PERGAMON

Vision Research 41 (2001) 397–413

Vision Research

www.elsevier.com/locate/visres

Evaluation of the NEI visual functioning questionnaire as an interval measure of visual ability in low vision

Robert W. Massof [a,*], Donald C. Fletcher [b,c]

[a] *Lions Vision Research and Rehabilitation Center, Wilmer Ophthalmological Institute, The Johns Hopkins University School of Medicine, 6th Floor, 550 North Broadway, Baltimore, MD 21205, USA*

[b] *Retina Consultants of Southwest Florida, Fort Myers, FL, USA*

[c] *Helen Keller Eye Research Foundation, 1201 11th Ave S., Suite 300, Birmingham, AL 35205, USA*

Received 12 August 1999; received in revised form 31 July 2000

Abstract

The National Eye Institute developed a visual functioning questionnaire (NEI-VFQ) designed to assess health-related quality of life of patients with visual impairments. The developers of the NEI-VFQ distributed the original 52 items into 13 different domains. The recommended method for scoring the NEI-VFQ is to linearly transform the sum of the ordinal ratings to each item within each domain to produce 13 scores. The major shortcoming of this scoring method is that sums of ordinal numbers do not necessarily generate valid measurement scales. However, Rasch models can be used to estimate interval measurement scales from ordinal responses to items. We administered 27 items from the 52-item NEI-VFQ to 341 patients with low vision. Rasch analysis was used to estimate the 'visual ability' required by each item for a particular response (item measures) and to estimate the 'visual ability' of each patient (person measures). The validity of the model was evaluated by examining the distributions of residuals for item and person measures. We observed that the 17 items we tested from the NEI-VFQ that require difficulty ratings produce a valid interval scale for low-vision patients. The estimated person measures of visual ability are linear with log MAR acuity. The ten items that require frequency or level of agreement ratings do not work together to produce a valid interval scale. Rather, these items appear to be confounded by other variables distributed in the patient sample (e.g. psychological state). The visual ability scale estimated from the 17 NEI-VFQ items is proportional to the visual ability scales estimated from two earlier studies that also elicited difficulty ratings from low-vision patients.

Keywords: Low vision; Visual impairment; Visual function; Quality of life; Rasch analysis; Item response model; Psychometric; Functional assessment; Questionnaire

1. Introduction

Visual acuity, visual fields, and other psychophysical tests of vision have been the traditional clinical measures of functional capability in patients with diseases of the visual system. Until recently, visual acuity, in particular, had been the primary variable in epidemiological studies of visual impairments and blindness (Tielsch, Sommer, Witt, et al., 1990) and in ophthalmologic clinical intervention trials (Macular Photocoagulation Study Group, 1991). Visual acuity also has served as the major criterion for establishing the medical necessity of cataract surgery (Cataract Management Guidelines Panel, 1993), defining restrictions on driving and other regulated activities (Shipp, 1998), and determining eligibility for blindness-related benefits (Social Security Administration, 1997). With the changing health care system placing strong emphasis on patient-based assessments of health care outcomes, psychophysical measures of visual impairments have come to be considered insufficient descriptors of visual function (Kupfer, 1996). Today, everyone from health care managers to agencies that fund research is demanding patient-based assessments of the impact of visual impairment on the patient's quality of life (Stoline & Weiner, 1993).

'Health-related quality of life' (HRQOL) is difficult to define and challenging to measure. Many investigators and clinicians have equated HRQOL with the

* Corresponding author. Fax: +1-410-9551829.
E-mail address: rmassof@lions.med.jhu.edu (R.W. Massof).

0042-6989/01/$ - see front matter
PII: S0042-6989(00)00249-2

ability to engage in everyday activities. Other investigators and clinicians wish to add general health and well-being, psychological and emotional state, and/or social and economic stress to the definition of HRQOL (McDowell & Newell, 1996). Several psychometric patient-based and provider-based rating instruments have been developed to assess each of these HRQOL domains (Beck, Ward, Mendelson, et al., 1961; Neugarten, Havighurst, & Tobin, 1961; Pfeffer, Kurosaki, Harrah, et al., 1982; Sherbourne & Stewart, 1991; Granger & Hamilton, 1993). Other assessment tools have been developed that try to include all or several of the HRQOL domains in a single instrument (Bergner, Bobbitt, Carter, et al., 1981; Ware & Sherbourne, 1992).

Over the past two decades, more than a dozen psychometric instruments have been developed, validated, and used to measure the impact of visual impairments on daily function (Massof & Rubin, 2000). All of these instruments elicit patients' judgments about the difficulty they have performing specific everyday activities. Patient responses are usually in the form of a difficulty rating for each item, ranging from dichotomous to ten-level scales. Rating descriptions include level of difficulty (Mangione, Phillips, Seddon, et al., 1992), frequency of problems (Sloane, Ball, Owsley, et al., 1992), level of agreement with a statement about problems (Mangione et al., 1998b), and level of disability (Elliott, Hurst, & Weatherill, 1990).

Recently, the National Eye Institute (NEI) contracted the Rand Corporation to develop a vision-specific HRQOL instrument, the NEI Visual Functioning Questionnaire (NEI-VFQ) (Ellwein, Fletcher, Negrel, & Thulasiraj, 1995; Mangione et al., 1998a; Mangione et al., 1998b). The April 1995 test version of the NEI-VFQ has 52 items (i.e. questions) grouped into three parts (some publications and documentation refer to a 51-item instrument (Mangione et al., 1998b), presumably because item 9 was not included in any subscale). Part 1 consists of 14 questions about general health and vision history. Part 2 has 20 questions about difficulty performing everyday activities, four questions about difficulty of driving, and six general questions about functional limitations. Part 3 has eight questions about coping with vision loss. Each item in part 1 requires a rating on one of five different scales. All but the last six items in part 2 require ratings of difficulty on a five-point scale (with a sixth possible response that is equivalent to 'not applicable'). The last six items in part 2 require ratings of frequency on a five-point scale and the eight items in part 3 also require ratings on a five-point scale, but of agreement with the statement in the item.

The developers of the NEI-VFQ propose that 13 different HRQOL subscales (i.e. domains) are sampled by the instrument: (1) general health, (2) general vision, (3) ocular pain, (4) near vision, (5) distance vision, (6) social functioning, (7) emotional well-being, (8) vision expectations, (9) role difficulties, (10) dependency, (11) driving, (12) color, and (13) peripheral vision. Each item (except item 9) contributes to only one domain. For one-half of the domains, only one to three questions define the subscale (Mangione et al., 1998b). Nothing has been published about how these domains were defined. A recent validation/reliability study demonstrated correlations among domains ranging from -0.04 to 0.85 (Mangione et al., 1998a). The average rank–order correlation among the subscales is 0.42. If the general health, ocular pain, and visual expectation subscales are removed, the average correlation increases to 0.62. This study did not include a factor analysis to validate the subscales.

The NEI-VFQ is scored by linearly transforming the patient's ratings for each item to values that range from 0 to 100. The average of the item scores for each subscale is computed to produce 13 domain scores, each of which ranges from 0 to 100. Some investigators linearly combine the domain scores to produce a single instrument score (Parrish, Gedde, Scott, et al., 1997).

Because respondents are instructed to do so, the respondent's ratings form an ordinal scale for each item. Although the same category labels are used for different classes of responses (e.g. a scale of one to five for level of difficulty and for frequency of performing an activity), one cannot assume that different response classes produce equivalent ordinal scales. Even if they did, or confining the discussion to a single response class, arithmetic operations cannot be performed on ordinal numbers (e.g. if we only know that $A > B$ and $B > C$, we do not know the relation of $B + C$ to A). Consequently, subscales that are defined as averages of item scores are uninterpretable. We cannot even assume that such subscales are nominal (i.e. equal subscale scores can be interpreted as equality of the attributes of interest).

The purpose of the NEI-VFQ is to provide an instrument for measuring visual impairment-related HRQOL. To qualify as a measurement, the instrument must produce an interval or a ratio scale (Wright & Linacre, 1989). Interval scales have an arbitrary origin (e.g. measurement of temperature on a Celsius or Fahrenheit scale); for ratio scales, the origin is absolute (e.g. the measurement of distance in centimeters or inches). The NEI-VFQ, or some subset of the items, might produce results that can be expressed as an ordinal scale. If so, through the use of item response models, an interval scale can be estimated from the ordinal scale. The purpose of this study is to evaluate items on the NEI-VFQ to determine if they can be used to estimate an interval vision disability scale for low-vision patients. First, however, it is necessary to review the basics of item response theory and the Rasch measurement

model — the specific item response model that we will use to evaluate the NEI-VFQ measurement scale.

2. Review of item response theory

Variables can be 'manifest' or 'latent'. Manifest variables can be observed and measured publicly. Most physical variables (e.g. length, weight, force, etc.) are manifest. Latent variables are inferred from judgments. Most psychological variables are latent; they are inferred from subjects' reports or observer judgments of subjects' behavior.

Psychometric instruments such as the NEI-VFQ obtain the visually impaired person's ratings of the difficulty he/she has in performing specific activities. Presumably, some activities require high visual ability to be performed with ease while other activities can be performed with ease despite low visual ability. For example, a mild vision impairment might make it impossible to thread a needle but have little impact on the person's ability to socialize with their friends. A severe vision impairment would not only render needle-threading impossible, but might also make it difficult to socialize. Thus, we would conclude that threading a needle requires a high level of the latent visual ability variable, whereas socializing requires a much lower level of visual ability. Each item on the instrument represents a specific activity that requires a specific level of visual ability to perform with ease. Thus, in principle, we should be able to order the items according to their required visual ability.

Item response models are designed to estimate the values of latent variables on an interval scale from item scores that form an ordinal scale. Item scores, or linear combinations of item scores, are called 'raw scores'. The recommended method of computing domain scores for the NEI-VFQ produces raw scores. If the raw scores form a unidimensional ordinal scale, then when the data are displayed with the items ordered according to item raw scores (the sum of subject responses to each item) and with the subjects ordered according to individual raw scores (the sum of each subject's responses across all items), the data matrix will conform to a Guttman scale (Guttman, 1950).

A Guttman scale means that item raw scores are monotonic with item difficulty, and test scores are monotonic with the subject's ability. The sum of scores across items for each person is the person raw score and the sum of scores across people for each item is the item raw score. If the raw scores conform to a Guttman scale, then when people are rank-ordered by person raw score and items are rank-ordered by item raw score, the person rankings will be the same for each item and item rankings will be the same for each person. There are likely to be inconsistencies with this rigid rule, but the overall statistical pattern of responses should agree with these expectations. The more closely the data agree with this Guttman scale, the more likely it is that the raw scores represent at least an ordinal scale.

Item response theory begins with an explicit definition of the latent variable that the instrument is supposed to measure, θ. This variable is an attribute of the patient and will have a unique value for each patient n, θ_n. Each item of the instrument requires a threshold value of θ to elicit a particular response from the patient 50% (or some other criterion percentage) of the time. The response threshold for item i, b_i, is in the same units as θ. The probability that patient n will give a particular response to item i can be modeled with Birnbaum's logistic:

$$P(\theta_{ni}) = c + \frac{d-c}{1+\mathrm{e}^{-a_i(\theta_n - b_i)}} \quad (1)$$

where c is the lower performance asymptote ($0 \le c < 1$), d is the upper performance asymptote ($0 < d \le 1$),and a_i controls the slope of the item response function (Birnbaum, 1968). The parameter c usually refers to chance performance, d is controlled by the rate of careless response errors, and a is the discriminability of the item. Other models have been proposed, some of which incorporate a skew parameter (Prentice, 1976). The choice of model for the item response function is not particularly important for our application of the theory. All models are designed to generate 'S-shaped' curves. They differ only in details that probably cannot be resolved by the data anyway.

In the case of the NEI-VFQ, there are no 'right' or 'wrong' answers. Therefore, $c = 0$ and $d = 1$ in the item response model of Eq. (1). If θ and b are in the same units, then item-dependent variations in the slope of the item response function must indicate different levels of measurement noise for different items. Measurement noise could be due to instability in θ, instability in b, or both. It also could be attributed to variables not under study and, therefore, not made explicit in the model. The parameter a soaks up the variability and creates the illusion of precision in the estimation of person and item values. Furthermore, an item-dependent slope parameter is inconsistent with the basic tenents of measurement theory because it implies that the measurement units vary across items (Wright, 1977). Therefore, if we define $a = 1$ and live with imprecision in the estimation, which, in any event, is probably real, then the estimates of the item response model can be interpreted as measurements of a single variable.

This simplified item response model is identical to the probabilistic measurement model developed independently by Georg Rasch, a Danish mathematician (Rasch, 1960; Andersen, 1973; Hambleton & Cook, 1977; Molenaar, 1995). Rasch deduced his model from principles of measurement theory and probably would

disapprove our association of his model with item response theory (Andersen, 1995). He proved that the person and item parameters (θ_n and b_i) are separable, and that item and subject raw scores are sufficient statistics to estimate the values of the subject and item parameters. Rasch models are used to estimate values on an interval scale from raw scores of psychometric instruments (the Rasch model is reviewed by Massof (1998)). Rasch models have been used extensively for the past 15 years to analyze data from functional assessment instruments in rehabilitation medicine (Fisher, Harvey, Taylor, et al., 1995).

To apply item response theory to the assessment of visual function, we must begin with an explicit definition of the variable we wish to measure. Most generally, we are interested in the patient's visual ability to function in everyday life. So, 'visual ability', α, will be our variable of interest. Like all variables, this variable is a theoretical construction.

Each patient, n, has a unique visual ability, α_n, that we wish to measure (the person parameter). To 'function in everyday life' means that the patient must perform a series of daily activities at an acceptable rate, while expending an acceptable amount of effort. Depending on the patient's visual ability, some activities will seem easy to perform and others will be difficult or impossible. Patients with more visual ability will be able to perform a greater number of activities with ease than will patients with less visual ability. (This simple statement is the premise underlying the traditional use of the test raw score as a 'measure' of functional ability.) Thus, we can think of each activity as requiring of the patient a specific level of visual ability (α), for the patient to be able to perform that activity with ease. The threshold value of α required to perform activity i with ease is the item parameter ρ_i.

To simplify the review of the Rasch model, consider only dichotomous responses. If patients respond that they can perform an activity with ease, we assign a score of 1 to that item; otherwise, we assign a score of 0. The probability that patient n will report that he/she can perform activity i with ease is:

$$P(1|\alpha_n,\rho_i) = \frac{e^{\alpha_n - \rho_i}}{1 + e^{\alpha_n - \rho_i}} \tag{2}$$

which is Eq. (1) with $a_i = 1$, $c = 0$, $d = 1$, $\theta_n = \alpha_n$, and $b_i = \rho_i$. The probability that patient n will report that he/she cannot perform activity i with ease is:

$$P(0|\alpha_n,\rho_i) = 1 - P(1|\alpha_n,\rho_i) = \frac{1}{1 + e^{\alpha_n - \rho_i}} \tag{3}$$

The odds that patient n will report that he/she can perform activity i with ease is:

$$\frac{P(1|\alpha_n,\rho_i)}{P(0|\alpha_n,\rho_i)} = e^{\alpha_n - \rho_i} \tag{4}$$

and the log of the odds ratio, or 'logit', is:

$$\ln \frac{P(1|\alpha_n,\rho_i)}{P(0|\alpha_n,\rho_i)} = \alpha_n - \rho_i \tag{5}$$

which isolates the parameters of interest.

The person and item parameters can be estimated from response odds ratios in the data set using a constrained form of Eq. (5). Because there are no free model parameters, the Rasch model is prescriptive rather than descriptive, i.e. either the data fit the model, or the assumptions of the model must be rejected for the data set. The model assumptions are: (1) the subjects used to test the model differ in visual ability, (2) the subjects' responses to items in the instrument depend only on visual ability, (3) subjects' responses are probabilistic and conditional on the subject's visual ability and the visual ability required to perform the activity with ease, and (4) the odds of performing an activity with ease increase monotonically with the difference between the subject's ability and the ability required to perform the activity with ease. If an item is sensitive to more than one variable (i.e. 'domain') distributed in the subject sample, then the item will appear to be noisy or an outlier when evaluating the fit of the data to the model. Similarly, if patients have another problem unrelated to vision that limits his/her ability to perform the activity, that person's response pattern will be identified as noisy or outlying relative to the expectations of the model.

The item response function model in Eq. (1), as well as the simplified version that is equivalent to the Rasch model in Eq. (2), applies to instruments that elicit dichotomous responses. Andrich (1978), Masters (1982) modified the Rasch model to make it applicable to polytomous rating scale instruments, the response scale that is used in the NEI-VFQ. The modified Rasch model assumes that ρ refers to the visual ability required to respond with rating category x to item i, i.e. ρ_{ix}, and assumes that Eq. (2) refers to the probability of subject n responding with rating category x rather than rating category $x-1$ to item i (see equations (A.1)–(A.8) in Massof (1998)). Everything else about the model is the same.

The power of the Rasch model is that it provides estimates of the variables of interest on an interval scale and allows one to test the validity of any psychometric instrument with an objective set of criteria. The tests of construct validity are the fit of the person measures to the model, and the correlations of person and item parameter values with other variables, compared with expected correlations. The tests of content validity are the fits of individual items to the model, the estimation errors of item parameter values, and the spacing and range of item parameter values, relative to the distribution of person parameter values. The tests of criterion validity are the estimation precision and distribution of

person parameter values, and the discriminability of person parameter values relative to external criteria. Reliability is assessed in the traditional way, i.e. test–retest consistency, and the dependency of item and person parameter values on test conditions and mechanics of administration. All of these validity and reliability tests are analytical and parametric, and can be compared objectively across instruments.

3. Application of Rasch models to visual function assessments

Becker, Lambert, Schulz, Wright, and Burnet (1985) developed an instrument to assess the 'activity level of the blind' (ALB) (note that Becker et al. named their instrument the 'functional activities questionnaire'; however, that name is the same as the pre-existing proxy-based assessment instrument of Pfeffer et al. (1982)). The ALB has two groups of items, 36 items that represent single specific 'skills' (e.g. 'read print'), and 38 items that represent more 'general activities' that require a range of skills (e.g. 'grocery shopping'). For the skill items, the three types of responses were 'difficulty', 'independence', and 'motivation to learn'. For the general activities items, the two types of responses were 'frequency' and 'feeling of loss' (i.e. rating of how much the performance of an activity is missed). For each type of response, the subject used a three-point rating scale, except for frequency, which had nine response categories.

The ALB was administered to 129 blind or visually impaired veterans who were waiting to be admitted to the Hines VA Blind Rehabilitation Center. Rasch analysis was used to estimate an interval scale for each type of response. For each response type, a value on an interval scale was estimated for each item (e.g. item difficulty) and for each person (e.g. capability). On the difficulty scale for the skill items, reading print was the most difficult item (disregarding the Braille items), and grooming hair and dressing oneself were the least difficult. Matching clothes, crossing a street, taking a taxi, and using the stove were all of intermediate difficulty on the interval scale.

More recently, Massof (1998) estimated an interval scale for 'visual ability to live independently' from a Rasch analysis of the responses of 445 low-vision patients to 24 items. Most of the items in that study describe activities that Becker et al. (1985) would classify as 'general activities', and subjects rated on a six-point scale the difficulty of performing each activity without assistance. Similar to the results of Becker et al., 'recreational reading' was the most difficult activity for low-vision patients and 'self care' was the least difficult. 'Watching television', 'managing personal finances', and 'outdoor recreational activities' were of average difficulty on the estimated interval scale.

Turano, Geruschat, Stahl, and Massof (1999) developed a 35-item questionnaire to assess the difficulty visually impaired patients have with mobility. Each item describes a specific mobility situation, and subjects rate on a five-point scale the difficulty of independent mobility under the described conditions. Rasch analysis was used to estimate an interval 'perceived visual ability for mobility' scale from the responses of 127 subjects with retinitis pigmentosa. The most difficult item was 'walking at night' and the least difficult was 'moving about in the home'. 'Using public transportation', 'being aware of another person's presence', and 'avoiding bumping into head-height objects' were items of average difficulty on the interval scale.

These three studies concur that valid interval scales for visual ability can be constructed from rating scale responses to items by visually impaired people. To the extent that many items on the NEI-VFQ are similar to items used in these three studies, we expect that an interval visual ability scale can be estimated from patient responses to the NEI-VFQ. However, within the framework of item response theory, the variable measured by an instrument is defined by the patient sample. The NEI-VFQ and nearly all other visual function instruments have questions about the difficulty of reading print intermingled, for example, with questions related to mobility difficulties. Retinitis pigmentosa patients are more likely to have mobility problems before reading problems, whereas the reverse is likely to be true for macular degeneration patients. Therefore, to increase the likelihood of estimating a valid interval scale for the NEI-VFQ, this study is limited to low-vision patients, most of whom have impaired central vision.

4. Methods

One of the authors (D.C.F.) routinely administers a subset of NEI-VFQ items to low-vision patients in his private practice as part of the history taken during the first visit evaluation. This study is a retrospective analysis of patient responses to those NEI-VFQ items for the first 341 patients who were administered the questionnaire. The study protocol was reviewed and approved by the Johns Hopkins human subject institutional review board.

4.1. Subjects

All subjects were private patients of D.C.F.'s low-vision rehabilitation service. None of the patients were participating in a research study. Patient age ranged from 11 to 94 years (median, 79 years). Primary diagnoses of visual system disorders were: age-related macular degeneration, 76%; diabetic retinopathy, 9%;

glaucoma, 5%; homonymous hemianopia post stroke, 2%; optic neuropathy, 2%; retinal vascular occlusive disease, 1%; retinitis pigmentosa, 1%; and other, 4%. Sixty percent were female. Forty-six percent of the patients had bilateral central scotomas (tested with the scanning laser ophthalmoscope). Entering corrected binocular visual acuity, measured with a rear-illuminated ETDRS chart, ranged from 20/20 to light perception (median, 20/200). Log-contrast sensitivity, measured with a LH Low Contrast Test, ranged from 0.1 to 2.2 (median, 1.2; normal, >1.65 for this age group; see, for example, Rubin, Adamsons, & Stark, 1993).

4.2. Procedure

Twenty-seven items from the field test version of the NEI-VFQ, representing nine of the 13 subscales, were administered by interview to the low-vision patients as a routine part of the initial evaluation. The same interviewer, a trained ophthalmic technician, was used for all patients. The interview occurred before the patient was seen by any other members of the clinic staff. Table 1 lists the items chosen from the NEI-VFQ along with the corresponding subscales.

To meet the time constraints of the clinic, yet cover the topics that were deemed important by the low-vision service, 25 items from the field test version of the NEI-VFQ were omitted. The omitted subscales were ocular pain (items 6 and 12), general vision (items 3 and 14), driving (items 35–38, very few of the patients reported that they drove), peripheral vision (item 23), and color vision (item 27). Twelve items were omitted because they were judged to be redundant with other items (4, 5, 10, 30, 39b, 39c, 39d, 39e, 39f, 40, 42, and 45), and two items (7 and 11) were omitted because too many patients required further explanation.

4.3. Rasch analysis

The patient responses to the 27 NEI-VFQ items were analyzed with BIGSTEPS (Linacre & Wright, 1997), an iterative computer program that estimates α for each patient and ρ for each item in logit units using the Masters–Andrich modification of the Rasch model for polytomous responses. This program employs a maximum unconditional likelihood estimation procedure with a correction for bias and returns model fit statistics in addition to parameter estimates (Wright & Masters, 1982).

5. Results

The initial analysis used patient responses to all 27 items. Table 2 lists the logit item measure, ρ_i, and the standard error of the estimate for each item. Positive logit values indicate that the item requires greater visual ability than required by the average of the items, and negative logit values indicate that the item requires less visual ability than the average (note that BIGSTEPS reports item and person measures with signs opposite to those reported in this study). Thus, question 44, 'people know too much about my personal business', requires the least amount of visual ability for patients to disagree with the item ($\rho = -0.83$), and question 39a, 'Accomplish less than you would have liked', requires the most visual ability for patients to disagree with the item ($\rho = 1.14$).

If any of the items are sensitive to more than one variable that is distributed in the patient sample, then the pattern of responses to those items will appear noisy or outlying relative to model expectations. Noise is assessed with the information-weighted fit statistic ('infit') which is the ratio of the mean (across patients) squared response residuals (relative to responses expected by the model) to the mean (across patients) squared residuals expected by the model. Outlying items are detected with the outlier-sensitive fit statistic ('outfit'), which is the mean ratio of the squared patient response residuals to the expected squared patient response residuals. These two weighted mean-square fit statistics can be normalized and expressed in model standard deviation units (Wright & Masters, 1982; Smith, 1986, 1991). The normalized item fit statistics are presented in Table 2 in the columns labeled infit $z$std and outfit $z$std. The expected values are 0, with a tolerance of ± 2 standard deviation units. Positive $z$std values indicate that response residuals exceed the expectations of the model, which means that the responses to the item are inconsistent with the assumptions of the model. Negative values indicate that response residuals are less than the expectations of the model, which implies some strong source of covariance that is shepherding item responses toward the expected value.

As illustrated graphically in Fig. 1, all but seven of the 27 items fall outside the fit statistics tolerance box of $\pm 2z$std. The most misfitting items (those where the infit or outfit $z\text{std} > 3$) are 'less privacy', 'rely on others', 'expect blindness', 'stay home', 'eyesight worse', 'frustrated', and 'accomplish less' (questions 44, 46, 9, 41, 8, 13, and 39a). All of these misfitting items use response categories that refer to patients' agreement with the truth or frequency of applicability of the item statement. All but two of the items that require patient ratings of the level of difficulty of the activity described in the item fall in the fit statistics tolerance box, or have negative infit and outfit $z$std values (filled circles in Fig. 1). Thus, in general, there is greater variability in responses, relative to model expectations, for items that require response ratings other than level of difficulty (open circles in Fig. 1) and less variability for items that require a difficulty rating (filled circles).

The results illustrated in Fig. 1 suggest that the different items in the NEI-VFQ are sampling a variety of uncorrelated variables distributed in the low-vision patient sample. This conclusion is consistent with the correlation matrix published in an earlier validation study (Mangione et al., 1998b). The items represented by the filled symbols in Fig. 1 are assigned to the near vision, distance vision, and social functioning domains. The intercorrelations among these domains in that earlier validation study range from 0.75 to 0.85. Thus, we can infer that, most likely, only one variable is sampled by the items that ask the patient to rate the level of difficulty of performing an activity. The other items represent the general health, mental health, vision ex-

Table 1
Items selected from the 52-item NEI-VFQ that were used in the present study[a]

Number	Item	Response	Domain
1	In general would you say your overall health is:	Quality	General health
8	I expect my eyesight to get worse than it is now	Agreement	Vision expectation
9	I expect to be completely blind at some time in the future	Agreement	Vision expectation
13	How much of the time do you feel frustrated because of your eyesight?	Frequency	Mental health
15	How much difficulty do you have reading ordinary print in newspapers?	Difficulty	Near vision
16	Wearing glasses, how much difficulty do you have reading the small print in a telephone book, on a medicine bottle, or on legal forms?	Difficulty	Near vision
17	How much difficulty do you have doing work or hobbies that require you to see well up close, such as cooking, sewing, fixing things around the house or using hand tools?	Difficulty	Near vision
18	Because of your eyesight, how much difficulty do you have playing cards or games like bingo or Monopoly?	Difficulty	Near vision
19	Because of your eyesight, how much difficulty do you have finding something on a crowded shelf?	Difficulty	Near vision
20	How much difficulty do you have reading street signs or the names of stores?	Difficulty	Distance vision
21	Because of your eyesight, how much difficulty do you have going down steps, stairs or curbs in the daytime	Difficulty	Distance vision
22	Because of your eyesight, how much difficulty do you have going down steps, stairs, or curbs in dim light or at night?	Difficulty	Distance vision
24	Because of your eyesight, how much difficulty do you have recognizing people you know from across a room?	Difficulty	Distance vision
25	Because of your eyesight, how much difficulty do you have seeing how people react to things you say?	Difficulty	Social functioning
26	Because of your eyesight, how much difficulty do you have figuring out whether bills you receive are accurate?	Difficulty	Near vision
28	Because of you eyesight, how much difficulty do you have doing things like shaving, styling your hair, or putting on make-up?	Difficulty	Near vision
29	Because of your eyesight, how much difficulty do you have in performing your normal social activities with family, friends, neighbours or groups (including church activities)?	Difficulty	Social functioning
31	Because of your eyesight, how much difficulty do you have visiting with people you do not know well in their homes, at parties, or in restaurants?	Difficulty	Social functioning
32	Because of your eyesight, how much difficulty do you have seeing and enjoying programs on TV?	Difficulty	Distance vision
33	Because of your eyesight, how much difficulty do you have taking part in active sports or other outdoor activities that you enjoy (like golf, bowling, jogging, or walking)?	Difficulty	Distance vision
34	Because of your eyesight, how much difficulty do you have going out to see movies, plays or sports events?	Difficulty	Distance vision
39a	Do you accomplish less than you would have liked?	Frequency	Role functioning
41	I stay home most of the time because of my eyesight	Aureement	Dependency
43	I have much less control over what I do because of my eyesight	Agreement	Mental health
44	Because of my eyesight, other people know too much about my personal business	Agreement	Dependency
46	Because of my eyesight, I have to rely too much on what other people tell me	Agreement	Dependency
47	I need a lot of help from others, because of my eyesight	Agreement	Dependency

[a] First column, item number in the 1995 field test version of the NEI-VFQ; second column, item as read to the patient; third column, type of response required of each item. Quality was rated on a five-point scale ranging from 'excellent' (1) to 'poor' (5). Agreement was rated on a five-point scale ranging from 'definitely true' (1) to 'definitely false' (5). Frequency was rated on a five-point scale ranging from 'all of the time' (1) to 'none of the time' (5). Difficulty was rated on the following five-point scale: 'no difficulty at all' (1), 'a little difficulty' (2), 'moderate difficulty' (3), 'extreme difficulty' (4), or 'stopped doing this because of your eyesight' (5). A sixth possible response on the difficulty scale was 'stopped doing this for other reasons or not interested in doing this', which was scored as missing data. The final column lists the domain that each item was assigned to by the instrument developers.

Table 2
Estimates of item measures and fit statistics from Rasch analysis applied to patient responses to all 27 items[a]

Item	Item number	ρ	Standard error	Infit $z$std	Outfit $z$std
Less privacy	44	−0.83	0.05	5.5	3.2
Social activities	29	−0.78	0.05	−3.7	−3.2
Rely on others	46	−0.65	0.05	4.6	2.6
Unfamiliar people	31	−0.63	0.05	−4.6	−3.8
Shaving or make-up	28	−0.48	0.05	−3.7	−3.1
Sports and outdoors	33	−0.48	0.05	−2.4	−2.0
Steps daytime	21	−0.47	0.05	−5.0	−3.0
Expect blindness	9	−0.44	0.05	3.6	7.2
Steps night	22	−0.36	0.05	−4.2	−2.5
TV programs	32	−0.26	0.05	−7.3	−5.7
Crowded shelf	19	−0.22	0.05	−5.2	−4.3
General health	1	−0.15	0.05	−2.4	0.0
Stay home	41	−0.02	0.05	8.7	8.1
Help from others	47	0.03	0.05	−0.8	−0.7
Movies or plays	34	0.05	0.06	−1.2	−1.6
Work or hobbies	17	0.06	0.05	−2.3	−2.4
Play games	18	0.13	0.05	−1.0	−1.4
Eyesight worse	8	0.15	0.05	9.9	9.9
Gauge reactions	25	0.3	0.05	2.5	1.7
Less control	43	0.36	0.05	2.2	0.4
Accurate bills	26	0.38	0.05	2.5	1.0
Recognize people	24	0.41	0.05	1.0	1.1
Street signs	20	0.45	0.05	−1.3	−0.7
Frustrated	13	0.47	0.05	5.0	3.6
Read ordinary print	15	0.88	0.06	−0.8	−1.5
Read small print	16	0.96	0.06	0.1	−1.0
Accomplish less	39a	1.14	0.07	3.6	2.1

[a] First column, item; second column, item number (see Table 1 for complete item description); third column, estimate of the visual ability required by the item (ρ); fourth column, standard error of the estimate of the item measure; fifth and sixth columns, information weighted (infit) and outlier sensitive (outfit) fit statistics expressed as normalized residuals.

pectations, role functioning, and dependency domains. General health and vision expectations correlate poorly with every other domain ($-0.05 \leq r \leq 0.30$), whereas the role functioning, mental health, and dependency domains correlate highly with each other and with the near vision, distance vision, and social functioning domains ($0.64 \leq r \leq 0.81$).

5.1. Analysis of NEI-VFQ items requiring difficulty ratings

Because of gross misfits to the model by many of the items that used agreement instead of difficulty rating scales, and because of observations from the interdomain correlations reported earlier that items in the general health and vision expectation domains are sampling uncorrelated variables, we excluded all ten items that required responses other than ratings of the level of difficulty to perform an activity (open circles in Fig. 1). The edited data were then re-analyzed with BIGSTEPS. The resulting item measures and fit statistics are displayed in Table 3.

As shown in Fig. 2, the item measures estimated from the edited data set are proportional to the item measures estimated with all the data included. The regression line has a slope of 1.34. This greater than unity slope can be interpreted as an increase in the steepness of the item response function for the edited data set (i.e. an increase in the value of a in Eq. (1) for all items). A steeper item response function implies greater precision in the estimate of the item parameter. This interpretation is confirmed by taking the ratios of the standard errors of the estimates, when normalized to be in the same units (i.e. divide the standard error for the edited data by 1.34). The standard error of the item measure for the edited data is 90% of the standard error for all the data.

Fig. 3 illustrates the infit and outfit values for the revised list of items in Table 3. All but three of the 17 items fall within, or very close to the $\pm 2z$std tolerance box. The most misfitting items are 'figuring out whether bills you receive are accurate' (item 26) and 'seeing how people react to things you say' (item 25). The large infit values, 5.3 and 3.8$z$std, respectively, indicate inconsistent responses to these items relative to patient response patterns to the other items. The outfit values for these two items are close to 2$z$std, suggesting that the misfit of these items can be attributed more to noise, than to extreme anomalous responses by a subset of patients. Such noise might be attributed to variation among

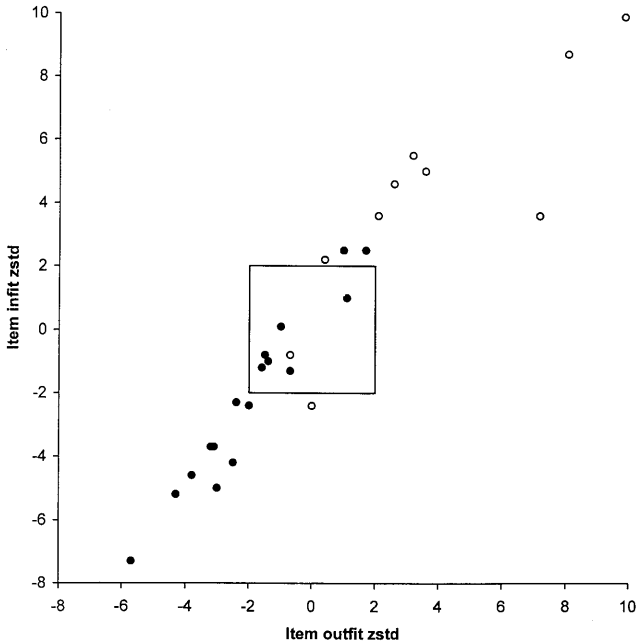

Fig. 1. Scatter plot of the information weighted (infit) and outlier sensitive (outfit) fit statistics for estimates of item measures. The infit and outfit values are expressed as z scores for the normalized distribution of response residuals (relative to Rasch model expectations). ●, items that require difficulty ratings; ○, items that require frequency or level of agreement ratings. The square bounds the 95% confidence limits ($\pm$2S.D.).

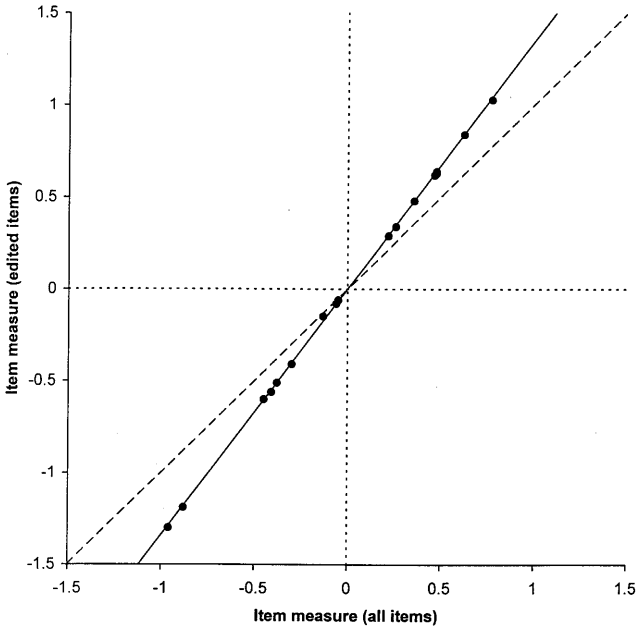

Fig. 2. Scatter plot of item measures estimated from responses to only the 17 items that required difficulty ratings (see Table 3) versus item measures estimated from responses to all 27 items (see Table 2). The regression line (solid line) passes through the origin, but is steeper than the identity line (dashed line) with a slope of 1.34.

patients in their interpretation of these questions and/or their dependence on vision to successfully perform these activities. One item, 'seeing and enjoying programs on TV' (item 32), fell well outside the expectations of the model at the other extreme. Both the infit and outfit values indicated that response residuals were much lower than those expected by the model. Lower than expected response variability suggests that responses to this item are strongly governed by a co-factor in the patient sample, such as visual acuity, and are less sensitive than other items to the effects of modifying variables, such as behavior modification to adapt to vision loss.

The item responses across patients was nearly evenly distributed for the five rating categories. Table 4 illustrates that each category was used an average of 1057

Table 3
Estimates of item measures and fit statistics from Rasch analysis restricted to response to the 17 items selected from the NEI-VFQ that required difficulty ratings

Item	Item number	ρ	Standard error	Infit $z$std	Outfit $z$std
Social activities	29	−1.03	0.06	−2.2	−2.1
Unfamiliar people	31	−0.84	0.06	−2.5	−2.5
Sports and outdoors	33	−0.64	0.06	0.1	0.3
Shaving or make-up	28	−0.63	0.05	−0.5	−0.8
Steps daytime	21	−0.62	0.05	−2.5	0.1
Steps night	22	−0.48	0.06	−1.0	1.7
TV programs	32	−0.34	0.05	−5.2	−3.5
Crowded shelf	19	−0.29	0.05	−2.3	−1.5
Movies or plays	34	0.06	0.07	1.6	0.4
Work or hobbies	17	0.08	0.05	2.7	1.7
Play games	18	0.15	0.06	2	1.2
Gauge reactions	25	0.41	0.06	3.8	2.3
Accurate bills	26	0.51	0.06	5.3	2.8
Recognize people	24	0.56	0.06	2.1	1.4
Read street signs	20	0.6	0.06	0.8	1.0
Read ordinary print	15	1.19	0.07	1.5	−0.2
Read small print	16	1.3	0.07	2.3	−0.1

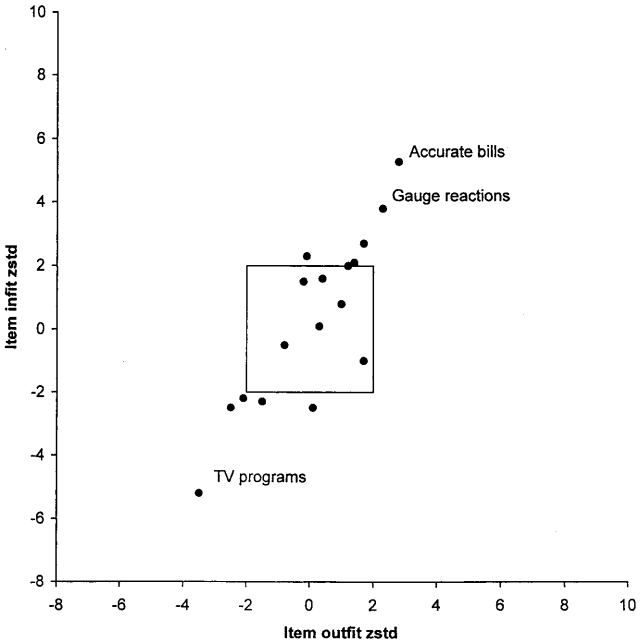

Fig. 3. Scatter plot of infit and outfit normalized squared and weighted residuals for item measures estimated from responses to the 17 items that required difficulty ratings. The square bounds the 95% confidence limits.

times. The infit and outfit $z$std values in the third and fourth columns of Table 4 indicate the use of all five response categories was consistent with the expectations of the model.

Within the framework of the Andrich–Masters modification of the Rasch model, one can think of each item of the NEI-VFQ as requiring a specific level of visual ability to elicit a particular rating response, x, from the patient (x ranges from 1 to 5). The required visual ability to elicit response x for item i is ρ_{ix} (Masters, 1982; Wright & Masters, 1982). In the Andrich (1978) model, $\rho_{ix} = \rho_i + \tau_x$, where τ_x refers to a criterion difference from the required visual ability for the item that is necessary to elicit rating x. Returning to Eq. (2), when $\alpha = \rho$ the probability of a particular response is 0.5. Thus, we can interpret the criterion

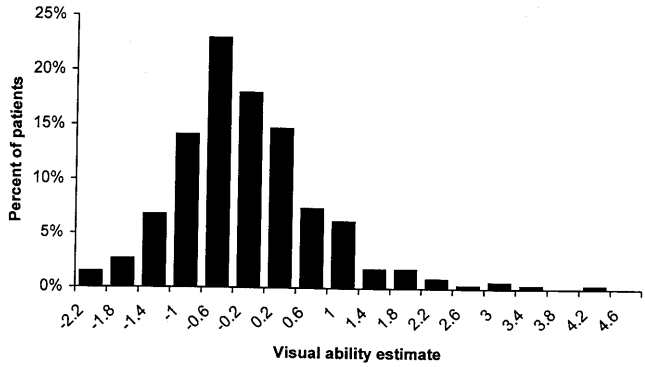

Fig. 4. Histogram of 341 person measures estimated from patient responses to the 17 items that required difficulty ratings.

difference as $\tau_x = \alpha - \rho$, which is the threshold difference between the patient's visual ability and the visual ability required by the item in order to elicit rating x (i.e. $P(x|\alpha,\rho) = 0.5$ when $\alpha - \rho = \tau_x$; see Massof, 1998).

The fifth column of Table 4 lists the estimated values of τ_x for each of the five rating categories. The sixth and seventh columns list the relative rating category boundaries in logit units. Since rating 1 has no lower bound and rating 5 has no upper bound, the values for τ_1 and τ_5 are estimates based on the average use of those categories. These values are enclosed in parentheses to identify them as unbounded estimates.

Fig. 4 illustrates the distribution of visual ability measures, α, for the sample of low-vision patients. This visual ability measure is expressed in logit units, relative to the mean of the required abilities for the 17 NEI-VFQ items. High values of α mean that patients have greater visual ability (are less disabled), and low values correspond to lower visual ability (more disabled). The average standard error of the estimate of α is 0.28 with a range in the standard error of 0.23–1.4.

The distributions of the infit and outfit $z$std statistics for estimates of α are presented in Fig. 5. Each data point represents the infit and outfit $z$std values for an individual patient. Eleven percent of the low-vision patients have fit statistics that fall outside the ± 2std

Table 4
Results of the analysis of the use of response categories for rating difficulty[a]

Response	Frequency	Infit $z$std	Outfit $z$std	τ_x	τ_{max}	τ_{min}
1	1071	1.7	1.55	(2.1)	∞	1.46
2	911	0.82	0.69	0.08	1.46	0.38
3	989	−0.68	−0.83	0.02	0.38	−0.35
4	1096	0.45	−0.18	−0.79	−0.35	−1.48
5	1216	−1.44	−0.94	(−2.15)	−1.48	$-\infty$

[a] First column, rating used by the patient (see caption to Table 1 for details); second column, frequency in the database that each response category was used; third and fourth columns, normalized weighted residuals (relative to model expectations) for the use of each rating category; fifth column, logit values for each response category that, when added to the item measure, define the item-response measure (i.e. ρ_{ix}). The values in parentheses are estimates based the average use of those categories since the extreme categories are unbounded. The final two columns list the relative logit values of the category boundaries.

tolerance box. The response patterns of these patients to the 17 NEI-VFQ items were inconsistent with the expectations of the model. A retrospective review of the histories of the 15 most misfitting patients (infit or outfit $z\text{std} > 3$) revealed that six of the patients were successful low-vision device users. These six patients had received previous low-vision rehabilitation services at another clinic. Despite poor visual acuities, these six patients reported that they had no difficulty on the items that asked about reading-related activities. Another three patients had central islands of vision with relatively good visual acuity that were surrounded by annular scotomas or severely restricted visual fields. These patients had difficulty with items that patients with worse acuity found easy, and vice versa. The remaining six misfitting patients had nothing in their histories that might account for their statistically anomalous responses to the items. Eliminating these 15 most misfitting patients from the database does not influence the estimation of item measures (ρ) or person measures (α).

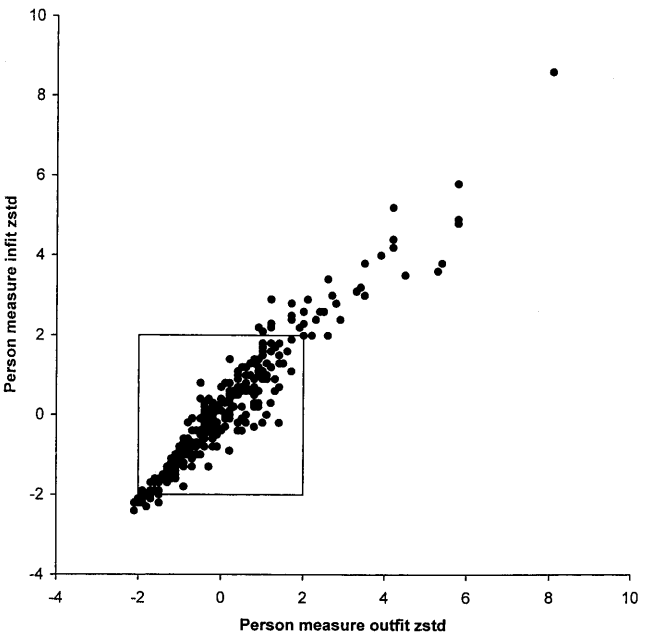

Fig. 5. Scatter plot of infit and outfit normalized fit statistics for estimates of person measures. Each point represents a different patient. The square bounds the 95% confidence limits.

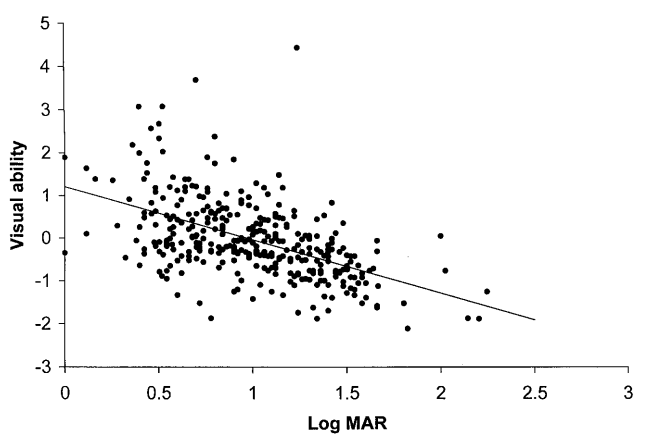

Fig. 6. Scatter plot of person measure estimates based on patient responses to the 17 items that required difficulty ratings vesrus corrected binocular visual acuity. Visual acuity is expressed as the logarithm of the minimum angle of resolution (log MAR): log MAR = 0 corresponds to a Snellen acuity of 20/20, log MAR = 1 corresponds to 20/200. The regression line (solid line) has a slope of −1.24 and an intercept of 1.2. The Pearson correlation coefficient is 0.523.

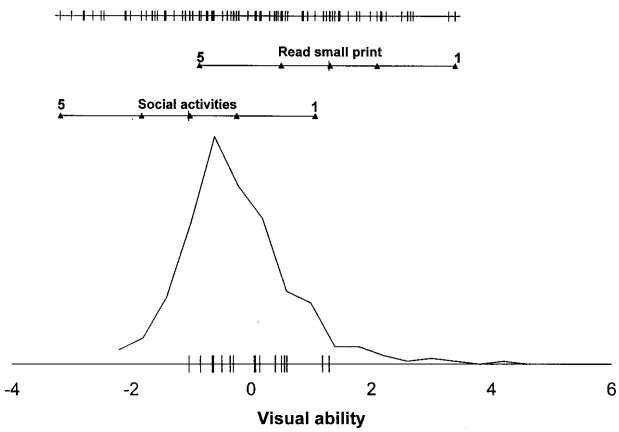

Fig. 7. Summary of the Rasch analysis estimates of an interval visual ability scale from the responses of low-vision patients to 17 items on the NEI-VFQ that require ratings of difficulty. The axis at the bottom of the figure is the interval visual ability logit scale. The tick marks on the axis are estimates of item measures (ρ) for each of the 17 items. The curve above the axis is the distribution of estimated person measures (same as the histogram in Fig. 4). The axis labeled 'Social activities' illustrates the range and logit values of difficulty ratings (1–5, represented by triangles; see Table 4) to the item requiring the least amount of visual ability (minimum value of ρ). The axis labeled 'Read small print' illustrates the range and logit values of difficulty ratings for the item requiring the most visual ability (maximum value of ρ). The tick marks on the axis at the top of the figure represent the logit values for every possible rating response for every item.

Visual acuity is the most obvious difference among patients that might account for the person measure distribution in Fig. 4. This expectation is confirmed by the scatter plot in Fig. 6 that illustrates a strong linear trend between visual acuity, expressed as the logarithm of the minimum angle of resolution (log MAR), and person measures of visual ability (α). The Pearson correlation is 0.523 and the slope and intercept of the regression line are −1.24 and 1.2, respectively. This slope translates to 0.12 logit per line of visual acuity measured on the ETDRS Bailey–Lovie chart (Ferris, Kassoff, Bresnick, & Bailey, 1982). The average standard error of the person measure is 0.21 logit. Thus, the 95% confidence interval for a difference in person measures corresponds to a three-line difference in visual acuity.

Fig. 7 summarizes the analysis of the responses of low-vision patients to the 17 items selected from the NEI-VFQ that require ratings of the level of difficulty the patient has in performing the activity. The axis at the bottom of the figure is the interval visual ability scale in logit units. The tick marks represent the required visual ability for each of the 17 items. Note that there is a large gap between read ordinary print ($\rho = 1.19$) and read street signs ($\rho = 0.6$). Also note that the required visual abilities for movies or plays ($\rho = 0.06$) and work or hobbies ($\rho = 0.08$) are nearly coincident, as are the required visual abilities for sports and outdoors ($\rho = -0.64$), shaving or make-up ($\rho = -0.63$), and steps in the daytime ($\rho = -0.62$).

The curve above the axis in Fig. 7 is the frequency distribution of visual ability measures for the 341 low-vision patients (same as the histogram in Fig. 4). The scales above the frequency distribution are the positions of the different rating categories for the item requiring the most visual ability (read small print) and for the item requiring the least visual ability (social activities). Using the terminology of the Andrich model, each triangle represents τ_x (from Table 4) added to ρ_i (from Table 2). The range of visual ability values from the triangle labeled 5 for social activities (-3.18) to the triangle labeled 1 for read small print (3.4) defines the measurement range of the 17 items chosen from the NEI-VFQ. This range nearly spans the range of person measures for this sample of low-vision patients.

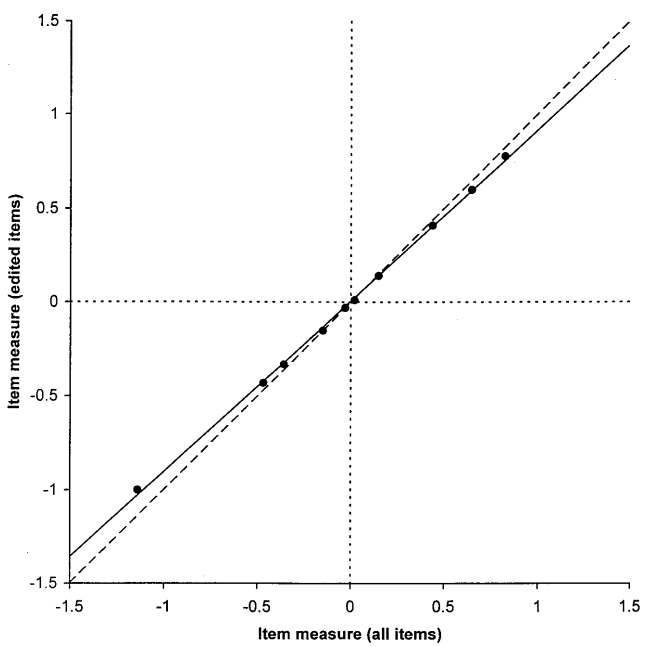

Fig. 8. Scatter plot of item measures estimated from patient responses to the ten items that required frequency or level of agreement ratings vesrus item measures estimated from patient responses to all 27 items. The regression line (solid line) passes through the origin and is shallower than the identity line (dashed line) with a slope of 0.91.

The tick marks on the axis at the top of Fig. 7 illustrate the positions on the visual ability scale of every possible response to each of the 17 items. It would appear that the 17 items chosen from the NEI-VFQ would produce a very precise measurement scale. However, estimations of measurement precision must also take into consideration the measurement error. If we use a criterion of three standard errors as a significant difference in measurements, then the ratings to the17 items can resolve 15 statistically significant steps in the scale ($G = 10.72$, $r = 0.99$; see Wright & Masters, 1982 for an interpretation of the separation [G] and reliability [r] indices). For this particular sample of patients, the person measure frequency distribution can be divided into four groups that have statistically significant differences in visual ability values ($G = 2.71$, $r = 0.88$).

5.2. Analysis of responses to the items edited from the data set

It is conceivable that responses to the items that were edited from the data set, which required responses other than difficulty ratings (open circles in Fig. 1), isolate another variable distributed within the patient sample. To test that hypothesis, we applied Rasch analysis to the patient responses to those ten items.

Fig. 8 illustrates that the item measures estimated when applying Rasch analysis to these ten items alone are proportional to the item measures estimated when the analysis was performed on all 27 items. The proportionality constant of 0.91 indicates that these ten items were a major source of variability in the original item measure estimates (i.e. the slopes of the item response functions for these items would be shallow). This result is consistent with the hypothesis that patient responses to these items are governed to a large extent by another latent variable(s) distributed in the patient sample.

The scatter plot in Fig. 9 illustrates the distribution of infit and outfit values for these ten items when analyzed separately. The most misfitting items are numbers 8 (eyesight worse), 9 (expect blindness), and 41 (stay home). The response patterns to items 47 (help from others), 43 (less control), and 1 (general health) were less variable than expected by the model (i.e. $z\text{std} < -2$). The four items that fell within the $\pm 2z\text{std}$ tolerance box (items 44 (less privacy), 46 (rely on others), 13 (frustrated), and 39a (accomplish less)) were closer to the margins of too little (items 13, 44, and 46) or too much (item 39a) variability. These observations are inconsistent with the hypothesis that patient responses to these ten items represent another single variable separate from that estimated from the responses to the 17 items listed in Table 3. Rather, it appears that these ten items are differentially sensitive to several different variables distributed in the sample of low-vision patients.

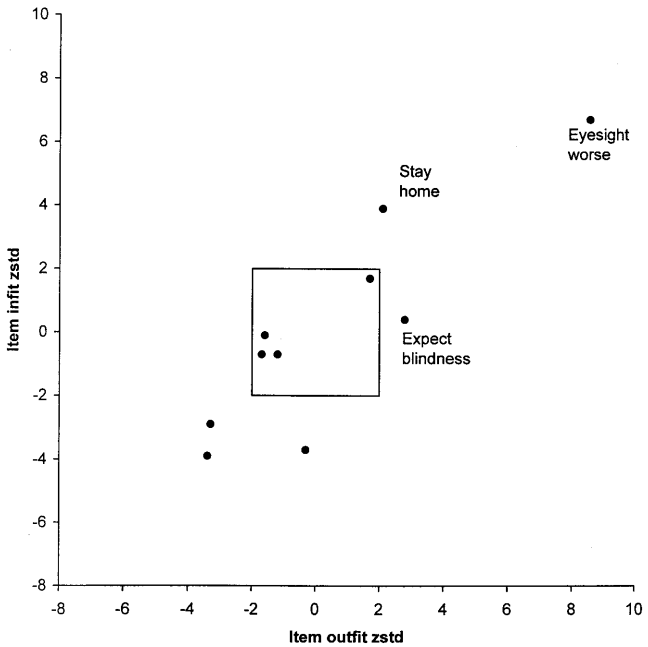

Fig. 9. Scatter plot of infit and outfit normalized fit statistics for item measures estimated from responses to the ten items that required frequency or level of agreement ratings. The square bounds the 95% confidence limits.

5.3. Comparison with other instruments

Based on the sample of items used in this study, we conclude that those items in the NEI-VFQ that require difficulty ratings can be used to generate a valid interval visual ability scale for low-vision patients. The items requiring difficulty ratings are those that constitute the near vision, distance vision, and social functioning domains. These three domains do not represent three different variables. Rather, all of the items in these domains can be used to estimate a single latent variable that is highly dependent on visual acuity (see regression in Fig. 6).

The ALB developed by Becker et al. (1985) and the ABS developed by Massof (1998) also required subjects to respond with difficulty ratings for items that described daily activities. Both studies estimated interval visual ability scales from responses of low-vision patients. In principle, the interval scales estimated from the two published data sets and the present data should be linear transformations of one another since the three instruments use the same types of responses and were applied to similar patient samples. The origins of the three interval scales may differ because the origin is arbitrarily set to the mean of the estimated item measures for each instrument. Also, the slopes of the psychometric item response functions for the three instruments may differ, depending on response variability to items not common to the three instruments (similar to the explanation of the relationships illustrated in Figs. 2–8).

Fig. 10a illustrates the relationship of item measures estimated from the ABS used in Massof's study to item measures estimated in the present study from responses to the NEI-VFQ items that required difficulty ratings. Each data point represents items on the two instruments that were similar in content. The slope of the regression line is 0.89 and the intercept is 0. Fig. 10b illustrates the relationship between estimated item measures from the present study for items from the NEI-VFQ that required difficulty ratings, and items with similar content in the ALB used in the Becker et al. (1985) study. The slope of the regression line is 1.49 and the intercept is 0.29. Finally, Fig. 10c illustrates the relationship between item measures for items with similar content in the ALB from the Becker et al. study and the ABS from the Massof (1998) study. The line drawn through the data was estimated from the regression lines for the ABS versus NEI-VFQ (Fig. 10a) and the ALB versus the NEI-VFQ (Fig. 10b). The slope is 1.34 and the intercept is 0.29. These linear relationships allow us to construct a common visual ability scale for the three studies.

Fig. 11 illustrates the items used in the three studies plotted on a common visual ability scale. Samples of items on the instruments that have similar content are labeled. There is strong agreement among the three instruments. The differences among instruments are in the distribution of intervals between items and the instrument's range on the scale.

6. Discussion

The present study demonstrates that items in part 2 of the NEI-VFQ can be used to estimate an interval scale of visual ability for patients with low vision. This latent visual ability variable is strongly related to visual acuity. Items from parts 1 and 3 of the NEI-VFQ can also be positioned on this scale, but patient responses to those items are confounded by other latent variables, such as coping ability, that increase noise and decrease the validity of the estimate of the item measure. The visual ability scale estimated from low-vision patient responses to part 2 of the NEI-VFQ is the same as the visual ability scales estimated from low-vision patient responses to the ABS and the ALB.

Despite the pioneering work of Becker et al. (1985), the application of item response theory and Rasch models to visual function assessments is still a new concept. The development of the NEI-VFQ followed an approach that might be considered traditional in the development of health care instruments (McDowell & Newell, 1996). This traditional approach accepts raw scores at face value and relies heavily on inferential arguments for validating scales. To a limited extent, it adheres to the tenets of classical test theory (Lord,

1980), but falls short by allowing respondents to answer items in part 2 with a response that is scored as missing data (i.e. 'stopped doing this for other reasons or not interested in doing this'). Even if it could be demonstrated that the instrument score orders patients according to the latent variable of interest, missing data distort the score and make it uninterpretable for individual patients. Also, as measures, raw scores are nonlinearly related to the variable of interest. That is, even if it could be demonstrated that the ordinal response ratings are spaced at even intervals, the extreme response categories (e.g. 1 and 5) are open-ended. The use of these response categories contribute to floor and ceiling effects in the data when they are averaged with responses to other items.

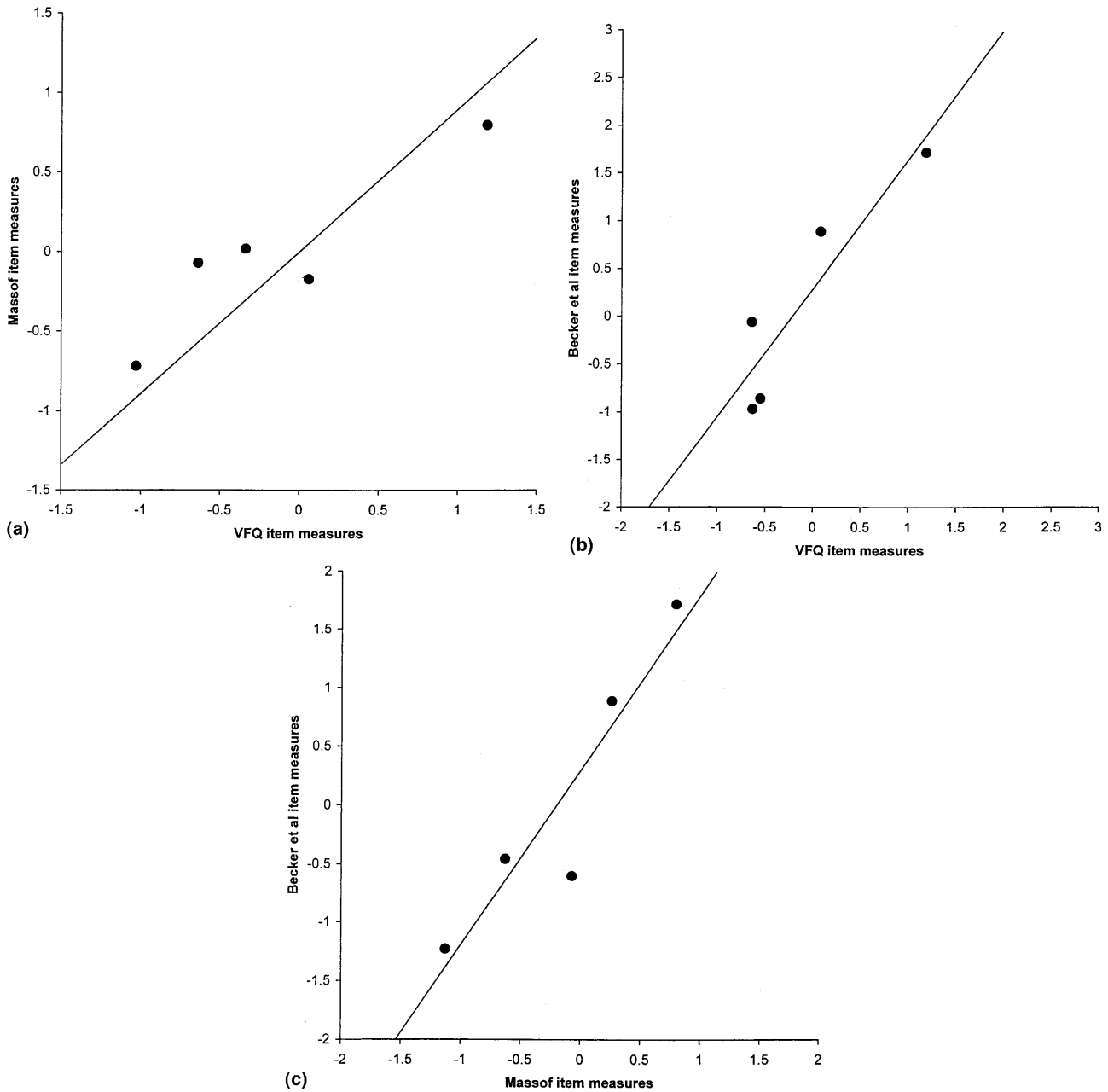

Fig. 10. Scatter plots comparing item measures for similar items estimated from low-vision patient responses with different instruments on different samples of low-vision patients. (a) Comparison of item measures estimated in Massof's study using the ABS to item measures for similar items estimated from the NEI-VFQ in the present study. The regression line (solid line) has a slope of 0.89 and passes through the origin. (b) Comparison of item measures estimated in the Becker et al. study using the ALB to item measures estimated from the NEI-VFQ in the present study. The slope of the regression line (solid line) is 1.49 and the intercept is 0.29. (c) Comparison of item measures estimated in the Becker et al. study using the ALB to item measures estimated in Massof's study using the ABS. The line through the data was estimated from the other two regression lines. The slope is 1.34 and the intercept is 0.29.

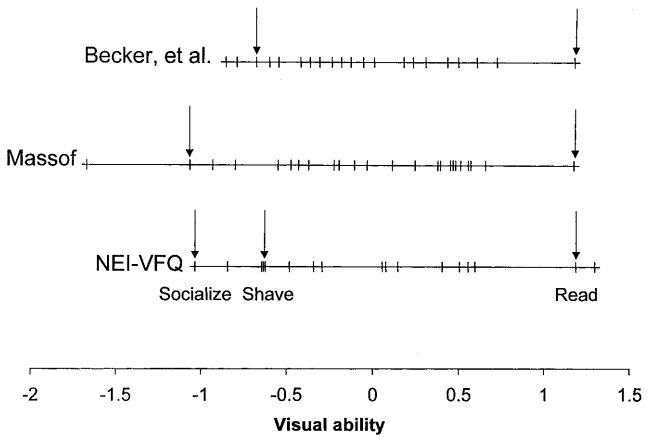

Fig. 11. Item measures estimated from low-vision patient difficulty ratings in the Becker et al. study using the ALB, Massof's study using the ABS, and the present study using the NEI-VFQ. The common visual ability scale was constructed from recalibrations of the original scales using the linear regressions illustrated in Fig. 10a–c. Samples of common items are marked with arrows and labeled.

Perhaps the greatest limitation of the traditional scoring method is that all of the items are given equal weight. A response of 'extreme difficulty' to an easy item carries the same weight as the same response to a difficult item; both are given a score of 4. However, in terms of the ability of the patient, those two responses are not equivalent. For the hard item, a score of 4 means hard things are hard to do (e.g. read small print), something that might be true even for a patient with mild visual impairment. For the easy item, a score of 4 means that easy things are hard to do (e.g. taking part in active sports and other outdoor activities), something that might only be true for patients with more severe visual impairment. Traditional scoring implicitly treats every item as equally difficult.

Rasch analysis assumes that only one variable is measured by the instrument. The NEI-VFQ was designed to make measurements in 13 different domains. The recommended scoring procedure generates 13 different scores. Although one could argue the value of making 13 different measurements on each patient, it is questionable whether this single multidimensional instrument is adequate, especially when only one to three items per domain are used to estimate half of the domain scores. Furthermore, the domains are not defined by the items, they are defined by the patients. Looking at the interdomain correlation matrix in Mangione et al. (1998b), it appears that at most there are four different factors: general health, ocular pain, vision expectations, and everything else. This expectation can be confirmed by performing a factor analysis on their correlation matrix.

Rasch analysis on our data demonstrated that the items requiring difficulty ratings (i.e. part 2) work together to measure a single latent variable (see Fig. 1), despite representing three different domains. The items requiring frequency or agreement ratings also can be used to estimate that single latent variable, but with less precision (see Fig. 8). These items most likely are more sensitive than the part 2 items to other variables distributed in the sample of patients, such as psychosocial traits. Including these items with the part 2 items reduces measurement precision and accuracy.

If patients differed only in visual acuity, then differences among patients in responses to the items would have to be attributed to the visual acuity differences. In this idealized case, even differences in opinions about the weather would produce an estimate of visual acuity. Other differences among patients, such as personality traits and living conditions, might differentially contribute to their responses to different items, but these conditions might not be as well distributed in the target population as is visual acuity. In this case, these other attributes, including other types of visual impairments such as visual field loss, would be manifested as contributing to errors in the measurement and would elevate the mean square fit statistic used to evaluate fits to Rasch models. For these other attributes to emerge as domains, they would have to rival visual acuity in their contributions to between-patient response variability. The strong proportionality and high correlation between log MAR and visual ability (Fig. 6) suggests that visual acuity is a strong factor for responses to items in part 2 for the low-vision clinic population. In effect, one could interpret the required ability for item 26, for example, as the average log MAR acuity that is required to figure out whether bills received are accurate. This result suggests that we may end up traveling full circle. That is, once we understand the relationship between functional limitations and visual acuity, visual acuity measures could be sufficient.

Rasch analysis offers an alternative to traditional scoring methods that enables one to estimate the latent variable of interest and to assess the performance of each item as a contributor to the measurement. Within this framework, items become important, not instruments. The appropriate strategy within the Rasch framework would be to calibrate individual items (i.e. estimate item measures for different target populations) and build an item bank. This strategy could lead to adaptive testing, e.g. a staircase-type procedure, using items that are meaningful to the individual patient. In other words, each patient could take a custom test that would precisely measure his/her functional capability in a common unit, even though different patients would respond to different items. The comparison of different instruments that is summarized in Fig. 11 illustrates that such item calibrations could be accomplished for visual function assessments.

Acknowledgements

This work was supported by a grant from the Multiple District 22 Lions Vision Research Foundation, Grant EY12045 of the National Eye Institute, National Institutes of Health, and by Retina Consultants of Southwest Florida. Edie Stern edited the manuscript.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Andersen, E. B. (1995). What Georg Rasch would have thought about this book. In G. H. Fischer, & I. W. Molenaar, *Rasch models: foundations*, *recent developments*, *and applications* (pp. 383–390). New York: Springer-Verlag.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Becker, S. W., Lambert, R. W., Schulz, E. M., Wright, B. D., & Burnet, D. L. (1985). An instrument to measure the activity level of the blind. *International Journal of Rehabilitation Research*, *8*, 415–424.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gibson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care*, *19*, 787–805.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Cataract Management Guidelines Panel. *Cataract in adults: management of functional impairment: clinical practice guideline number 4*. Rockville, MD: US Department of Health and Human Services, PublicHealth Service, Agency for Health Care Policy and Research. AHCPR Publication No. 93-0542, February 1993.

Elliott, D. B., Hurst, M. A., & Weatherill, J. (1990). Comparing clinical tests of visual function in cataract with the patient's perceived visual disability. *Eye*, *4*, 712–717.

Ellwein, L. B., Fletcher, A., Negrel, A. D., & Thulasiraj, R. D. (1995). Quality of life assessment in blindness prevention interventions. *International Ophthalmology*, *18*, 263–268.

Ferris, F. L., Kassoff, A., Bresnick, G. H., & Bailey, I. (1982). New visual acuity charts for clinical research. *American Journal of Ophthalmology*, *94*, 91–96.

Fisher, W. P., Jr, Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: a common language of functional assessment. *Archives of Physical Medicine and Rehabilitation*, *76*, 113–122.

Granger, C. V., & Hamilton, B. B. (1993). The Uniform Data System for Medical Rehabilitation report of first admissions for 1991. *American Journal of Physical Medicine and Rehabilitation*, *72*, 33–38.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction*. Princeton, NJ: Princeton University Press.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, *14*, 75–96.

Kupfer, C. (1996). The expanded role of randomized clinical trials. *American Journal of Ophthalmology*, *122*(6), 883–885.

Linacre, J. M., & Wright, B. D. (1997). *A user's guide to BIGSTEPS Rasch-model computer program*. Chicago, IL: MESA Press.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.

Macular Photocoagulation Study Group (1991). Laser photocoagulation of subfoveal neovascular lesions in age-related macular degeneration: results of a randomized clinical trial. *Archives of Ophthalmology*, *109*, 1220–1231.

Mangione, C. M., Berry, S., Spritzer, K., Janz, N. K., Klein, R., Owsley, C., & Lee, P. P. (1998a). Identifying the content area for the 51-item National Eye Institute Visual Function Questionnaire. *Archives of Ophthalmology*, *116*, 227–233.

Mangione, C. M., Lee, P. P., Pitts, J., Gutierrez, P., Berry, S., & Hays, R. D. (1998b). Psychometric properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). *Archives of Ophthalmology*, *116*, 1496–1504.

Mangione, C. M., Phillips, R. S., Seddon, J. M., Lawrence, M. G., Cook, E. F., Dailey, R., & Goldman, L. (1992). Development of the 'Activities of Daily Vision Scale': a measure of visual functional status. *Medical Care*, *30*, 1111–1126.

Massof, R. W. (1998). A systems model for low vision rehabilitation. II. Measurement of vision disabilities. *Optometry and Vision Science*, *75*, 349–373.

Massof, R.W., Rubin, G.S. (2000). Visual function assessment questionnaires. Survey of Ophthalmology 2000 (in press).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McDowell, I., & Newell, C. (1996). *Measuring health: a guide to rating scales and questionnaires* (2nd ed.). New York: Oxford University Press.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer, & I. W. Molenaar, *Rasch models: foundations*, *recent developments*, *and applications* (pp. 3–14). New York: Springer-Verlag.

Neugarten, B. L., Havighurst, R. J., & Tobin, S. S. (1961). The measurement of life satisfaction. *Journal of Gerontology*, *16*, 134–143.

Parrish, R. K., Gedde, S. J., Scott, I. U., Feuer, W. J., Schiffman, J. C., Mangione, C. M., & Montenegro-Piniella, A. (1997). Visual function and quality of life among patients with glaucoma. *Archives of Ophthalmology*, *115*, 1447–1455.

Pfeffer, R. I., II, Kurosaki, T. T., Harrah, C. H., Chance, J. M., & Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology*, *37*, 323–329.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, *32*, 761–768.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rubin, G. S., Adamsons, I. A., & Stark, W. J. (1993). Comparison of acuity, contrast sensitivity, and disability glare before and after cataraact surgery. *Archives of Ophthalmology*, *111*, 56–61.

Sherbourne, C. D., & Stewart, A. L. (1991). The MOS Social Support Survey. *Social Science and Medicine*, *32*, 705–714.

Shipp, M. D. (1998). Potential human and economic cost-savings attributable to vision testing policies for driver license renewal, 1989–1991. *Optometry and Vision Science*, *75*, 103–118.

Sloane, M. E., Ball, K., Owsley, C., Bruni, J., & Roenker, D. (1992, pp. 26–29). The visual activities questionnaire: developing an instrument for assessing problems in everyday visual tasks. In *OSA technical digest of noninvasive assessment of visual systems*, vol. 1. Washington, DC: Optical Society of America.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*, 359–384.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*, 541–565.

Social Security Administration. Listing of impairments. *20 CFR*. 1997, vol §404.1525.

Stoline, A. M., & Weiner, J. P. (1993). *The new medical marketplace: a physician's guide to the health care system in the 1990s*. Baltimore, MD: Johns Hopkins University Press.

Tielsch, J. M., Sommer, A., Witt, K., et al. (1990). Blindness and visual impairment in an American urban population. The Baltimore Eye Survey. *Archives of Ophthalmology*, *108*, 286–290.

Turano, K. A., Geruschat, D. R., Stahl, J. W., & Massof, R. W. (1999). Perceived visual ability for independent mobility in persons with retinitis pigmentosa. *Investigative Ophthalmology and Visual Science*, *40*, 865–877.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*, 473–483.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*, 857–860.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Education Measurement*, *14*, 97–116.