An Evaluation of the Reliability and Validity of the Visual Functioning Questionnaire (VF-11) Using Rasch Analysis in an Asian Population

Ecosse L. Lamoureux,1,2 Konrad Pesudovs,3 Julian Thumboo,4 Seang-Mei Saw,5,6 and Tien Y. Wong1,2,6

PURPOSE. To determine whether the VF-11 is a valid scale to measure visual functioning in an Asian population with vision impairment.

METHODS. Participants from the Singapore Malay Eye Study (SiMES) took part. Visual functioning was assessed by using the VF-11 (a modified version of VF-14 for an Asian population). Rasch analysis was performed on 618 participants with presenting visual acuity < 6/12 in the better eye.

RESULTS. Disordered thresholds were initially evident, indicating that the categories were difficult to discriminate and required category collapsing (from 5 to 4) for nine items. The removal of two misfit items related to driving resulted in a fit of the VF-9 data to the Rasch model ($\chi^2 = 50.5, df = 27, P = 0.005$). There were no more misfit items. The person separation reliability value was 0.82 which demonstrates that the VF-9 has sufficient ability to discriminate between at least two groups of participants with different levels of visual functioning.

The VF-9 significantly differentiated patients stratified by visual acuity demonstrating adequate criterion validity. All items were free of differential item functioning, and there was no evidence of multidimensionality. Targeting of person ability and item difficulty was suboptimal, although this is inevitable in a population-based survey where most people would not be disabled.

CONCLUSIONS. Although the Rasch-modified VF-9 scale achieved fit to the Rasch model, its suboptimal targeting suggests that the instrument does not have the range of items to assess the impact of vision impairment across the severity spectrum of vision loss in this population. (Invest Ophthalmol Vis Sci. 2009;50:2607–2613) DOI:10.1167/iovs.08-2359

Although objective vision impairment measures such as visual acuity are important, a comprehensive assessment of ophthalmic outcomes should also include measurements from the patient’s point of view.1 Patient-centered functioning is usually assessed with questionnaires, and these have shown that impaired vision significantly reduces activities associated with participation in society and religion, mobility, and visually intensive tasks, and increases the need for community help, family support, and nursing home placement.2–12

A major limitation common to several studies reporting on visual functioning or quality of life (QoL) has been the use of a mean or summary score. Summary scoring, termed Likert scoring, allocates an ordinal assignment of a numerical value to a participant’s response and assumes a score based on an interval scale. The validity of such summary scores has been questioned by Rasch analysis—usually referred as the Rasch measurement model.1,13–17

The model assumes that the probability that a respondent will select an item is a logistic function of the relative distance between the item location and the respondent’s location on a linear scale. In other words, the probability that a person will affirm a category within an item is a logistic function of the difference between the person’s level of, for example, visual functioning and the level of visual functioning necessary to perform this item, and only a function of that difference. Thus, Rasch analysis is taken as a criterion for the structure of the responses that should be satisfied rather than a simple statistical description of the responses. Both person ability and item difficulty must refer to one trait being measured (i.e., visual functioning) which supports the concept of unidimensionality, which is central to the Rasch model. Once the data fit the Rasch model, estimates of measures on an interval scaling are provided1,18–20 that improve the accuracy of scoring and remove measurement noise.1,18–20

The transformed overall score can then be used in analysis of variance and regression more readily than the raw total score which has floor and ceiling effects.

To date, however, few vision-specific scales have been validated by Rasch analysis, and even fewer questionnaires have been validated with Rasch analysis for use in Asia, the world’s largest continent. An estimated 22 million people are blind and 67 million with low vision currently live in Asia.21 The burden of vision loss in Asia is anticipated to increase with an aging population this century and an understating of the impact on visual impairment on daily function using a psychometrically validated scale is needed for clinical and research proposes.

The Singapore Malay Eye Study (SiMES) was undertaken to determine the prevalence and impact of visual impairment and major eye diseases in urban Asian populations. The 11-item Visual Functioning (VF-11) Questionnaire, a modified version of the 14-item Visual Functioning (VF-14) questionnaire,22 was used to determine the impact of compromised vision on visual functioning. In this article, using Rasch analysis, we determined whether the VF-11 possesses the required psychometric characteristics to measure visual functioning in an Asian pop-
interviewers were fluently bilingual. With the participant’s consent, given a choice of being interviewed in either Malay or English. All back-translated into English by two different interpreters conversant in Malay and/or English. The questionnaire was translated into Malay and difficulty; 3, a great difficulty).

Nine of the VF-11 scale items were rated on a numeric scale ranging from 0 (no difficulty) to 4 (unable to perform activity). The remaining two driving items had three response options (1, no difficulty; 2, a little difficulty; 3, a great difficulty).

Trained research assistants conducted face-to-face interviews in Malay and/or English. The questionnaire was translated into Malay and back-translated into English by two different interpreters conversant in both Malay and English. During administration, the participant was given a choice of being interviewed in either Malay or English. All interviewers were fluently bilingual. With the participant’s consent, randomly selected interviews were recorded for periodic review by the investigators for quality control purposes.

**Rasch Analysis**

Rasch analysis was undertaken to determine specific aspects of validity, reliability, and measurement of the VF-11. We used Rasch analysis software (RUMM2020, 2003; RUMM Laboratory, Perth Australia) and the Andrich rating scale to determine whether the data fit the Rasch model.27 Content validity was evaluated by using person and item fit residual statistics, where it is expected that the mean and SD values approximate 0 and 1, respectively. An overall item–trait interaction score (γ²) with a statistically nonsignificant Bonferroni-adjusted p value indicates fit to the model and that hierarchical ordering of the items (i.e., from difficult to easy) is consistent across all levels of visual functioning. An estimate of overall scale functioning is the person separation reliability (PSR) index.28 This estimate is closely linked to the targeting of the scale, as it differentiates the number of statistically distinct groups of respondents that can be identified on the trait.29 A PSR index of 0.7 determined by the RUMM software is the equivalent of a G value of 1.5, representing the ability to distinguish two distinct strata of person ability.29 A value of 0.9 is equivalent to a G value of 3, with the ability to distinguish four strata of person ability. PSR indices (similar to Cronbach’s α) of ≥0.7 are suitable for group use.31

We also assessed the disordered thresholds which occur when participants have difficulty discriminating between the response options. A disordered threshold means literally that a category expected to be “harder” than an adjacent category was actually “easier,” but often represents interchangeability of categories. Disordered thresholds indicate a lack of the invariance of items, which is the ratio of difficulties between any pair of items, remains constant across the ability levels of respondents. Category collapsing is often the solution to disordered thresholds. Differential item functioning (DIF) testing was undertaken to determine whether different groups within the sample (age, sex, types of eye disease, degree of visual impairment, and comorbidity), despite equal levels of functioning, respond differently to individual items.

Finally, the unidimensionality of the scale is assessed with a principal components analysis of the residuals. This allows for a test of the local independence of the items. This test implies that once the Rasch factor has been taken into account, there should be no further associations between the items other than random ones. The absence of any
meaningful pattern in the residuals is deemed to support the assumption of unidimensionality and the evidence of construct validity. Unidimensionality, in addition to the other adequate fit statistics of the data to the Rasch model, would support the specific validity, reliability, and measurement characteristics of the VF-11. To maximize the retention of the initial character of the VF-11 a minimalist approach was taken to item and scale changes such that only those changes essential for satisfactory scale functioning were made.

The unit of measurement in Rasch analysis is logits, which is the natural logarithm of the odds of success in choosing a response. Tasks of average difficulty are assigned 0 logits. Tasks with above-average difficulty get a negative logit score, and tasks with below-average difficulty get a positive logit score. Person ability is defined as 0 logits, tasks with below-average difficulty are assigned 0 logits. Tasks with above-average difficulty get a negative logit score. For a well-targeted scale, a sample size of 108 can give 99% confidence that the person estimate is within ±0.5 logits if the scale is not well targeted (i.e., <15% or >85% endorsement rate), as was anticipated in our population-based study with a high number of individuals with mild visual disability, then the sample size required for accurate estimation increases to 243. Consequently, the sample size of 618 participants in the present study was considered large enough to give good precision, regardless of the targeting of the sample.

**RESULTS**

**Participants’ Characteristics**

The characteristics of 618 participants with presenting visual acuity < 6/12 in the better eye are shown in Table 2. The mean ± SD age was 65.4 ± 10.5 years, and most of the participants were female (n = 384; 62.1%). More than two thirds (n = 263; 42.6%) reported at least one comorbid condition, and 67.1% (n = 427) had moderate or worse presenting visual acuity in the better eye.

Initial examination of the person fit statistics showed that half of them (n = 311, 50.3%) were considered to have a pattern of extreme responses. They either reported that they had no difficulty with the nine items or the items were not applicable. Further analysis found that participants with extreme responses had significantly better vision than did those with a nonextreme pattern of responses (logMAR 0.51 vs. 0.64, P < 0.001). They also had a significantly greater proportion of participants with undercorrected refractive error (55% vs. 45%, P < 0.05) as the cause of decreased visual acuity. Removal of these participants, however, did not improve the overall item–trait interaction statistics. Nonetheless, considering that those with extremes scores did not contribute to the estimates of the item locations, the validation of the VF-11 was essentially undertaken by using data from the remaining 312 participants.

**Fit of the VF-11 Data to the Rasch Model**

For ease of interpretation of scores of the VF-11 rating scale, the scoring was reversed for the Rasch analysis giving higher scores to participants with high visual ability. The Andrich rating scale was used to determine whether the VF-11 data fitted the Rasch model. One rating scale was used for the nine items with a five-response scale and another one with the two driving items with three response options. Rasch analysis of the VF-11 data initially showed a lack of fit to the Rasch model with a significant item–trait interaction probability (total $\chi^2 = 114.4, df = 44; P = 0.000000$). Examination of the pattern of item thresholds revealed disordered thresholds for the nine items with the five response options but not for the two driving items. As shown in Figure 1, the response category 3 was not consistently chosen by the participants (i.e., that the difficulty of a higher threshold was lower than that of its adjacent lower threshold), indicating that response

![Figure 1](https://iovs.arvojournals.org/pdfaccess.ashx?url=/data/journals/iovs/932960/ on 11/25/2018)
category 3, a little difficulty, was never the most likely response of any participant and was therefore considered a superfluous response option. This result necessitated that categories 2, moderate difficulty, and 3, a little difficulty, be collapsed (Fig. 1, right), which resulted in ordered thresholds for all items. Consequently, scores for the nine items were recorded from 4, 3, 2, 1, and 0 into 3, 2, 2, 1, and 0.

Rescoring improved the overall item-trait interaction probability (total $\chi^2 = 92.4$, $df = 44$; $P = 0.000027$), although it still remained statistically significant, indicating that the VF-11 did not function within the Rasch model. Examination of the item fit statistics showed that item 11, difficulty in driving at night time because of vision, and item 10, difficulty in driving during the day because of vision, showed misfit to the model expectation with probabilities of 0.00011 and 0.00026, respectively, which are less than the Bonferroni-adjusted $\alpha$ value of 0.004. These two driving items also recorded the worst response rates (15.5% for both). Misfit items are often indicative that these items do not measure the same construct as other items of the scale. In this particular case, there may be reasons that older patients have difficulty with driving that are not related to their vision (i.e., fear of driving). Item 11 was removed first, followed by item 10.

After item deletion, the VF-9 (since the instrument now has 9 items, not 11, we will hereafter refer to it as the VF-9) showed fit to the Rasch model with a Bonferroni-adjusted nonsignificant item-trait interaction total $\chi^2$ probability ($\chi^2 = 50.5$, $df = 27$, $P = 0.005$). The mean (SD) items and persons fit residuals were $-0.57$ (0.82) and $-0.35$ (0.79), respectively, where optimal fit of data to the model would have a mean of 0 and an SD of 1. All items showed fit residuals $< 2.5$ with a probability $> 0.0045$, indicating no significant deviation from the model (Table 3). The PSR was 0.82, which demonstrates that the VF-9 was able to discriminate between at least two groups of participants with different levels of visual functioning.

### Table 3. Fit Indices of the VF-9 Scale

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>Standard Error</th>
<th>Item Fit Residuals</th>
<th>Degrees of Freedom</th>
<th>$\chi^2$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Difficulty seeing stairs</td>
<td>$-0.657$</td>
<td>0.124</td>
<td>$-1.279$</td>
<td>260.050</td>
<td>7.749</td>
<td>0.051</td>
</tr>
<tr>
<td>2. Difficulty reading street signs</td>
<td>0.535</td>
<td>0.092</td>
<td>0.174</td>
<td>257.490</td>
<td>4.296</td>
<td>0.231</td>
</tr>
<tr>
<td>3. Difficulty recognizing friends</td>
<td>0.669</td>
<td>0.089</td>
<td>$-0.052$</td>
<td>258.340</td>
<td>2.047</td>
<td>0.563</td>
</tr>
<tr>
<td>4. Difficulty watching television</td>
<td>$-0.295$</td>
<td>0.113</td>
<td>$-0.162$</td>
<td>259.190</td>
<td>4.032</td>
<td>0.258</td>
</tr>
<tr>
<td>5. Difficulty cooking</td>
<td>$-1.947$</td>
<td>0.221</td>
<td>0.015</td>
<td>213.150</td>
<td>1.227</td>
<td>0.746</td>
</tr>
<tr>
<td>6. Difficulty playing games</td>
<td>$-1.459$</td>
<td>0.467</td>
<td>$-1.145$</td>
<td>34.960</td>
<td>1.356</td>
<td>0.716</td>
</tr>
<tr>
<td>7. Difficulty reading newspaper</td>
<td>0.782</td>
<td>0.107</td>
<td>$-1.632$</td>
<td>186.720</td>
<td>10.590</td>
<td>0.014</td>
</tr>
<tr>
<td>8. Difficulty filling out lottery forms</td>
<td>0.889</td>
<td>0.104</td>
<td>$-1.127$</td>
<td>185.020</td>
<td>10.795</td>
<td>0.009</td>
</tr>
<tr>
<td>9. Difficulty reading small print</td>
<td>1.483</td>
<td>0.096</td>
<td>0.305</td>
<td>173.080</td>
<td>3.441</td>
<td>0.328</td>
</tr>
</tbody>
</table>

### Targeting

The person-item threshold map (Fig. 2) shows the persons and items on the same calibrated logit scale for only 307 participants with nonextreme scores. The upper part of the graph is the distribution of persons, and the lower half is the distribution of the thresholds from the items. The map shows a skewed spread of items across the range of respondents’ scores suggesting a suboptimal targeting of the patients to the VF-9 item thresholds. Overall, a large number of participants had no difficulty undertaking the hardest items. The mean $\pm$ SD person location logit value ($2.07 \pm 0.99$ logits) substantiates that overall the questionnaire was not optimally targeted and that the participants had a higher level of visual functioning than the average of the scale items (which would be 0 logit).

Overall, the three most difficult items in the VF-9 were difficulty reading small print (1.48 logits), difficulty in filling out lottery forms (0.89 logits), and difficulty reading newspaper (0.78 logits; Table 3). Conversely, the three least difficult items were difficulty cooking, difficulty playing games, and difficulty seeing stairs, with logit scores of $-1.95$, $-1.46$, and $-0.66$, respectively. A similar hierarchal item order has been recorded in prior studies.35,36 The response rates for these nine items ranged from 98.5% (watching TV) to 64.4% (playing games).

![FIGURE 2.](https://iovs.arvojournals.org/pdfaccess.ashx?url=/data/journals/iovs/932960/)
Differential Item Functioning

DIF was tested for age (40–49, 50–59, 60–69, and 70–80 years); sex (male, female); types of eye disease (cataract, age-related maculopathy, diabetic retinopathy, glaucoma, undercorrected refractive error, and other), degree of visual impairment (mild, moderate, or severe), and comorbidity (yes and no). We found no evidence of DIF on any of these factors for all items of the VF-9. For example, for the item difficulty seeing stairs, there was no clear systematic difference in the pattern of responses between those with mild, moderate, or severe vision impairment ($P = 0.82$; Fig. 3, top) or between those with or without other comorbidities ($P = 0.64$; Fig. 3, bottom).

Cultural DIF

To test for culture-specific DIF between Singapore Malays and Western subjects, we compared our VF-9 item measures to previously published VF-14 item measures from patients with low vision in the United States. As shown in Figure 4, there was substantial agreement between the two sets of item measures, except for recognizing friends ($r = 0.50$ with that item included and $r = 0.79$ with it excluded). The line through the data is the principal component (bivariate regression with orthogonal fit; slope $= 1.63$ and intercept $= 0.21$) when recognizing friends was excluded. This result suggests strong agreement between U.S. and Singapore Malay patients with low vision in item measure calibration. Population-specific DIF may be evident for the item recognizing friends.

Unidimensionality

PCA of the residuals identified two subsets of items consisting of the highest positive (item $8 = 0.75$, item $9 = 0.70$, and item $7 = 0.59$) and negative loading items (item $6 = -0.33$, item $3 = -0.57$, and item $2 = -0.66$). Only $3.36\%$ of estimates were found to be significantly different for these participants. These values are less than the recommended cutoff of $5\%$, and therefore no evidence of multidimensionality was detected. This finding confirms the internal construct validity of the VF-9 and
that the instrument measuring the underlying trait (vision function) that it purports to measure. These results collectively show that the VF-9 is a unidimensional scale and the overall score conforms to an interval scaling.

**Criterion Validity**

The criterion validity of the Rasch-scaled VF-9 was tested by assessing its ability to discriminate between participants with mild (VA < 6/12–6/18), moderate (<6/18–6/60), or severe (<6/60) vision loss. There was a significant difference between the three groups (ANOVA; $F_{2, 300} = 5.73; P = 0.003$) with poorer visual acuity being associated with worse visual functioning (mean of 2.32, 2.00, and 1.55 logits for mild, moderate, and severe visual impairment, respectively).

**Scoring of the VF-9 Questionnaire**

Other investigators wishing to use the VF-9 can use the validation data detailed in this article to convert raw scores into Rasch person measures without having to perform Rasch analysis. We have generated a Rasch scoring key in a spreadsheet (Excel; Microsoft, Redmond, WA; available on request from ecosse@unimelb.edu.au) that converts category responses into 36 item–category calibrations, to create Rasch measurement estimates.

**DISCUSSION**

In this article, using Rasch analysis, we investigated whether the original VF-11 possesses the required psychometric characteristics to measure visual functioning in an Asian population of Singaporean Malays with low vision. There was evidence of disordered thresholds initially, which necessitated that the response options be collapsed from five to four categories for 9 of the 11 items. In addition, the two “driving” questions were removed from the scale, as they showed evidence of misfit. Subsequently, the remaining nine items were found to fit the Rasch model which indicates that the VF-9 is an appropriate scale to measure visual functioning in this population. We demonstrated that the VF-9 is a unidimensional scale and is free of DIF for several important disease related factors, including culture. The scale possesses interval scaling and good person separation reliability. There was evidence, however, of suboptimal targeting of person ability to item difficulty which suggests that further work is needed to optimize the performance of the instrument.

There have been previous attempts to reduce the original VF-14 questionnaire to fewer items. Uusitalo et al.37 created a shortened instrument, the VF-7, by analyzing the results of the VF-14 administered to patients with cataract and then selecting seven items that correlated best with patient satisfaction. Similarly, to improve efficiency, Mallinson et al.35 demonstrated the use of Rasch measurement in six subsets of seven items of the original VF-14. Although our initial choice of 11 items was motivated by items that were considered relevant to the Singaporean culture, examination of our PSR index suggests that the VF-9 did not lose measurement precision in the shortened version. Our PSR index was 0.82, which indicates that the scale can distinguish between several distinct strata of person ability and has good reliability. In addition, this value compares favorably with that of the VF-14 (0.84) and indicates that the two scales are comparable and the shortening of the original VF-14 did not affect the measurement precision of the VF-11.35,36 Similarly, we found that the items difficulty reading small print and difficulty cooking were considered to be the most and least difficult items, respectively, on the VF-9. Similar findings were noted after the VF-14 underwent Rasch analysis, which indicates the comparability of the item measures of the two scales.35,36

Although it is difficult to compare vision-specific areas of activity limitation between Singaporean Malays and other populations because of the use of different scales and psychometric evaluation techniques, it can be hypothesized that activities associated with near and distance reading are consistently rated as the most difficult to undertake. For example, in Australians with low vision, it was also found that after Rasch analysis of the Impact of Vision Impairment (IVI) Questionnaire, the two most difficult items were reading ordinary size print and reading labels or instructions on medicines.13 Similarly, after Rasch analysis of the Activities of Daily Vision Scale (ADVS) in British patients awaiting cataract surgery reading signs was found to be the most challenging activity.34 A parallel result was found in people with visual impairment living in East Timor who rated the most difficult item of the Rasch-analyzed Timor-Leste vision-specific QoL scale as being the reading things such as a newspaper, bible, or book.39 These findings suggest that difficulty associated with reading activities is consistent across people with low vision, irrespective of their cultural differences. Our findings also highlight the need for the development and validation of an item bank that would enhance our ability to compare visual functioning and quality of life across cultures and significantly open up the field of low vision to international comparison studies.

The application of Rasch analysis to the VF-9 has allowed greater scrutiny of the performance of the response scale that would not have been possible with the traditional approach to test development (i.e., classic test theory). Singaporean Malays with low vision consistently did not use the response option a little difficulty on any of the nine items of the VF-9. The nonuse of this option necessitated the collapsing of the response categories with the adjacent one (i.e., moderate difficulty), which subsequently produced a consistent endorsement of all response options across items. The nonutilization of categories is not unusual in scales with many response options, or when the labeling of options is too similar to each other, which can be confusing or open to misinterpretation. Our finding is consistent with other vision-specific scales that have benefited from a shortening of their response scales after Rasch analysis.14,38,40–42

The person–item threshold map showed that numerous participants recorded a positive person logit score, which suggests that the person’s level of functioning is higher than the mean required level of functioning for the items. Considering that only a small percentage of participants (4.2%) were considered to have severely impaired vision, it is possible that our low-vision sample was not impaired enough to experience substantial visual functioning difficulties. On the other hand, it could be argued that since almost two thirds (64.9%) of our sample were considered to have moderate vision impairment (categorized as presenting visual acuity ranging from <6/18 to 6/60), a better targeting was anticipated. Perhaps the reason for this result is the inclusion of refractive error as the dominant cause of poor visual acuity, because low myopia may cause a decrease in VA, but enable good reading vision, and thus may not cause as much visual disability as cataract or macular degeneration. Regardless, the poor targeting of VF-9 may indicate the instrument does not include the content required to fully assess the impact of vision impairment across the severity spectrum of vision loss in this population. Future investigations are therefore needed to determine whether the inclusion of items of greater difficulty improves the targeting of patients with low vision in this population.

In conclusion, there are very little published data on vision-specific QoL or vision functioning scales in Asia that have been validated with Rasch analysis, which is increasingly considered
the gold standard of modern psychometric scale validation. The Rasch-scaled VF-9 in this Singaporean Malay population demonstrated internal consistency and criterion validity, with no evidence of DIF and multidimensionality. However, targetting was suboptimal, suggesting that the VF-9 is not suitable for population-based studies and that additional items are needed to optimize the instrument’s performance.

References