# Comparison of Scoring Approaches for the NEI VFQ-25 in Low Vision

**Bradley E. Dougherty, OD, MS** **Mark A. Bullimore, MCOptom, PhD, FAAO**

The Ohio State University, College of Optometry, Columbus, Ohio

## Abstract

**Purpose**—The aim of this study was to evaluate different approaches to scoring the NEI VFQ-25 in patients with low vision including: scoring by the standard method, by Rasch analysis, and by use of an algorithm created by Massof to approximate Rasch person measure. Subscale validity and use of a seven-item short form instrument proposed by Ryan et al were also investigated.

**Methods**—NEI VFQ-25 data from 50 patients with low vision were analyzed using the standard method of summing Likert-type scores and calculating an overall average, Rasch analysis using Winsteps software, and the Massof algorithm in Excel. Correlations between scores were calculated. Rasch person separation reliability and other indicators were calculated to determine the validity of the subscales and of the seven-item instrument.

**Results**—Scores calculated using all three methods were highly correlated, but evidence of floor and ceiling effects was found with the standard scoring method. None of the subscales investigated proved valid. The seven-item instrument showed acceptable person separation reliability and good targeting and item performance.

**Conclusions**—Though standard scores and Rasch scores are highly correlated, Rasch analysis has the advantages of eliminating floor and ceiling effects and producing interval-scaled data. The Massof algorithm for approximation of the Rasch person measure performed well in this group of low vision patients. The validity of the subscales VFQ-25 should be reconsidered.

## Keywords

low vision; NEI VFQ-25

The NEI VFQ-25[1] is the most widely used instrument for the assessment of vision-related quality of life. It has been used in many large-scale studies, most frequently using the scoring method published with the original instrument.[2-9] The original published scoring method involves assigning an integer value to the subject's responses, summing those values, and then generating a composite score between 0 and 100 that is meant to be a measure of the subject's visual ability. The published scoring methods also allow for groups of items to be scored as subscales that are intended to be used as an indicator of ability in specific areas such as near visual tasks or distance visual tasks.

This published scoring method has been criticized because it does not produce interval-scaled estimates of visual ability.[10] Standard scoring of Likert-style scale data assumes that each rating category (e.g. little difficulty) has the same value across all items, and the difference between each category is exactly the same. Use of a composite score and

Corresponding author: Bradley E. Dougherty, The Ohio State University, College of Optometry, 338 West 10th Ave, Columbus, OH 43210, bdougherty@optometry.osu.edu.

statistical analyses that assume interval, rather than ordinal, scaled data is not a valid method for analyzing subject response data.[10-11]

Several authors have advocated the use of Rasch analysis for producing interval-scaled estimates of visual ability from Likert scale data.[12-15] Rasch analysis uses subjects' responses to the items of an instrument to produce person measures meant to serve as an indicator of visual ability for each subject and item measures which are indicators of the difficulty of each item.[11] Other important indicators of instrument performance are generated including person separation reliability—higher values representing better ability to discriminate among subjects. Values of at least 0.8 are considered acceptable.[16]

The need for dedicated software, which is not always intuitive, to analyze survey data with Rasch analysis may deter some researchers. Massof recently published a scoring algorithm that allows for the approximation of the person measure produced by Rasch analysis of VFQ scores for patients with low vision without the use of specialized software.[17] This algorithm is meant to make the use of interval-scaled scoring of VFQ data easier and more accessible. The approximation should be quite useful to investigators who have small sample sizes and would not be able to obtain reliable estimates with standard Rasch analysis software, and also to clinicians and researchers who would want to obtain person measure estimates as data are being collected.

Subsets of items on the NEI VFQ-25 have been used to create shorter instruments. For example, Ryan et al proposed the use of a seven-item version of the NEI VFQ-25 for use in the evaluation of low vision rehabilitation programs.[18] They selected seven items previously demonstrated to be responsive to clinical low vision rehabilitation programs,[19] performed Rasch analysis on the responses of 490 subjects to the instrument, and concluded that the instrument was adequate for use as an outcome measure for vision rehabilitation. Likewise, the developers of the NEI VFQ-25 specified a number of subscales, each of which is a subset of items that are grouped and scored separately. These subscales are intended to serve as an indicator of more specific aspects of visual function. For instance, there is a distance vision subscale and near vision subscale. The near vision subscale includes item numbers five, six, and seven. Appendix items A3, A4, and A5 are also labeled as near vision subscale items. Each of these items was included in 51-item instrument from which the VFQ-25 was derived, as well as in the 39-item instrument.[20] There has been interest recently in evaluating whether these subscales are valid. O'Connor et al found that, of the five subscales that had enough items to perform Rasch analysis, none of the NEI VFQ-25 subscales were valid in a low vision population.[21]

The Massof algorithm was developed using data from a large group of patients with low vision but has not, to the authors' knowledge, been used to analyze NEI VFQ-25 scores in any other published studies involving other groups of patients with low vision. The primary aim of the present study was to test the Massof algorithm in an independent group of patients with low vision and compare the measures generated to those produced using dedicated Rasch analysis software, as well as to show the results of both scoring methods compared to the standard scoring method. Generally, authors present NEI VFQ-25 data using either the standard method or Rasch analysis, but not both. This is understandable, but it may leave the reader wondering where the similarities and differences in the two approaches may occur. We used the Massof algorithm, Rasch analysis software, and standard composite scoring approaches in a sample of patients with low vision and present the data in a way that allows for comparison of the methods. A secondary aim was to evaluate the Rasch person separation reliability of the seven-item version of the instrument and the near vision and distance vision subscales in this group of patients with low vision.

## Methods

### Subjects

Data from previous studies of low vision rehabilitation programs in which the NEI VFQ-25 was administered to patients with low vision were analyzed for the present study[22]. Institutional Review Board approval was obtained for the original studies and for the present study. Data from 50 patients who had previously given their informed consent were analyzed. The age of patients ranged from 60 to 89 years. Visual acuity ranged from 20/40 to 20/525.

### Analysis of 24-Item Instrument and Massof Algorithm

The set of items from the NEI VFQ-25 that was analyzed comprised 24 items that were both available from previous studies and also included in the paper describing the Massof algorithm.[17] Appendix items A4, A5, A6, and A8 were available and included in the analysis. Items that that did not deal with visual function (1, 3, 4, and 19) and items that did not have five response categories (2, A1, and A2) were excluded from the analysis. Table 1 contains a list of items included in each subset of NEI VFQ-25 items that was analyzed.

Three methods of scoring the 24 items were used. These were:

- the standard method recommended by the developers of the NEI VFQ-25
- Rasch analysis using WINSTEPS software
- the Massof algorithm to approximate Rasch analysis

The standard method for scoring the instrument requires that a score from 0 to 100 is assigned to the numerical response for each item, with 0 being the lowest possible score and 100 being the highest.[1] To generate a composite score, the scores for each individual item are averaged so that a score ranging from 0 to 100 was produced.

WINSTEPS software was used to calculate Rasch measures. Numerical responses for each item were recoded so that 0 was assigned as the lowest possible response and 4 as the highest. The ranking of response categories was reversed when necessary so that lower scores always represented lower levels of visual functioning. The software was used to calculate person measures and item measures for each subject and item, respectively.

Instructions for using the Massof algorithm and a sample Excel spreadsheet are included in the original paper by Massof.[17] Subject responses for each item are input into a spreadsheet which contains item measures that were generated with Rasch analysis by Massof using a large sample of patients with low vision. The spreadsheet also contains average functional reserve values that were generated from the same sample of patients with low vision and coefficients for an inverse hyperbolic tangent function for dealing with floor and ceiling effects.

It should be noted that care must be taken to assign the average functional reserve values in the proper order. Users must be careful that the ranking of response categories for all items is such that the most negative responses correspond to the most negative average functional reserve value in the Excel spreadsheet for the algorithm and so on for the rest of the response categories.

### Analysis of Seven-Item VFQ

The seven-item instrument described by Ryan et al contains items 5, 6, 8, 14, A3, A4, and A8 from the NEI VFQ-25. Targeting, person separation, and item performance were

evaluated for these items using Rasch analysis. As recommended by the Ryan et al [18], the categories "no difficulty at all" and "a little difficulty" were combined into a single category for the analyses.

### Analysis of Subscales

In addition, performance of the near vision and distance vision subscales was investigated using Rasch analysis. The person separation index value and person separation reliability were used to evaluate the appropriateness of use of this subscale. One cause for concern regarding the use of subsets of items from a larger instrument which contain only a few items is that the reliability of the measurements will be low.[23] Since participants in this study had responded to all three near subscale appendix items in addition to the standard 25-item instrument, we analyzed the near subscale data in two ways — as a three-item subscale composed of only the near items contained in the standard 25-item instrument (5, 6, and 7), and as a six-item subscale which also contained the three near appendix items (A3, A4, and A5). Our hypothesis was that the addition of the three appendix near vision items would increase the person separation reliability for the subscale.

### Rasch Analysis and Data Analysis

WINSTEPS version 3.69 was used to perform Rasch analysis using the Andrich rating scale model.[24] Person and item separation reliability indices and Rasch fit statistics (item infit mean squares) were used to examine measurement reliability and agreement of the observations with the expectations of the Rasch model. Principal components analysis (PCA) of the residuals was performed to further investigate dimensionality. [11, 25] Differential item functioning (DIF) was evaluated by age (older or younger than the median age) and gender with values greater than 0.5 logits considered significant.[25] Instruments and subscales were evaluated using the parameters proposed by Pesudovs et al. [16] Spearman correlation coefficients were calculated to examine the relationship between the three scoring methods of the 24-item instrument.

## Results

Table 2 contains results from Rasch analysis of instruments and subscales that were evaluated. Person measures obtained from Rasch analysis for the 24-item instrument ranged from -1.29 to +1.50 logits. Mean person measure was 0.13 logits. The mean value for item infit mean squares was close to the model expectation of one for all instruments. The Massof approximation person measures ranged from −1.20 to +1.00 logits. Mean person measure approximation was −0.11 logits. Scores obtained with the standard VFQ-25 scoring algorithm ranged from 19-85. Consistent with previous studies,[12] several individual items were found to misfit the Rasch model with item infit mean squares outside of the acceptable range. These included items involving driving (15 and 16), finding items on a shelf (7), and frustration with vision (21).

Scores for the 24-item instrument obtained using all three approaches were highly correlated. The correlation between Rasch person measure and Massof approximation was 0.996. The correlation between Rasch person measure and NEI VFQ-25 standard score was 0.997. The calculated NEI VFQ-25 composite score is plotted against Rasch person measure in Figure 1 along with the test characteristic curve produced by WINSTEPS, which shows the expectations of the Rasch model. It is evident from the figure that the standard composite score will underestimate the ability of the most visually able and overestimate the ability of the least visually able. Person measures computed with the Massof algorithm are plotted against Rasch person measures from WINSTEPS in Figure 2.

Principal components analysis of residuals indicated a lack of unidimensionality, with only 42.8% of the variance accounted for by the principal component. Typically, multidimensionality is considered likely when less than 60% of the variance is accounted for by the principal component.[25] Variance accounted for by the first contrast was 3.4 Eigenvalue units, with 2.9 units explained by the second contrast. Six items loaded positively (>0.4) onto the first contrast and were related to: dependency [stay home because of sight (20), rely too much on what others tell me (23), and need a lot of help (24)], social function [visiting with people (13)], or mental health [less control over what you do (22) and frustrated by sight (21)]. Three items loaded positively onto the second contrast and were related to near tasks: bills (A4), shaving, styling hair, makeup (A5), and reading the newspaper (5).

Two items demonstrated differential item functioning by age: seeing people reacting (11) and driving at night (16), with logit values of 0.73 and 1.51, respectively. One item displayed differential item functioning by gender: need a lot of help (24), with a logit value of 0.53. Older participants rated items 11 and 16 as relatively easier than did younger participants, while males rated item 24 as relatively easier than females.

The seven-item instrument was well-targeted, with an average person measure of −0.06. Person separation reliability was 0.80, which was less than that of the 24 item set and corresponds to the minimum acceptable value for person separation reliability. [16] Infit mean squares for all seven items were between 0.7 and 1.3, suggesting no misfit of items. Person separation reliability values for the near subscale, near subscale including appendix items, and the distance subscale were all inadequate (Table 2).

## Discussion

We have compared several approaches to scoring the NEI VFQ-25 in patients with low vision. In comparing Rasch analysis, the standard scoring method, and the Massof algorithm for approximating Rasch person measure, we found that all three methods are highly correlated. The standard scoring method, in addition to requiring assumptions about interval data that are incorrect, is subject to ceiling and floor effects. Both Rasch analysis person measure and the Massof algorithm approximation of Rasch person measure allow for theoretically sound analysis of VFQ data and are resistant to ceiling and floor effects.

Although Rasch analysis has been demonstrated in a number of published studies to be preferable to the standard scoring method for NEI VFQ-25 data[12-13, 26], the use of the standard method remains quite common in the current literature. This is potentially attributable to the fact that Rasch analysis software is required to convert the raw scores to interval-scaled data. The software, while not particularly expensive or complicated, is not widely used. The Massof algorithm, which requires only a simple spreadsheet program, should be an effective tool for making the use of interval-scaled VFQ scoring more widespread in low vision research. The extra time and effort required to convert raw scores to interval-scaled is minimal. This study provides more evidence of the algorithm's effectiveness in analyzing survey data from low vision patient populations.

We found the seven-item instrument of Ryan et al to be well-targeted to our patient group and all seven items demonstrated acceptable fit statistics. Person separation reliability, a key indicator of an instrument's ability to discriminate between persons of different abilities, was less than that of the 24-item instrument but still acceptable. These findings indicate that this short instrument may be useful in low vision research. It has the advantage of being quicker to administer. Items contained in the seven item instrument chosen in part because they were shown to be responsive to low vision rehabilitation by Stelmack et al.[19] It should be noted

that the instrument was not developed using Rasch analysis, which many would argue is a superior way to develop new instruments.

In contrast, neither the near or distance subscales of the NEI VFQ-25 were found to be valid. Inclusion of the three appendix items of the near vision subscale to form a six item subscale, however, produced values of person separation reliability that were improved from the three-item version and approached acceptable levels. Nonetheless, our findings are consistent with those of O'Connor et al and Marella et al[21, 27] and suggest that the subscales cannot be interpreted as valid measures.

Our analyses also suggest a problem with the unidimensionality of the NEI VFQ-25. Even with a number of items removed from the analysis because they are not related to visual functioning, principal components analysis of the Rasch residuals suggests that the instrument tested is multidimensional. This is problematic because the use of a composite score requires that only a single construct is being measured. The results of our principal components analysis have much in common with those recently published by Marella et al in that several of the items loading positively onto the first contrast are from the "dependency" and "mental health" content area.[27] In addition, our PCA also identified 3 near vision items loading onto a second contrast. Marella et al proposed treating the NEI VFQ-25 as an instrument with two scales, "visual functioning" and "socioemotional", to deal with the issue of multidimensionality.

A limitation of this study is the relatively small sample size, which resulted in standard errors of Rasch measurements that are larger than would be likely with a larger sample. These larger standard errors are particularly evident in the analyses of the subscales and 7-item VFQ, which contain fewer items (Table 2). The increased variability of the measurements which results from the relatively small sample size limits the generalizability of our results.

Some readers may look at the high correlation of standard scores with Raasch person measures in Figure 1 and wonder whether Rasch analysis is necessary for the analysis of VFQ scoring, but it should be noted that this is a best-case scenario in that items not pertaining directly to visual functioning were excluded from this analysis. Many of the benefits of Rasch analysis occur at the instrument development stage by enabling the researcher to eliminate poorly fitting or redundant items so that all items on the final instrument are measuring the same construct in an efficient and valid manner. Rasch analysis may also increase the sensitivity of an instrument by enabling the researcher to identify differences between populations that are not apparent using standard analysis.[13]

## Acknowledgments

## References

1. Mangione CM, Lee PP, Gutierrez PR, Spritzer K, Berry S, Hays RD. Development of the 25-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol 2001;119:1050–8. [PubMed: 11448327]

2. Miskala PH, Bass EB, Bressler NM, Childs AL, Hawkins BS, Mangione CM, Marsh MJ. Surgery for subfoveal choroidal neovascularization in age-related macular degeneration: quality-of-life findings: SST report no. 12. Ophthalmology 2004;111:1981–92. [PubMed: 15522363]

3. Miskala PH, Hawkins BS, Mangione CM, Bass EB, Bressler NM, Dong LM, Marsh MJ, McCaffrey LD. Responsiveness of the National Eye Institute Visual Function Questionnaire to changes in

visual acuity: findings in patients with subfoveal choroidal neovascularization—SST Report No. 1. Arch Ophthalmol 2003;121:531–9. [PubMed: 12695250]

4. Clemons TE, Gillies MC, Chew EY, Bird AC, Peto T, Figueroa M, Harrington MW. The National Eye Institute Visual Function Questionnaire in the Macular Telangiectasia (MacTel) Project. Invest Ophthalmol Vis Sci 2008;49:4340–6. [PubMed: 18586874]

5. Cruess A, Zlateva G, Xu X, Rochon S. Burden of illness of neovascular age-related macular degeneration in Canada. Can J Ophthalmol 2007;42:836–43. [PubMed: 18026200]

6. Ruiz-Moreno JM, Coco RM, Garcia-Arumi J, Xu X, Zlateva G. Burden of illness of bilateral neovascular age-related macular degeneration in Spain. Curr Med Res Opin 2008;24:2103–11. [PubMed: 18547463]

7. Suner IJ, Kokame GT, Yu E, Ward J, Dolan C, Bressler NM. Responsiveness of NEI VFQ-25 to changes in visual acuity in neovascular AMD: validation studies from two phase 3 clinical trials. Invest Ophthalmol Vis Sci 2009;50:3629–35. [PubMed: 19255158]

8. Chang TS, Bressler NM, Fine JT, Dolan CM, Ward J, Klesert TR. Improved vision-related function after ranibizumab treatment of neovascular age-related macular degeneration: results of a randomized clinical trial. Arch Ophthalmol 2007;125:1460–9. [PubMed: 17998507]

9. Bressler NM, Chang TS, Fine JT, Dolan CM, Ward J. Improved vision-related function after ranibizumab vs photodynamic therapy: a randomized clinical trial. Arch Ophthalmol 2009;127:13–21. [PubMed: 19139332]

10. Massof RW, Rubin GS. Visual function assessment questionnaires. Surv Ophthalmol 2001;45:531–48. [PubMed: 11425359]

11. Bond, TG.; Fox, CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2007.

12. 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. [PubMed: 11164454]

13. 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. [PubMed: 16105674]

14. 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. [PubMed: 15557851]

15. Elliott DB, Pesudovs K, Mallinson T. Vision-related quality of life. Optom Vis Sci 2007;84:656–8. [PubMed: 17700329]

16. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. Optom Vis Sci 2007;84:663–74. [PubMed: 17700331]

17. Massof RW. An interval-scaled scoring algorithm for visual function questionnaires. Optom Vis Sci 2007;84:689–704. [PubMed: 17700325]

18. Ryan B, Court H, Margrain TH. Measuring low vision service outcomes: Rasch analysis of the seven-item National Eye Institute Visual Function Questionnaire. Optom Vis Sci 2008;85:112–21. [PubMed: 18296928]

19. Stelmack JA, Stelmack TR, Massof RW. Measuring low-vision rehabilitation outcomes with the NEI VFQ-25. Invest Ophthalmol Vis Sci 2002;43:2859–68. [PubMed: 12202503]

20. Mangione CM, Berry S, Spritzer K, Janz NK, Klein R, Owsley C, Lee PP. Identifying the content area for the 51-item National Eye Institute Visual Function Questionnaire: results from focus groups with visually impaired persons. Arch Ophthalmol 1998;116:227–33. [PubMed: 9488276]

21. O'Connor, PM.; Keeffe, JE.; Pesudovs, K.; Marella, M.; Lamoureux, EL. Invest Ophthalmol Vis Sci. Vol. 50. 2009. Comparing the psychometric performance of the Impact of Vision Impairment (IVI) and the National Eye Institute Visual Functioning Questionnaire-25 (NEI VFQ-25). E-Abstract 3203

22. Dougherty BE, Martin SR, Kelly CB, Jones LA, Raasch TW, Bullimore MA. Development of a battery of functional tests for low vision. Optom Vis Sci 2009;86:955–63. [PubMed: 19593240]

23. Mallinson T, Stelmack J, Velozo C. A comparison of the separation ratio and coefficient alpha in the creation of minimum item sets. Med Care 2004;42:I17–24. [PubMed: 14707752]

24. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.

25. Linacre, J. A User's Guide to Winsteps [computer program]. Chicago: Winsteps; 2009.

26. Massof RW. Application of stochastic measurement models to visual function rating scale questionnaires. Ophthalmic Epidemiol 2005;12:103–24. [PubMed: 16019693]

27. Marella M, Pesudovs K, Keeffe J, O'Connor PM, Rees G, Lamoureux EL. The psychometric validity of the NEI VFQ-25 for use in a low vision population. Invest Ophthalmol Vis Sci 2010;51 Epub ahead of print.
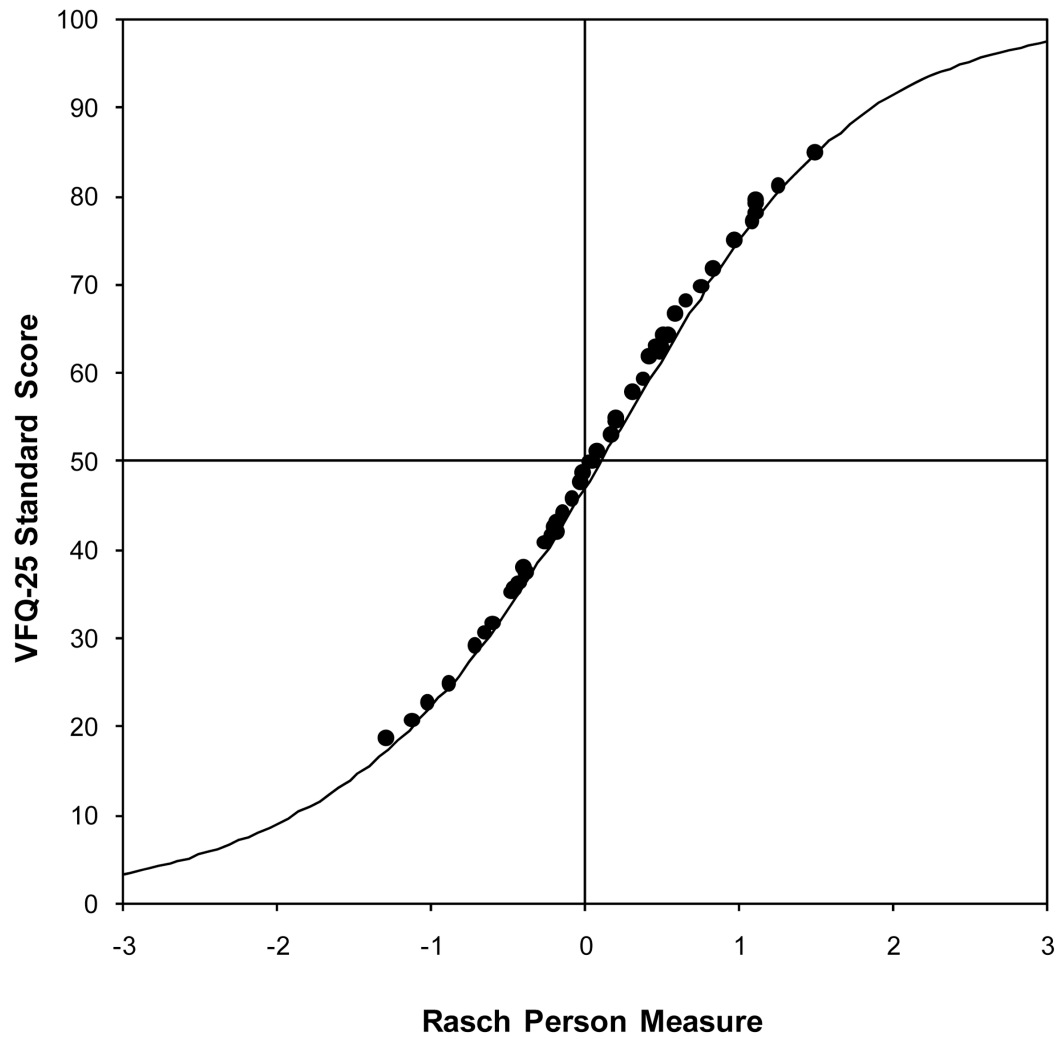
**Figure 1.**
Standard NEI VFQ-25 score is plotted against the Rasch person measure in logits generated
using WINSTEPS. The solid line is the test characteristic curve from WINSTEPS, which
represents the Rasch model prediction of standard score for given person measures.
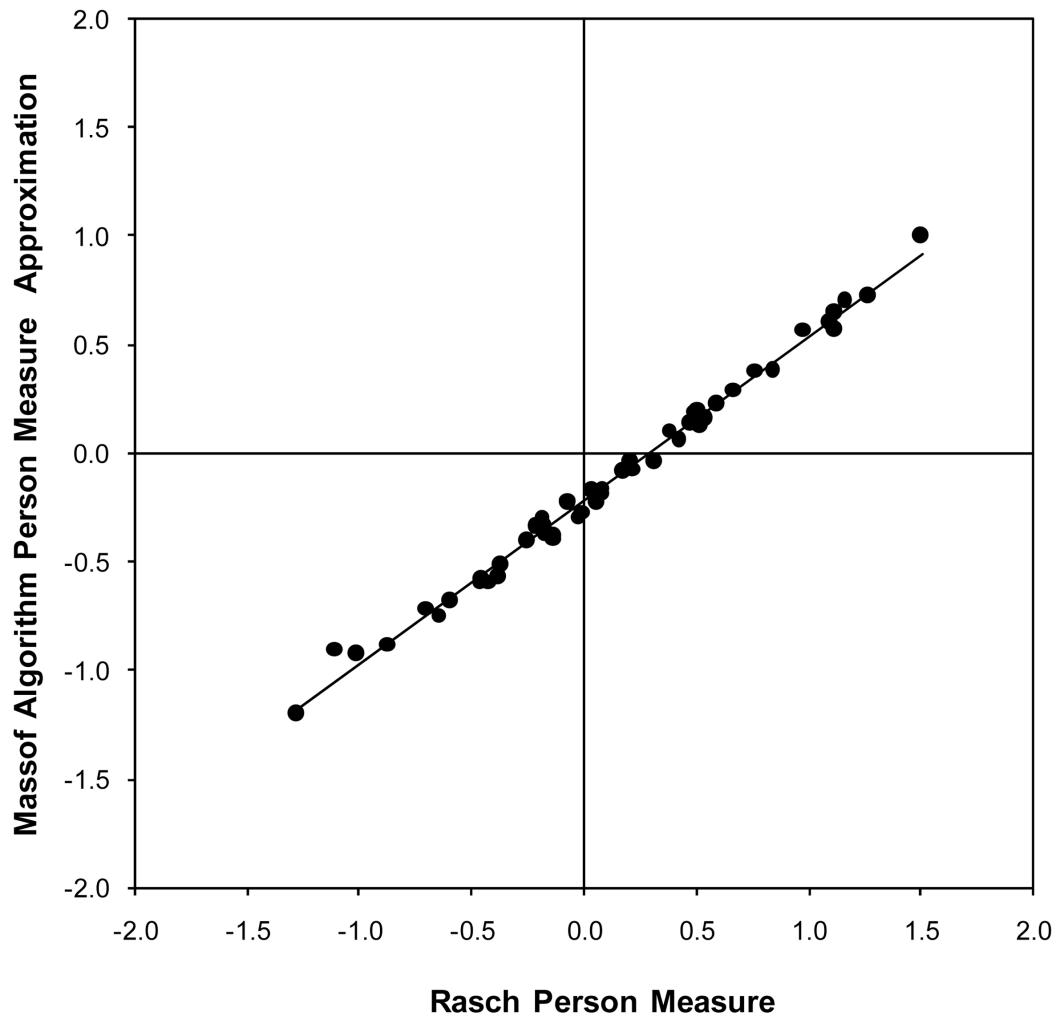
**Figure 2.**
The approximation of Rasch person measure generated using the Massof algorithm is plotted against Rasch person measure from WINSTEPS in logits.

**Table 1**

Items included in each instrument or subscale analyzed.

| Instrument (or Subscale) | NEI VFQ-25 Items Included |
|---|---|
| Massof algorithm (24 items used) | 5-18, 20-25, A4, A5, A6, A8 |
| Seven-item instrument18 | 5, 6, 8, 14, A3, A4, A8 |
| Near subscale | 5, 6, 7 |
| Near subscale with appendix items | 5, 6, 7, A3, A4, A5 |
| Distance subscale | 8, 9, 14 |

**Table 2**

Summary statistics from Rasch analysis using WINSTEPS for different instruments or subscales are shown. Mean item measure is set to zero for all instruments. Mean person measures near zero indicate good targeting of an instrument. Person and item separation reliability values greater than 0.8 are acceptable.[16] The model expectation for item infit mean square is one.

| Instrument (or subscale) | Mean Rasch Person Measure (SE) | Mean Rasch Item Measure (SE) | Person Separation Reliability | Item Separation Realiability | Item Infit Mean Squares |
|---|---|---|---|---|---|
| 24-item | +0.13 (0.09) | 0.00 (0.12) | 0.87 | 0.91 | 1.04 |
| Seven-item | −0.06 (0.23) | 0.00 (0.33) | 0.80 | 0.93 | 0.99 |
| Near Subscale | −0.54 (0.22) | 0.00 (0.59) | 0.57 | 0.93 | 0.99 |
| Near Subscale w/ appendix | −0.55 (0.17) | 0.00 (0.34) | 0.72 | 0.94 | 0.99 |
| Distance Subscale | −0.17 (0.22) | 0.00 (0.31) | 0.51 | 0.80 | 0.99 |