simple Pearson correlation coefficient, but in this case, a poor correlation e.g., <0.3 is the desired result. More complex hypotheses of concurrent and discriminant validity could also be set. For example, a new cataract specific visual activity limitation instrument could be hypothesized to correlate very well with an existing cataract specific visual activity limitation instrument, less well with an ophthalmic QoL instrument and least well with a general health QoL instrument. Such a hypothesis can avoid the 0.3 cutoff, as the correlations may well be of the order of 0.7, 0.5, and 0.3, respectively, and therefore provide good criterion-related validity evidence for both convergent and discriminant validity.

Predictive validity determines whether the instrument can make accurate predictions of future outcomes. For example, can a score on a visual activity limitation instrument be used to predict the need for cataract surgery? This may be worthwhile because people could be prioritized for examination based on instrument scores and some people with minimal activity limitation could be spared a costly comprehensive eye examination. Again, a simple Pearson correlation coefficient (assuming a normal distribution, alternatively a Spearman rank correlation) is the appropriate test and a correlation of >0.3 is an appropriate cut-off, although for predictive validity a very high correlation is not disadvantageous. For a dichotomous outcome, a significant χ^2 or odds ratio would be appropriate.

Concurrent validity illustrates an instrument's ability to distinguish between groups that it should theoretically be able to distinguish.[22] Critically, both are measured at the same time, rather than one being measured at a future time. For example, an instrument designed for a particular condition should be able to discriminate between groups with and without a condition. Testing such a hypothesis is often the easiest contribution to construct validity. The instrument is administered to two groups, one with the condition, one without. For simplicity, equivocal cases are not included in the analysis. Although this provides weak evidence of validity because it may be the equivocal cases where the instrument may be most needed (assuming there is a needs-based reason for developing the new instrument). The results become more powerful when discriminating between two groups that are very similar.

Validity demonstrates that the instrument measures the construct that it was intended to measure, and relates well to other measures of the same or similar constructs. It does not, however, show that the construct is consistently captured across respondents, time, and, setting.

Reliability. *Reliability* is the consistency of the instrument in measuring the same construct over different administrations, but does not indicate validity, as it makes no assumption that the correct construct was measured in the first place. Reliability generally examines the proportion of the total variance that is attributable to true differences among subjects. Total variance includes the true differences and measurement error. That measurement error is considered to result from time, rater, and content selection.[26] Reliability is a very important quality of an instrument as unreliability detracts from validity. For example, if a test has poor reliability such that test results correlate poorly with retest results, it is unlikely that results from the test will correlate highly with gold standard measures, so that its concurrent and convergent validity will also be impaired by poor reliability.

The reliability of an instrument can be explored using many methods, which can be classified broadly into two categories: single administration and multiple administrations. Single administration methods include split half and internal consistency tests, for example Cronbach's alpha. These methods, however, are really examining 'internal consistency reliability', which indicates unidimensionality (as discussed above) rather than reliability. In particular, claims of very good instrument reliability based on very high Cronbach's alpha values (>0.90) can be downgraded as they are more indicative of redundancy in the instrument. It is important that Cronbach's alpha is not overemphasized as a measure of reliability and that the other attributes of reliability are reported. Multiple administration methods include test-retest, alternate forms (intermode), and interobserver reliability (not appropriate for self-administered instruments) and are typically calculated using the Pearson product-moment correlation coefficient (r), the intraclass correlation coefficient (ICC),[26,59] Bland-Altman limits of agreement,[60,61] or kappa statistics.

The *ICC* is defined as the ratio of the between-groups variance to the total variance. Thus it is a measure of agreement and it is valid to be used as such when there is no intrinsic ordering of two measures under comparison, e.g., in test-retest reliability.[62] The ICC, is dependent on the range of responses, so care must be taken with the population in question.[63]

The Bland-Altman limits of agreement (LoA) is the range of values over which 95% of the differences between two measures should lie.[60,61] This is a simple method to perform, and is applicable to many situations as long as the units of measurement (e.g., Diopters for refractive error etc) are the same (for reliability testing the units of measurement are essentially the same). The advantage of this approach is that it is robust to large data ranges and can detect and manage bias. Interpretation of whether a limit of agreement is a good or a bad result requires clinical context. Therefore, a disadvantage of this approach lies in interpretability if the scale of the instrument is unfamiliar. For an LoA result showing that the reliability of subjective refraction is ± 0.50 D, an optometrist or related clinician will readily understand the precision of the measurement, but other people would not know whether this was good or bad without an appreciation of typical values for the scale.

Kappa statistics should be used when comparing categorical data.[64] This statistic is designed to indicate the agreement between two measurers using the same nominal scale, but corrected for agreement that would occur by chance. Kappa varies from -1 to 1 where 0 is equivalent to agreement occurring by chance. Kappa of 0.81 or greater represents "almost perfect agreement", and between 0.61 and 0.80 represents "substantial agreement".[65] A Kappa statistic ≥ 0.70 is desirable for reliability testing of instrument responses. A weighted Kappa statistic is designed for ordinal categorical data such as that seen with instrument response scales where greater penalty is given for pairs with greater disagreement over scale categories. Kappa weighted with the quadratic weighting scheme is mathematically identical to the ICC.[66] Notably, Kappa statistics depend upon the prevalence of the characteristic under study so are not directly comparable across measures.

In addition to the above tests, Rasch analysis also provides *person and item separation reliability indices*, indicating the overall performance of an instrument. It is the ratio of the true variance in the estimated measures to the observed variance and indicates the number of distinct person strata that can be distinguished.[36] There are a number of versions of separation including the Person Separation Index (PSI) or person separation reliability, which can range from 0 to 1, with high values indicating better reliability. A PSI value of 0.8 is the equivalent of a G value (person separation ratio) of 2.0, representing the ability to distinguish three distinct strata of person ability.[58,67] A value of 0.9 is equivalent to a G value of 3, with the ability to distinguish four strata of person ability. Item separation reliability should also be reported with 0.8 being the cutoff for both in terms of acceptability.

Other Important Indicators: Responsiveness and Interpretation. *Responsiveness* is the extent to which the instrument can detect clinically important changes over time.[68,69] This can be studied in patients who are known to undergo a change in status over a time frame, e.g., before and after cataract surgery. The perspective of what constitutes a "clinically important" change is given by the minimum clinically important difference (MID). The MID indicates the smallest difference in score that can be perceived as beneficial by the subject. This is calculated relative to a difference reported by a patient. For example, one could ask cataract patients: "By how much has the operation improved your vision?" and provide the options: "made it worse", "not at all", "a little", "quite a bit", "a lot". The score change in the instrument of interest that equates to a change in status from one step to the next on this

question can be used to calculate the MID with receiver operating characteristic analysis. The MID ideally should be larger that the LoA of test-retest reliability of the instrument, as this means that the reliability of the test does not interfere with detection of the MID. Although this criterion may not always be achieved, a MID comparable to the LoA is still scored as a positive result (Table 1).

To demonstrate that an instrument is responsive to an intervention, the mean change e.g., with cataract surgery needs to be greater than the MID. Responsiveness can be expressed by a number of statistics: Effect Size, the difference between pre and post operative score divided by the preoperative standard deviation; standardized response mean (SRM), the mean of the change scores divided by the standard deviation of the change scores; and Responsiveness Statistic (RS), the difference between pre and post operative score divided by the standard deviation of retest score. Effect size, SRM and RS are considered to be large if >0.80.[70] For each of these measures, convention holds that effect sizes of 0.20 to 0.49 are considered small; 0.50 to 0.79 are moderate, and 0.80 or above are large.[70]

Interpretation indicates the degree to which scores on a measure can be considered meaningful. To ensure interpretation of an instrument, the instrument should be tested on a representative target population whose demographics are fully described. Normative scores and the minimum clinically important difference (see responsiveness) should be described. The amount of interpretation information that should be described depends on the purpose of an instrument. For example, an instrument intended for cataract surgery probably need only report normative data (means and SDs) for typical populations of bilateral and second eye surgery cases. Although one could perhaps argue that cataract only and cataract and comorbidity populations should also be described. Contrast this to an instrument designed for use across all ophthalmic conditions; normative data would need to be provided for a great many eye diseases. Data may also need to be provided for subgroups other than disease: e.g., age, gender, socioeconomic status. Scores before and after important interventions e.g., cataract surgery should also be provided.

Recommendations for Instrument Selection

In this article, we have presented a range of methods and analysis techniques for developing and validating instruments and scales. These guidelines are intended to help investigators understand what determines instrument quality and to assist interpretation of articles detailing instrument development. Once the basic principles of psychometric methods are understood, we recommend that researchers wishing to include a QoL measure in a study or clinical trial, and not wishing to develop and validate their own instrument, use the following instrument selection process and the quality criteria presented in Table 1 to guide their selection of an appropriate instrument.

Instrument Selection Process.

1. Be sure that the content area of the instrument suits the purpose of your study.
2. Be aware of what it was developed for and whom it was developed on and not just assume that it will work on your sample. Be aware of cultural differences.
3. Check that appropriate item selection and reduction processes were used and that the final number of items in the instrument is not too large as to represent a burden to respondents.
4. Check the scaling for whether adding scores is justified statistically. Note that some traditionally developed instruments can be Rasch scaled to provide a more sensitive and effective (although perhaps not ideal) measurement. Score-to-measure tables that provide a cross-walk between total raw scores and Rasch measures for some traditionally developed instruments, such as the ADVS, RSVP,[11] NEI-VFQ,[54] may be published, or available on request from researchers who have investigated the performance of such instruments within the Rasch model.
5. Check that the validity and reliability of the instrument are adequate for your purposes.
6. Check for useful interpretation and responsiveness data that correspond to your intended purpose.

It is likely that many existing instruments will not have been tested in all the ways recommended herein. By necessity, these quality assessment criteria must be comprehensive. However, existing instruments which have not been tested on certain criteria are not necessarily flawed, just untested. Such instruments may give useful information, but should be used with caution.

CONCLUSION

The quality assessment criteria proposed herein may be useful to guide new instrument development, redevelopment of existing instruments or assessment of existing instruments whether for choosing an instrument for use or as part of a formal review of instruments. Questionnaire research is a dynamic field, with the importance of item response theory, particularly Rasch analysis, gaining prominence in recent ophthalmic instruments.[71] We have sought to represent this progress in these quality assessment criteria while remaining inclusive of traditional methods. These quality criteria should be considered as a proposal, and we acknowledge that debate over the appropriateness of these criteria will likely occur. However, we welcome this debate as we believe it can only lead to the evolution of better quality assessment criteria and in turn better assessment of patient-centered outcomes.

ACKNOWLEDGEMENTS

We thank Professor Peter Fayers, Department of Public Health, University of Aberdeen, for his initial guidance on traditional methods for quality assessment criteria reported in this article. We also thank Dr. Trudy Mallinson for her helpful advice on this manuscript.

Received May 2, 2007; accepted June 6, 2007.

REFERENCES

1. Likert RA. A technique for the measurement of attitudes. Arch Psychol 1932;140:1–55.
2. de Boer MR, Moll AC, de Vet HC, Terwee CB, Volker-Dieben HJ, van Rens GH. Psychometric properties of vision-related quality of life questionnaires: a systematic review. Ophthal Physiol Opt 2004;24:257–73.
3. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for

measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
4. Day H, Jutai J, Campbell KA. Development of a scale to measure the psychosocial impact of assistive devices: lessons learned and the road ahead. Disabil Rehabil 2002;24:31–7.
5. Jutai J, Day H, Woolrich W, Strong G. The predictability of retention and discontinuation of contact lenses. Optometry 2003;74: 299–308.
6. Day HY, Jutai J, Woolrich W, Strong G. The stability of impact of assistive devices. Disabil Rehabil 2001;23:400–4.
7. Day H, Campbell KA. Is telephone assessment a valid tool in rehabilitation research and practice? Disabil Rehabil 2003;25:1126–31.
8. Vitale S, Schein OD, Meinert CL, Steinberg EP. The refractive status and vision profile: a questionnaire to measure vision-related quality of life in persons with refractive error. Ophthalmology 2000;107: 1529–39.
9. Schein OD. The measurement of patient-reported outcomes of refractive surgery: the refractive status and vision profile. Trans Am Ophthalmol Soc 2000;98:439–69.
10. Schein OD, Vitale S, Cassard SD, Steinberg EP. Patient outcomes of refractive surgery. The refractive status and vision profile. J Cataract Refract Surg 2001;27:665–73.
11. Garamendi E, Pesudovs K, Stevens MJ, Elliott DB. The Refractive Status and Vision Profile: evaluation of psychometric properties and comparison of Rasch and summated Likert-scaling. Vision Res 2006; 46:1375–83.
12. Nichols JJ, Mitchell GL, Saracino M, Zadnik K. Reliability and validity of refractive error-specific quality-of-life instruments. Arch Ophthalmol 2003;121:1289–96.
13. Nichols JJ, Twa MD, Mitchell GL. Sensitivity of the National Eye Institute Refractive Error Quality of Life instrument to refractive surgery outcomes. J Cataract Refract Surg 2005;31:2313–8.
14. McDonnell PJ, Mangione C, Lee P, Lindblad AS, Spritzer KL, Berry S, Hays RD. Responsiveness of the National Eye Institute Refractive Error Quality of Life instrument to surgical correction of refractive error. Ophthalmology 2003;110:2302–9.
15. Hays RD, Mangione CM, Ellwein L, Lindblad AS, Spritzer KL, McDonnell PJ. Psychometric properties of the National Eye Institute-Refractive Error Quality of Life instrument. Ophthalmology 2003;110:2292–301.
16. McDonnell PJ, Lee P, Spritzer K, Lindblad AS, Hays RD. Associations of presbyopia with vision-targeted health-related quality of life. Arch Ophthalmol 2003;121:1577–81.
17. Pesudovs K, Garamendi E, Elliott DB. The Quality of Life Impact of Refractive Correction (QIRC) Questionnaire: development and validation. Optom Vis Sci 2004;81:769–77.
18. Garamendi E, Pesudovs K, Elliott DB. Changes in quality of life after laser in situ keratomileusis for myopia. J Cataract Refract Surg 2005; 31:1537–43.
19. Pesudovs K, Garamendi E, Elliott DB. A quality of life comparison of people wearing spectacles or contact lenses or having undergone refractive surgery. J Refract Surg 2006;22:19–27.
20. Lamoureux EL, Pallant JF, Pesudovs K, Hassell JB, Keeffe JE. The Impact of Vision Impairment Questionnaire: an evaluation of its measurement properties using Rasch analysis. Invest Ophthalmol Vis Sci 2006;47:4732–41.
21. Lamoureux EL, Ferraro JG, Pallant JF, Pesudovs K, Rees G, Keeffe JE. Are standard instruments valid for the assessment of quality of life and symptoms in glaucoma? Optom Vis Sci 2007;84:789–96.
22. Trochim WMK. The Research Methods Knowledge Base, 2nd ed. Cincinnati, OH: Atomic Dog Publishing; 2000.
23. Guion RM. Content validity: the source of my discontent. App Psychol Meas 1977;1:1–10.
24. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use, 3rd ed. Oxford: Oxford University Press; 2003.
25. Seiler LH. The 22-item scale used in field studies of mental illness: a question of method, a question of substance, and a question of theory. J Health Soc Behav 1973;14:252–64.
26. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press; 1987.
27. World Health Organization. The International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001.
28. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 2004;38:327–33.
29. La Grow S. Predicting perceived quality of life scores from the NEI-VFQ-25. Optom Vis Sci 2007;84:785–8.
30. Valderas JM, Alonso J, Prieto L, Espallargues M, Castells X. Content-based interpretation aids for health-related quality of life measures in clinical practice. An example for the visual function index (VF-14). Qual Life Res 2004;13:35–44.
31. Uusitalo RJ, Brans T, Pessi T, Tarkkanen A. Evaluating cataract surgery gains by assessing patients' quality of life using the VF-7. J Cataract Refract Surg 1999;25:989–94.
32. Steinberg EP, Tielsch JM, Schein OD, Javitt JC, Sharkey P, Cassard SD, Legro MW, Diener-West M, Bass EB, Damiano AM, Steinwachs DM, Sommer A. The VF-14. An index of functional impairment in patients with cataract. Arch Ophthalmol 1994;112:630–8.
33. Krueger RA. Focus Groups: A Practical Guide for Applied Research, 2nd ed. Thousand Oaks, CA: Sage Publications; 1994.
34. Caudle LE, Williams KA, Pesudovs K. The Eye Sensation Scale: an ophthalmic pain severity measure. Optom Vis Sci 2007;84: 752–62.
35. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004;7 (Suppl 1):S22–6.
36. Mallinson T. Why measurement matters for measuring patient vision outcome. Optom Vis Sci 2007;84:675–82.
37. Wright BD Fundamental Measurement for Psychology. In: Embretson SE, Hershberger SL, eds. The New Rules of Measurement: What Every Psychologist and Educator Should Know. Mahway, NJ: Lawrence Erlbaum;1999:65–104.
38. Massof RW. The measurement of vision disability. Optom Vis Sci 2002;79:516–52.
39. Lamoureux EL, Pallant JF, Pesudovs K, Rees G, Hassell JB, Keeffe JE. The impact of vision impairment questionnaire: an assessment of its domain structure using confirmatory factor analysis and Rasch analysis. Invest Ophthalmol Vis Sci 2007;48:1001–6.
40. Linacre JM. Structure in Rasch residuals: why principal components analysis? Rasch Meas Trans 1998;12:636. Available at: http://www.rasch.org/rmt/rmt122m.htm. Accessed June 8, 2007.
41. Massof RW, Ahmadian L, Grover LL, Deremeik JT, Goldstein JE, Rainey C, Epstein C, Barnett GD. The Activity Inventory (AI): An adaptive visual function questionnaire. Optom Vis Sci 2007;84: 763–74.
42. Massof RW, Hsu CT, Baker FH, Barnett GD, Park WL, Deremeik JT, Rainey C, Epstein C. Visual disability variables. II. The difficulty of tasks for a sample of low-vision patients. Arch Phys Med Rehabil 2005;86:954–67.
43. Massof RW, Hsu CT, Baker FH, Barnett GD, Park WL, Deremeik JT, Rainey C, Epstein C. Visual disability variables. I: the importance and difficulty of activity goals for a sample of low-vision patients. Arch Phys Med Rehabil 2005;86:946–53.

44. Stelmack J, Massof RW. Using the VA LV VFQ-48 in low vision rehabilitation. Optom Vis Sci 2007;84:705–9.
45. Stelmack JA, Szlyk JP, Stelmack TR, Demers-Turco P, Williams RT, Moran D, Massof RW. Psychometric properties of the Veterans Affairs Low-Vision Visual Functioning Questionnaire. Invest Ophthalmol Vis Sci 2004;45:3919–28.
46. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: L. Earlbaum, 2001.
47. Stelmack J, Szlyk JP, Stelmack T, Babcock-Parziale J, Demers-Turco P, Williams RT, Massof RW. Use of Rasch person-item map in exploratory data analysis: A clinical perspective. J Rehabil Res Dev 2004;41:233–41.
48. Mangione CM, Phillips RS, Seddon JM, Lawrence MG, Cook EF, Dailey R, Goldman L. Development of the 'Activities of Daily Vision Scale'. A measure of visual functional status. Med Care 1992;30: 1111–26.
49. Pesudovs K, Garamendi E, Keeves JP, Elliott DB. The Activities of Daily Vision Scale for cataract surgery outcomes: re-evaluating validity with Rasch analysis. Invest Ophthalmol Vis Sci 2003;44:2892–9.
50. Linacre JM. Size vs. significance: Standardized chi-square fit statistic. Rasch Meas Trans 2003;17:918. Available at: http://www.rasch.org/rmt/rmt171n.htm. Accessed May 25, 2007.
51. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. Med Care 2004;42:I25–36.
52. Massof RW, Fletcher DC. Evaluation of the NEI visual functioning questionnaire as an interval measure of visual ability in low vision. Vision Res 2001;41:397–413.
53. Massof RW. An interval-scaled scoring algorithm for visual function questionnaires. Optom Vis Sci 2007;84:689–704.
54. Massof RW. Application of stochastic measurement models to visual function rating scale questionnaires. Ophthalmic Epidemiol 2005; 12:103–24.
55. Thomee R, Grimby G, Wright BD, Linacre JM. Rasch analysis of Visual Analog Scale measurements before and after treatment of Patellofemoral Pain Syndrome in women. Scand J Rehabil Med 1995;27:145–51.
56. Pesudovs K, Noble BA. Improving subjective scaling of pain using Rasch analysis. J Pain 2005;6:630–6.
57. Andrich D. A rating scale formulation for ordered response categories. Psychometrika 1978;43:561–73.
58. Wright BD, Masters GN. Rating Scale Analysis. Chicago: MESA Press; 1982.
59. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol 1991;44:381–90.
60. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1: 307–10.
61. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60.
62. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337–40.
63. Patton N, Aslam T, Murray G. Statistical strategies to assess reliability in ophthalmology. Eye 2006;20:749–54.
64. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Stat Med 2002;21:2109–29.
65. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
66. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613–9.
67. Fisher W Jr. Reliability statistics. Rasch Meas Trans 1992;6:238. Available at: http://www.rasch.org/rmt/rmt63i.htm. Accessed June 8, 2007.
68. Brozek JL, Guyatt GH, Schunemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes 2006;4:69.
69. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 2004;57:898–910.
70. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–68.
71. Pesudovs K. Patient-centred measurement in ophthalmology—a paradigm shift. BMC Ophthalmol 2006;6:25.

Konrad Pesudovs
NH&MRC Centre for Clinical Eye Research
Department of Ophthalmology
Flinders Medical Centre
Bedford Park. SA 5042, Australia
e-mail: Konrad.Pesudovs@flinders.edu.au