This is a repository copy of *Refinement of the Child Amblyopia Treatment Questionnaire (CAT-QoL) using Rasch analysis*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/146357/

Version: Accepted Version

Article:
Carlton, J. orcid.org/0000-0002-9373-7663 (2019) Refinement of the Child Amblyopia Treatment Questionnaire (CAT-QoL) using Rasch analysis. Strabismus, 27 (2). pp. 66-77. ISSN 0927-3972

https://doi.org/10.1080/09273972.2019.1601743

This is an Accepted Manuscript of an article published by Taylor & Francis in Strabismus on 23/04/2019, available online:
http://www.tandfonline.com/10.1080/09273972.2019.1601743

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Refinement of the Child Amblyopia Treatment Questionnaire (CAT-QoL) using Rasch analysis

Corresponding Author

Jill Carlton

Health Economics and Decision Science (HEDS)

School of Health and Related Research (ScHARR)

University of Sheffield

Regent Court

30 Regent Street

Sheffield

S1 4DA

United Kingdom

j.carlton@sheffield.ac.uk

Tel: 0114 2220799

Fax: 0114 2724095

Keywords: Child Amblyopia Quality of Life Rasch analysis

ABSTRACT

Aims or Purpose

The Child Amblyopia Treatment Questionnaire (CAT-QoL) was developed using a "bottom-up" methodological approach. Interviews with children with amblyopia identified items (questions) and response levels to be tested in a draft questionnaire consisting of 11 items (sad, feeling on face, hurt, doing schoolwork, cross, how other children treat you, doing things, worried, upset with family, playing with friends, happy). This study describes the refinement of the descriptive system for the CAT-QoL instrument using the application of Rasch analysis.

Methods

A multi-centre pilot study was conducted, and data collected from 342 participants. Participants were asked to self-complete the appropriate treatment version of the CAT-QoL questionnaire Socio-demographic and clinical data was collected by the clinician using a standardised proforma. A "measure" of child's health was obtained from the parent by asking how they would rate their child's health over the previous week. Rasch analysis techniques were applied to refine the questionnaire. Rasch was used to examine response categories and collapse item response levels, identify poorly performing items, and explore local dependency of items.

Results

A total of 331 subjects were included in the study sample, however only 315 were accepted into the RUMM program as a number of subjects had missing questions responses on the CAT-QoL. RUMM also excluded a further 41 subjects as these demonstrated extreme responses. Disordered response categories were found for each item, requiring adjacent response levels to be combined. This was applied to all items, and the model fit was re-examined. Two items were found to have poor fit (cross and happy) and were removed from the measure and the model fit was re-examined. No statistically significant differential item functioning (DIF) was found for any item, using person factors of age, sex, or general health. Two items showed some dependency (worried and upset with family), and the poorer fitting item was subsequently

removed (upset with family). This resulted in a refined CAT-QoL instrument that consists of 8-items, each with 3-level response scales.

Conclusion

The refined CAT-QoL instrument includes the following items: sad, feeling on face, hurt, doing work at school, how other children treat you, doing things, worried, and playing with friends). The CAT-QoL can be Rasch scored, with a range of 0-16 where a greater value indicates a worse quality of life (or greater impact of treatment on the individual). The CAT-QoL may be useful in determining how amblyopia treatment affects children, and offers an alternative to generic patient reported outcome measures.

INTRODUCTION

The Child Amblyopia Treatment Questionnaire (CAT-QoL) is a paediatric disease-specific patient-reported outcome measure (PROM) for amblyopia. Designed for children aged 4-7 years, it was created through a number of methodological stages.[1-4] A "bottom-up" development approach was adopted, and children's responses were used to determine the content of the draft descriptive system. Interview data directly informed the items (questions) of the instrument; the response levels of the instrument; and the wording and layout of the draft instrument. This approach ensures good content and face validity. Seven treatment-specific versions of the draft questionnaire were created (patch; drops; glasses; patch and drops; patch and glasses; glasses and drops; glasses, patch and drops), with each version worded slightly differently to reflect the type of treatment the child is undertaking. The draft questionnaire contains eleven items that are marked on a 5- or 6-part response scale.[4] An example is shown in Figure 1. The items include; sad, feeling on face, hurt, doing schoolwork, cross, how other children treat you, doing things, worried, upset with family, playing with friends, and happy.

PROMs provide a mechanism of measuring health or health-related quality of life (HRQoL), and use rating scales to assess the variable. How the rating scales and items are defined and selected may differ depending upon the theoretical approach applied during their development. The classic approach (e.g. classical test theory) assumes that a total test score is made up of multiple items, and that this score is made up of both a "true component" and a "random error component".[5] The response levels of the items will have an assigned value, and can be thought of as categorical responses. Two main assumptions are made with this approach: that equal intervals exist between each response category; and that each item of the instrument has the same difficulty. The appropriateness of these assumptions have been questioned, and techniques developed to address such issues. Rasch analysis is a mathematical technique that converts categorical responses to a continuous latent scale using a logit model.[6] Rasch analysis converts item responses into a continuous latent scale covering the full severity range, and positions individual responses on the scale. Item responses are assumed to be a function of the location of both the person and the item on the logit scale.[7] The fundamental principle of the Rasch model is that "the outcome of an encounter between a person and an item is governed by the product of the ability of the person and the easiness of the item".[8] The easier an item is, the more likely it will be passed; and the more able the person, the more likely they will pass an

item compared to a less able person.[9] Rasch analysis can be used in PROM development to address the appropriate number of response levels for items, and to identify poorly performing items. The application of Rasch analysis in the development and refinement of ophthalmology questionnaires is becoming increasingly common, with application to the Adult Strabismus-20 (AS-20)[10], Ocular Comfort Index (OCI)[11], Ocular Surface Disease Index (OSDI)[11], Brief Impact of Vision Impairment (IVI) questionnaire[12], and the Keratoconus Outcomes Research Questionnaire (KORQ)[13], to name but a few. This study describes the application of Rasch analysis to refine the descriptive system for the CAT-QoL instrument.

When evaluating an instrument using Rasch analysis the following are explored: Overall model fit; Individual person fit and item fit with the Rasch model; Thresholds; Differential item functioning; and Local independence. The overall fit of the model for the scale is given by a Chi-Square Item-Trait Interaction statistic. This is calculated by adding the chi-square values for the individual scale items. Statistical significance is determined using the associated summated degrees of freedom. A non-significant value indicates that there is no substantial deviation from the model; and that the hierarchical ordering of the items is consistent across all levels of the underlying trait. When looking at the overall model fit, a non-significant probability is desirable. This means that our observed scores (i.e. participant responses) are not different from the model (what we expect). If misfit is found in the model, (that is to say that the observed and expected scores differ) then this should be investigated. The misfit may be the result of misfitting respondents or misfitting items. If the items and persons fit the model, we would expect a mean of 0, and a standard deviation (SD) of 1. If the observed values differ from these then the individual person fit and individual item fit should be explored.

The individual person fit is explored by examining the Person Separation Index (PSI). The PSI figure provides an indication of the power of the instrument to be able to discriminate amongst respondents with different levels of the trait being measured. So in this study, the PSI is an indication of how the CAT-QoL is able to discriminate between respondents who have different severity levels of amblyopia. A value of 0.7 is the minimum accepted level of PSI. This value indicates that the measure is able to statistically differentiate between 2 groups of patients.[14] A value of 0.8 represents the ability of the measure to statistically differentiate at least 3 ability

groups. A value of 0.9 would indicate the ability of the measure to discriminate between 4 or more groups.[15] If items are misfitting then this is demonstrated by two statistics: a Fit Residual value of 2.5 or more and a significant Chi-square probability value.[16] Misfitting items may be due to: inconsistent use of the response options (disordered thresholds); or item bias across groups of respondents (differential item functioning); or multidimensionality (local independence). Cronbach's Alpha may also be used to assess the reliability of the measure. This ranges from 0 to 1 with 0.70 being the lowest level of acceptability.[14]

Threshold refers to the point between two response categories where either response is equally probable (e.g. the point where the probability of scoring a 0 or a 1 is 50/50). However, Rasch analysis may reveal there to be disordered thresholds. That is, there is inconsistent use of the response thresholds. Simply put, respondents are not answering the items in a way that was expected. It occurs when respondents have difficulty in discriminating between the response options. This may be because there are too many options to choose from, or that the labelling of the response options is confusing. To investigate disordered thresholds, responses to an item are inspected on a category probability curve. If disordered thresholds are found, this can be addressed by collapsing adjacent category response levels for that item. After doing this, the model fit needs to be re-evaluated again.

Differential item functioning (DIF) is a form of item bias across groups of respondents. It occurs when different groups within the same sample, despite equal levels of the underlying characteristic, respond in a different manner to an individual item. There are two different types of DIF; uniform DIF (where one group shows a consistent systematic difference in their responses to an item, across the whole range of the attribute being measured) or non-uniform DIF (which occurs when the differences between groups varies across the level of the attribute). There are different methodological approaches that can be taken if DIF is found. In the case of instrument development, the presence of DIF may be influence the removing of that item from the instrument. DIF can be detected both statistically and graphically. An ANOVA is performed for each of the items, comparing the scores across each level of the "person factor" and across different levels of the trait (class intervals). Uniform DIF is indicated by a significant main effect

for the person factor (for example, gender). Non-uniform DIF is indicated by a significant interaction effect (person factor X interval).

Local dependency is another potential source of misfit within a scale. This is where a person's response to one item in the scale will have a bearing upon their response to another, different item within the same scale. Local dependency is assessed by looking at how the residual correlations for each item correlate with the residuals of every other item. There is no current consensus as to the values that indicate local dependency among items. However, a residual correlation between 0.2 and 0.3 above the average of all item residual correlations is thought to be problematic.[17]

The application of Rasch analysis in instrument development is not uncommon. However, there is no widely accepted approach as to what element to consider first in the inclusion/exclusion of items.

MATERIALS AND METHODS

Patient Cohort

Data used in this study was collected from nine sites across England, United Kingdom (UK). Inclusion criteria was that used during the development of the draft descriptive system.[3;4] The study was approved by the National Health Service Research Ethics Committee for Airedale, United Kingdom (UK), (REC Ref: 07/Q1201/5), and followed the tenets of the Declaration of Helsinki. Written parent/guardian consent was obtained prior to data collection. Each participant was asked to self-complete the appropriate treatment version of the CAT-QoL questionnaire. Item responses were scored from 0-4 (or 0-5 where appropriate) as indicated on Figure 1. Socio-demographic and clinical data was collected by the clinician using a standardised proforma (see Supplementary Material). A "measure" of child's health was obtained from the parent by asking how they would rate their child's health over the previous week. Response options included excellent, very good, good, fair and poor.

Rasch Analysis

The following steps were undertaken using an iterative approach using Rasch Unidimensional Measurement Models (RUMM2020).[18] An iterative approach was undertaken in the analysis. Figure 2 shows the methodological stages of the Rasch analysis. The acceptability criteria recommended by RUMM2020 are as follows.

Items were then assessed to investigate the goodness of fit to the Rasch model. This was done by assessing fit residuals and item-trait interactions. Fit residuals estimate the amount of divergence between the expected and observed responses and are investigated for both respondents and items. Divergence residuals > 2.5 are considered high, and so respondents outside of these levels are removed from analysis. When all the respondents fit the model, items are checked using the same criteria. Items with residuals > 2.5 are excluded. Once all items and persons fit the model, we would expect a mean of 0, and a standard deviation (SD) of 1.[7] The overall fit of the model for the scale is given by a Chi-Square Item-Trait Interaction statistic (X^2). This is calculated by adding the chi-square values for the individual scale items. Statistical significance is determined using the associated summated degrees of freedom. A non-significant value (> 0.01) indicates that there is no substantial deviation from the model; and that the hierarchical ordering of the items is consistent across all levels of the underlying trait. When looking at the overall model fit, a non-significant probability is desirable. This means that our observed scores (i.e. participant responses) are not different from the model (what we expect).

The item-fit, person-fit residuals, and item trait interactions are assessed for the overall model. The unidimensionality of the dimension is investigated by calculating independent t-tests comparisons of person estimates generated by different subsets of valid items. If a scale is unidimensional, then at least 95% of the t-tests will be non-significant.[7] Local dependency is also assessed by looking at how the residual correlations for each item correlate with the residuals of every other item. There is no current consensus as to the values that indicate local dependency among items. However, a residual correlation > 0.3 above the average of all item residual correlations is thought to be problematic.[17] Item range is examined, considering the range on the logit scale, and the spread at logit 0. A large range indicates that an item covers a

fuller range of the severity of the underlying construct being measured.[7] It is desirable for the range to include values above and below 0, as this means that the item covers both more severe and less severe cases, respectively. Spread at logit 0 relates to the spread of response at the average item severity, and again a higher spread indicates a better item coverage.

RESULTS

Study Sample

The socio-demographic details of the study sample are shown in Table 1. In total, 342 subjects were recruited into the study. There was missing clinical data for some participants. These were excluded from the sample (n=11), leaving 331 participants in any subsequent analysis. One hundred and eighty nine (57%) were male, and 142 (43%) were female. The age range of the study sample (4-8 years) is reflective of that seen clinically, with the majority of children on amblyopia treatment aged between 5 and 7 years. The range of interocular difference in VA (logMAR) at the time of the study was 0 – 1.75, with a mean of 0.20 and median of 0.175. Participants were rated in terms of their amblyopia severity at the time of the validation study. The definitions chosen were informed by previous studies by the PEDIG group, whereby mild was categorised as 0≥0.30 logMAR, moderate 0.31≥0.60 and severe >0.61 logMAR.[19-21] Table 2 shows the clinical demographics of the study sample, in terms of type of amblyopia, strabismus, and refractive error present. The majority of participants were in excellent or very good general health (as reported by parents). The majority of participants received the Glasses only version of the CAT-QoL or the Patch and Glasses version (n=145 and n=173 respectively).

Rasch analysis

A total of 331 subjects were included in the study sample, however only 315 were accepted into the RUMM program as a number of subjects had missing questions responses on the CAT-QoL. RUMM also excluded a further 41 subjects as these demonstrated extreme responses. These people are removed from the analysis for the purpose of calculating the item (and person) parameters, for they do not provide any useful information as they sit at either the floor or the ceiling of the scale. The extremes are removed only for the parameter estimation procedures. The “extreme respondents” are given a location on the scale (however, this is a “guess” as the

scale does not have the measurement points to be any more precise). This procedure is run automatically within the RUMM program.

Table 3 shows the summary fit statistics after the Rasch model had been applied. Initial analysis (initial) showed that the Person-Fit Residual is acceptable. There is a low mean value, and the SD is close to 1. The Item fits Residual shows a high mean, and a high SD. This suggested misfitting items. The Chi-Square Interaction Probability Statistic was at an acceptable level (<0.05). The test for unidimensionality (t-test score percentage) was below the accepted criteria (<5%).

Disordered response categories were found for each of the CAT-QoL items, and there were different levels of disordered response categories for each item. The category probability curves for the 11-item CAT-QoL instrument are shown in Supplementary Material B. The number of response levels was reduced for each item, with an attempt to ensure that the maximum number of response levels remained. The number of response levels for each item after the collapsing of adjacent categories is shown in Table 4. The majority of items (n=7) allowed 3-response level thresholds. Two items had a 4-response level threshold, and two items had only 2-response level thresholds. The aim was to apply the same number of response level thresholds for each item (to aid consistency throughout the overall instrument to reduce the complexity of the task for the respondent). To create a measure with a 2-response level threshold would result in an instrument with a "yes/no" type response option. This is not desirable as the final instrument would how low levels of sensitivity to detect changes in QoL. It was therefore decided to collapse the item response levels to 3-level responses for each of the 11 items. Introducing a common 3-response level improved the fit of the model (Table 3 – stage 1). The Chi-Square Interaction Probability Statistic remained acceptable (<0.05). The test for unidimensionality t-test score increased to 3.28% (although this remained within acceptable levels). The Person-Fit Residual was acceptable, and the SD was virtually at 1. The Item Fit Residual was acceptable, however the SD was high. This suggested that there were still some items that were misfitting.

Two items were shown to have poor fit. Item 5 (cross) showed a high negative Fit Residual outside of the accepted criteria. Item 11 (happy) showed a high positive Fit Residual and high Chi-Square probability which were both outside of accepted criteria. A decision was made to remove these items from the instrument. Removal of the items improved the fit of the model (Table 3 – stage 2). The Person-Fit Residual is still acceptable with a low mean value, and the SD is virtually at 1. The Item Fit Residual shows an acceptable mean, the SD is now acceptable (virtually at 1). This suggests that the items are fitting the Rasch model. The test for unidimensionality t-test score decreased to 1.54% (a value of < 5% is suggestive that the instrument is unidimensional[17]). Individual Item Fit was further explored. None of the remaining items showed any significant Fit Residuals (see Table 5).

The remaining items were explored to establish whether there was any item bias across groups of respondents by assessing the presence of DIF. No statistically significant DIF was present for the person factors of age, sex, the presence of any other health condition, CAT-QoL version, or child's health as reported by the parent. Local dependency of the items was explored by observing the correlation scores between the items (Table 6). Two items showed some dependency (worried and upset with family). A decision was made to omit one of these items. Upset with family was chosen as this had a higher fit residual value.

Removing Item 9 (upset with family) altered the fit of the model (as shown in Table 4 – stage 3). The Person-Fit Residual was acceptable with a low mean value and the SD is close to 1. The Item Fit Residual showed an acceptable mean, and the SD is now closer to 1 (than Stage 2). This suggests that overall the items are working. The test for unidimensionality t-test score has decreased, to 1.53% and so demonstrated "more" unidimensionality. Individual Item Fit was further explored. None of the items show any significant Fit Residuals. The items were then explored to establish whether there was any item bias across groups of respondents by assessing the presence of Differential Item Functioning (DIF). No statistically significant DIF was present for the person factors of age, sex, the presence of any other health condition, or CAT-QoL version of child's health as reported by the parent for Item 9 (upset). Local dependency of the items was explored by observing the correlation scores between the items. No local dependency was found for any of the remaining items.

After the removal of the three items (cross, happy and upset with family), the goodness of fit the Rasch model was re-evaluated (X^2=44.47; p-value=0.07; Item Fit (SD)=-0.200(0.825); Person Fit (SD)=-0.233(0.881); Person Separation Index=0.74). Table 7 shows the Rasch analysis summary for the individual items in the final CAT-QoL instrument. It can be seen that the items included in the final CAT-QoL instrument do demonstrate good coverage.

The final 8-item CAT-QoL instrument

The refined CAT-QoL instrument consists of 8-items, each with 3-level response scales. It includes the following items: sad, feeling on face, hurt, doing work at school, how other children treat you, doing things, worried, and playing with friends. An example of the final questionnaire is shown in Figure 3.

Scoring of the final 8-item CAT-QoL instrument

The results of the Rasch analysis were used to re-score the final CAT-QoL items to that the score they provide is an ordinal scale. Rescoring of the CAT-QoL instrument was achieved using the formula: y=m + (s * Logit score). The "m" and "s" values can then be used to transform the logit score into the desired 0 to 16 interval scale score, using the formula y= m + (s * Logit score). In the case of the CAT-QoL, the original scale was scored 0-16, therefore the "wanted" range of person scores = 0 to 16. The "current" range of persons scores observed in the study was -3.60 to 3.48 (given in logits). The results are shown in the conversion table, Table 8. The CAT-QoL instrument is scored summatively. Individual item responses are scored from 0 to 2 (least to worst) meaning the instrument has a range of 0-16. The summative score is then converted into a Rasch score (as shown in Table 8). It should be noted that this conversion chart can only be used when there is no missing data from an individual. It can only be used when complete data is present. For example, if an individual scored 14 (raw data score) this would be the equivalent of 12.6 on the re-scored measure. The final CAT-QoL scores range from 0-16, where a greater score indicates a worse quality of life (or greater impact of treatment on the individual).

DISCUSSION

This study describes the application of Rasch analysis with the primary aim of refining the CAT-QoL instrument. The results demonstrated that the refined CAT-QoL instrument offers good range, and coverage. The process of item selection for PROMs can be subjective. Some are driven by theory, and utilise factor analysis to pre-defined domains (informed by clinicians or literature). The results of factor analysis can be used as a basis to accept/reject items.[22] Other studies have used Rasch analysis, in conjunction with clinical input and psychometric assessment during the item selection process.[7;23;24] Both of these techniques adopt a "top-down" approach, where clinical opinion is imposed upon the content of the instrument. The application of Rasch techniques in this study has continued the ethos of the overall aim of this study, of developing an instrument for children, by children. Refinement of the descriptive system was informed by their responses, with the results of the analysis directing which items and response levels to keep in the instrument. As discussed, Rasch analysis transforms categorical responses to a linear scale.

Previous studies have shown that young children are able to reliably report upon their own health.[25-27] However, considerations must be made when designing PROM instruments for the paediatric population. The number of items included in an instrument contributes to the response burden of the task. The refined CAT-QoL instrument consists of eight items, which is smaller than other self-report PROM instruments, such as the PedsQL.[28;29] Results of the Rasch analysis revealed disordered response categories for each CAT-QoL item. This could suggest that children aged 4-7 years are not able to make the distinction between 5- or 6-level responses, and may interpret some of the response levels to mean the same conceptually. Although children aged 8 years have been shown to accurately use a 5-part or 7-part response scale, the target population for the CAT-QoL is younger than this (children aged 4-7 years).[30] The number of response level options in the refined CAT-QoL instrument is appropriate for the target population, with each item consisting of three response level options. It is anticipated that this lower task complexity will contribute to good acceptability and completion rates. The relatively poor Person Separation Index (Table 3) could be described as being low (however, still above acceptable level of >0.70). This may be another indication of the difficulty of children responding to questionnaires such as the CAT-QoL. It was not possible to determine how well participants

were able to complete the CAT-QoL questionnaire without help from others. It can only be assumed that the responses were self-report (rather than proxy-reported).

The study is not without limitations: the main being the size of the study sample. The optimum number for Rasch analysis is 500.[2931] However, a number of 300 is considered sufficient, and previous studies have also been limited to this number.[10] It may also be argued that the use of Rasch to determine selection of items is not appropriate. Mulhern et al postulated that this approach selects items with the best statistics, and that these may not best reflect the HRQoL of the patient.[7] However, as the development of items for the CAT-QoL was driven largely by the interviews with children; the face validity of the instrument is already high. Rasch analysis resulted in the removal of only three items suggested following analysis of the qualitative data. The remaining items in the CAT-QoL instrument cover a wide variety of aspects of HRQoL.

Furthermore, there are only small numbers of respondents in the "severe" category group. A number of reasons may account for this. The first is that of categorization: subjects were categorized into severity groups as used by the PEDIG studies.[19-21] However, this categorization is arbitrary, and is not universally accepted. Secondly, respondents in poorer health may not have participated in the study. It was outside of the scope of the study to document reasons for children not agreeing to participate. Another reason for the small number of respondents in the "severe" category is that amblyopia may have been detected (and therefore treatment initiated), at an earlier age, thereby reducing the potential of "severe" amblyopia in the available study population.

In conclusion, the methods applied have further refined a paediatric disease-specific PRO for amblyopia. Previous stages of development have ensured good content and face validity of the instrument.[3;4] Here, quantitative techniques were applied to select items and response levels for the final descriptive system. The refined CAT-QoL instrument (see Supplementary Material) offers an alternative to generic measures to measure the HRQoL implications of amblyopia treatment from a child's perspective. Further research is required to examine the psychometric properties of the instrument, examining both reliability and validity.

ACKNOWLEDGEMENTS

Data for this study was collected at the following sites: Bradford Teaching Hospitals NHS Foundation Trust; Doncaster and Bassetlaw NHS Foundation Trust; Harrogate and District NHS Foundation Trust; Maidstone and Tunbridge Wells NHS Trust; Medway NHS Foundation Trust; Sandwell and West Birmingham Hospitals NHS Trust; Sheffield Children's NHS Foundation Trust; Sheffield Teaching Hospitals NHS Foundation Trust; The Mid Yorkshire Hospitals NHS Trust. The author gratefully acknowledges the support from orthoptists and clinicians at each of the collaborating sites.

This work is produced by the author under the terms of Personal Development Award research training fellowship issued by the NIHR. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, The National Institute for Health Research or the Department of Health.

Disclosure of Interest

The authors report no conflict of interest.

Table 1 Study sample socio-demographics

	Number of subjects (%)
Age	
4 years	5 (1.5)
5 years	145 (43.8)
6 years	123 (37.2)
7 years	56 (16.9)
8 years	2 (0.6)
Ethnicity	
White	238 (71.9)
Mixed	9 (2.7)
Asian	46 (13.9)
Black	7 (2.1)
Other ethnic group	4 (1.2)
Not stated	27 (8.2)
Presence of any Co-morbidities	
Yes	30 (9.1)
No	301 (90.9)
Interocular difference at time of validation study	306 (92.4)
Mild (0 ≥ 0.3)	226 (73.9)
Moderate (0.31 ≥ 0.60)	70 (22.9)
Severe (> 0.61)	10 (3.3)

Amblyopia treatment history	
Glasses now*	324 (97.9)
Glasses previously*	0
Patching now*	170 (51.4)
Patching previously*	70 (21.1)
Atropine now*	6 (1.8)
Atropine previously*	11 (3.3)
Health State (parental report)	
Excellent	210 (63.4)
Very good	73 (22.1)
Good	23 (6.9)
Fair	4 (1.2)
Poor	3 (0.9)
Not answered	18 (5.4)
CAT-QoL Version issued	
Patch	9
Drops	0
Glasses	145
Patch and Drops	0
Patch and Glasses	173
Glasses and Drops	0
Glasses, Patch and Drops	0

TOTAL	**331**

Mild amblyopia 0 ≥ 0.3 logMAR

Moderate amblyopia 0.31 ≥ 0.60 logMAR

Severe amblyopia > 0.61 logMAR

* not mutually exclusive

Table 2 Type of amblyopia, strabismus and refractive error present in validation study (n=331)

Condition	N (%)
Type of amblyopia	
Strabismic	105 (31.7)
Anisometropic	86 (26.0)
Mixed	69 (20.8)
Microtropic amblyopia	35 (10.9)
Other	36 (10.6)
Type of Strabismus	
No strabismus present	124 (37.5)
Esotropia	103 (31.1)
Esotropia with microtropia	17 (5.1)
Exotropia	6 (1.8)
Microtropia	63 (19.0)
Intermittent	18 (5.4)
Type of Refractive Error	
No refractive error	7 (2.1)
Hypermetropia	100 (30.2)
Myopia	6 (1.8)
Astigmatism	4 (1.2)
Anisometropia	64 (19.3)
Mixed refractive error	150 (45.3)

	Item Location		Person Location		Item Fit Residual		Person Fit Residual		Chi Square Interaction			Person Separation Index	Unidimensionality T-tests		
Stage	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**With extremes α**	**N° of significant tests**	**Out of**	**Percentage at < 5%**
Initial	0	0.250	-0.986	0.826	-0.834	1.605	-0.314	0.850	99.12	44	0.00000	0.827	3	274	1.09
1	0	0.526	-1.682	1.292	-0.263	1.998	-0.216	0.921	99.55	44	0.00001	0.7998	9	274	3.28
2	0	0.492	-1.781	1.317	-0.295	1.044	-0.250	0.927	55.69	36	0.02368	0.7712	4	260	1.54
Final	0	0.507	-1.715	1.285	-0.200	0.825	-0.233	0.881	44.47	32	0.07028	0.7424	4	261	1.53

Table 3 Summary statistics of validation study data: log of Rasch approach

Criteria

- Chi-square Item Trait Interaction Probability Statistic: should be above 0.05
- Mean Person-Fit Residual value and standard deviation (SD): mean should be close to 0; SD close to 1
- Mean Item-Fit Residual value and SD: mean should be close to 0; SD close to 1
- Person Separation Index (PSI): should be > 0.7
- Unidimensionality (Percentage at < 5%): value should be less than 5%

Table 4 Maximum number of possible response levels after reducing response levels for each item

Item	**Number of Level Thresholds**
1 (sad)	3
2 (feeling on face)	4
3 (hurt)	4
4 (doing schoolwork)	3
5 (cross)	2
6 (children treating you)	3
7 (doing things)	3
8 (worried)	3
9 (upset with family)	3
10 (playing with friends)	3
11 (happy)	2

Table 5 Individual Item Fit scores after removing Item 5 (cross) and Item 11 (happy)

Item	Location	Standard Error	Fit Residual	Degrees of Freedom	Chi-Square Value	Degrees of Freedom	Probability
1 (sad)	-0.468	0.105	-1.636	229.61	11.208	4	0.024328
2 (feeling on face)	-0.845	0.114	0.338	226.97	3.271	4	0.513464
3 (hurt)	0.144	0.12	-0.084	229.61	4.743	4	0.314667
4 (doing schoolwork)	0.029	0.117	0.552	222.57	4.888	4	0.298973
6 (children treating you)	0.159	0.119	0.808	228.73	4.657	4	0.3243
7 (doing things)	-0.371	0.106	-1.155	224.33	3.479	4	0.481055
8 (worried)	0.26	0.122	-0.61	227.85	7.263	4	0.122648
9 (upset with family)	0.268	0.123	-1.787	226.09	12.662	4	0.013051
10 (playing with friends)	0.823	0.141	0.92	225.21	2.523	4	0.640563

Table 6 Person Item Residual Correlation Matrix after removal of Item 5 (cross) and Item 11 (happy) (Stage 2)

Item	Item 1	Item 2	Item 3	Item 4	Item 6	Item 7	Item 8	Item 9
Item 1 - sad								
Item 2 - feeling on face	0.007							
Item 3 - hurt	-0.162	-0.091						
Item 4 - doing schoolwork	-0.116	-0.106	-0.156					
Item 6 - children treating you	-0.174	-0.148	-0.05	-0.197				
Item 7 - doing things	-0.193	-0.188	-0.092	-0.08	-0.171			
Item 8 - worried	-0.056	-0.208	-0.201	-0.104	-0.196	-0.14		
Item 9 - upset with family	-0.066	-0.205	-0.171	-0.174	-0.127	-0.028	**0.146†**	
Item 10 - playing with friends	-0.179	-0.247	-0.105	-0.131	0.014	-0.092	-0.046	-0.114

† outside of accepted criteria

(average of residual correlations = -0.128

Table 7 Rasch analysis summary for final 8-item CAT-QoL instrument

Item	**Item Range (logit values)**	**Fit Residual**	**X^2 p-value**	**Spread at logit**
Sad	0.223 to -1.078	1.771	0.041	0.44 to 0.75
Feeling on face	-2.29 to 0.699	0.111	0.129	0.33 to 0.91
Hurt	-1.281 to 1.643	-0.334	0.213	0.16 to 0.78
Doing schoolwork	-0.699 to 0.824	0.173	0.622	0.30 to 0.67
Children treating you	-0.447 to 0.829	0.571	0.377	0.30 to 0.61
Doing things	-0.529 to -0.136	-1.011	0.448	0.53 to 0.63
Worried	-0.387 to 0.927	-0.007	0.249	0.28 to 0.60
Playing with friends	0.018 to 1.684	0.666	0.232	0.16 to 0.50

Table 8 Rescoring of CAT-QoL instrument

CAT-QoL Raw Score	Person Scores	Interval Level Equivalences	Rounded Interval Level Equivalent Score
0	-3.60	-0.00000020	0.0
1	-2.65	2.14689245	2.1
2	-1.98	3.66101674	3.7
3	-1.51	4.72316363	4.7
4	-1.13	5.58192069	5.6
5	-0.81	6.30508453	6.3
6	-0.52	6.96045176	7.0
7	-0.25	7.57062125	7.6
8	0.01	8.15819187	8.2
9	0.27	8.74576249	8.7
10	0.55	9.37853085	9.4
11	0.84	10.03389808	10.0
12	1.16	10.75706192	10.8
13	1.53	11.59322011	11.6
14	1.98	12.61016926	12.6
15	2.61	14.03389807	14.0
16	3.48	15.99999976	16.0

References

1. Carlton J, Kaltenthaler E. Amblyopia and quality of life: a systematic review. Eye. 2011; 25(4):403-13.
2. Carlton J. Clinicians' perspectives of health-related quality of life (HRQoL) implications of amblyopia: a qualitative study. Br Ir Orthopt J. 2011;8:18-23.
3. Carlton J. Identifying potential themes for the Child Amblyopia Treatment Questionnaire. Optom Vis Sci. 2013;90:867-73.
4. Carlton J. Developing the draft descriptive system for the Child Amblyopia Treatment Questionnaire (CAT-QoL): a mixed methods study. Health Qual Life Outcomes 2013;11:174.
5. Kline TJB. Classical test theory: Assumptions, equations, limitations, and item analyses. In: Psychological Testing: A practical guide to design and evaluation. Sage Publications; 2005: 91-106
6. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med. 2003; 35:105-15.
7. Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, et al. Improving the measurement of QALYs in dementia: Developing patient- and carer-reported health state classification systems using Rasch analysis. Value Health. 2012;15(2):323-33.
8. Wright BD, Panchapakesan N. A procedure for sample-free analysis. Educational and Psychological Measurement. 1969;29(1):33-48.
9. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health. 2004;Supplement 1:S22-6.
10. Leske DA, Hatt SR, Liebermann L, Holmes JM. Evaluation of the Adult Strabismus-20 (AS-20) questionnaire using Rasch analysis. Invest Ophthalmol Vis Sci. 2012 May 4;53(6):2630-9.
11. McAlinden C, Gao R, Wang Q, Zhu S, Yang J, Yu A, Bron AJ, Huang J. Rasch analysis of three dry eye questionnaires and correlates with objective clinical tests. Ocul Surf. 2017 Apr;15(2):202-210.
12. Fenwick EK, Man RE, Rees G, Keeffe J, Wong TY, Lamoureux EL. Reducing respondent burden: validation of the Brief Impact of Vision Impairment questionnaire. Qual Life Res. 2017 Feb;26(2):479-488.
13. Khadka J, Schoneveld PG, Pesudovs K. Development of a Keratoconus-Specific Questionnaire Using Rasch Analysis. Optom Vis Sci. 2017;94(3):395-403.
14. Fisher W. Reliability statistics, separation, strata statistics. Rasch Measurement Transactions 1992;6(3):238.

15. Bland JM, Altman DG. Cronbach's alpha. Br Med J 1997;314(7080):572.
16. Andrich D. Rasch Models for Measurement. London: Sage Publications; 1988.
17. Tennant A. An introduction to Rasch analysis using RUMM 2030. 2011.
18. Rasch Unidimensional Measurement Models (RUMM) 2020©. 1997.
19. Pediatric Eye Disease Investigator Group. The clinical profile of moderate amblyopia in children younger than 7 years. Arch Ophthalmol. 2002;120(3):281-7.
20. Repka MX, Beck RW, Holmes JM, Birch EE, Chandler DL, et al. A randomized trial of patching regimens for treatment of moderate amblyopia in children. Arch Ophthalmol. 2003;121(5):603-11.
21. Pediatric Eye Disease Investigator Writing Committee, Rutstein RP, Quinn GE, Lazar EL, Beck RW, Bonsall DJ, et al. A randomized trial comparing Bangerter filters and patching for the treatment of moderate amblyopia in children. Ophthalmol. 2010;117(5):998-1004.
22. Wolffsohn JS, Cochrane AL. Design of the Low Vision Quality-of-Life Questionnaire (LVQOL) and measuring the outcome of low-vision rehabilitation. Am J Ophthalmol. 2000;130(6):793-802.
23. Mulhern B, Rowen D, Jaboby A, Marson T, Snape D, Hughes D, et al. The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy Behav. 2012;24(1):36-43.
24. Rowen D, Brazier JE, Young T, Gaugris S, Craig BM, King MT, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-30. Value Health. 2011;14(5):721-31.
25. Juniper EF. Health-related quality of life in asthma. Curr Opin Pulm Med. 1999; 5(2):105-10.
26. Annett RD. Assessment of health status and quality of life outcomes for children with asthma. J Allergy Clin Immunol. 2001; 107(5):S473-81.
27. Connolly MA, Johnson JA. Measuring quality of life in paediatric patients. Pharmacoeconomics. 1999; 16(6):605-25.
28. Varni JW, Seid M, Rode CA. The PedsQL: Measurement Model for the Pediatric Quality of Life Inventory. Med Care. 1999; 37(2):126-139.
29. Varni JW, Seid M, Kurtin PS. The PedsQL 4.0: Reliability and validity of the Pediatric Quality of Life Inventory Version 4.0 Generic Core Scales in healthy and patient populations. Med Care. 2001; 39(8): 800-812.
30. Riley AW, Forrest CB, Rebok GW, Starfield B, Green BF et al. The Child Report Form of the CHIP-Child Edition: reliability and validity. Med Care. 2004; 42(3):221-31.
31. Linacre JM. Investigating rating scale category utility. J Outcome Meas. 1999;3:103-22.

Figure 1 Example of 11-item CAT-QoL (patch version)

Item		Item Score
Sad	My patch has not made me feel sad	0
	My patch has made me feel a little bit sad	1
	My patch has made me feel a bit sad	2
	My patch has made me feel quite sad	3
	My patch has made me feel very sad	4
Feeling of your patch on your face (like sticky, or itchy)	The feel of my patch has not bothered me	0
	The feel of my patch has bothered me a little bit	1
	The feel of my patch has bothered me a bit	2
	The feel of my patch has bothered me a lot	3
	The feel of my patch has really bothered me	4
Hurt	My patch did not hurt me	0
	My patch hurt me a little bit	1
	My patch hurt me a bit	2
	My patch hurt me quite a bit	3
	My patch hurt me a lot	4
	My patch really hurt me	5
Doing work at school (like reading and writing)	My patch has not made it hard to do my work	0
	My patch made it a little bit hard to do my work	1
	My patch made it a bit hard to do my work	2
	My patch made it quite hard to do my work	3
	My patch made it very hard to do my work	4
Cross	My patch did not make me feel cross	0
	My patch made me feel a little bit cross	1
	My patch made me feel a bit cross	2
	My patch made me feel quite cross	3
	My patch made me feel very cross	4
How other children have treated you (like laughing at you, or calling you	Children have not laughed at me or called me names	0
	Children have laughed at me or called me names a little bit	1
	Children have laughed at me or called me names a bit	2

names) because of your patch	Children have laughed at me or called me names quite a bit	3
	Children have laughed at me or called me names a lot	4
	Children have really laughed at me or called me names	5
Doing things (like playing on the computer, colouring, playing games, watching TV)	My patch has not made it hard to do things	0
	My patch has made it a little bit hard to do things	1
	My patch has made it a bit hard to do things	2
	My patch has made it quite hard to do things	3
	My patch has made it very hard to do things	4
Worried	My patch has not made me feel worried	0
	My patch has made me feel a little bit worried	1
	My patch has made me feel a bit worried	2
	My patch has made me feel quite worried	3
	My patch has made me feel very worried	4
Upset	My patch has not made me feel upset	0
	My patch has made me feel a little bit upset	1
	My patch has made me feel a bit upset	2
	My patch has made me feel quite upset	3
	My patch has made me feel very upset	4
Playing with my friends	My patch has not stopped me playing with my friends	0
	My patch has stopped me playing with my friends a little bit	1
	My patch has stopped me playing with my friends a bit	2
	My patch has stopped me playing with my friends quite a bit	3
	My patch has stopped me playing with my friends a lot	4
	My patch has really stopped me playing with my friends	5
Happy	My patch has not made me feel happy	0
	My patch has made me feel a little bit happy	1
	My patch has made me feel a bit happy	2

	My patch has made feel quite happy	3
	My patch has made me feel very happy	4

Figure 2 Methodological stages of the Rasch analysis

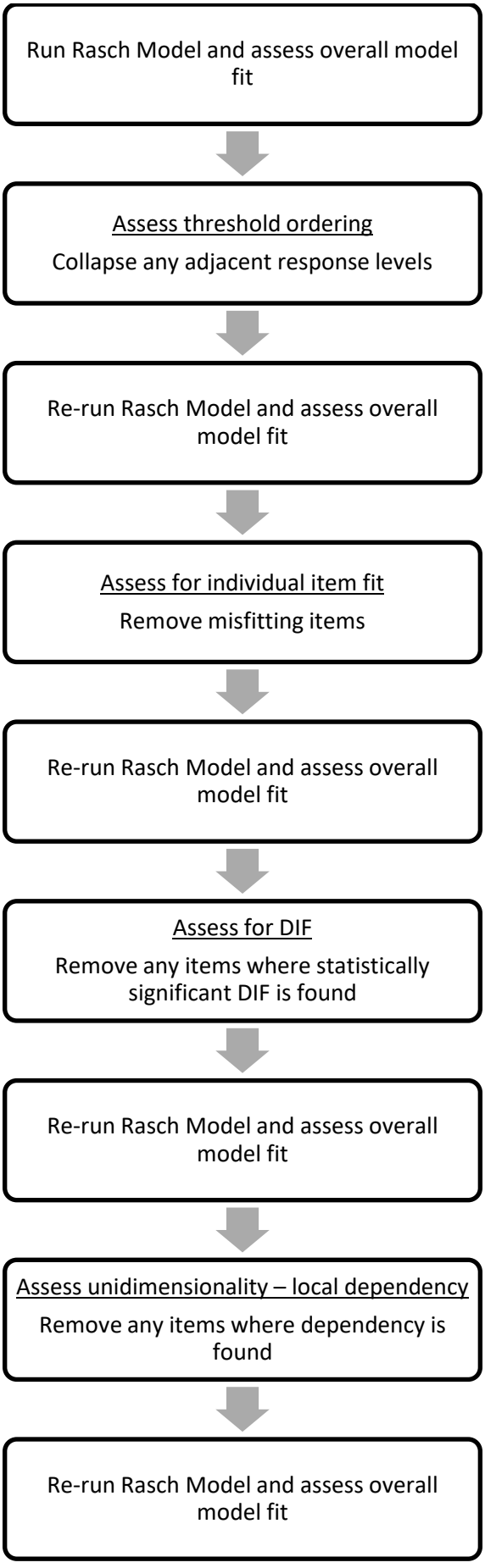

Run Rasch Model and assess overall model fit
Assess threshold ordering
Collapse any adjacent response levels
Re-run Rasch Model and assess overall model fit
Assess for individual item fit
Remove misfitting items
Re-run Rasch Model and assess overall model fit
Assess for DIF
Remove any items where statistically significant DIF is found
Re-run Rasch Model and assess overall model fit
Assess unidimensionality – local dependency
Remove any items where dependency is found
Re-run Rasch Model and assess overall model fit

Figure 3 Final 8-item CAT-QoL (patch version)

Item		Item Score
Sad	My patch has not made me feel sad	0
	My patch has made me feel a little bit sad	1
	My patch has made me feel very sad	2
Feeling of your patch on your face (like sticky, or itchy)	The feel of my patch has not bothered me	0
	The feel of my patch has bothered me a bit	1
	The feel of my patch has bothered me a lot	2
Hurt	My patch did not hurt me	0
	My patch hurt me a bit	1
	My patch hurt me a lot	2
Doing work at school (like reading and writing)	My patch has not made it hard to do my work	0
	My patch made it a bit hard to do my work	1
	My patch made it very hard to do my work	2
How other children have treated you (like laughing at you, or calling you names) because of your patch	Children have not laughed at me or called me names	0
	Children have laughed at me or called me names a bit	1
	Children have laughed at me or called me names a lot	2
Doing things (like playing on the computer, colouring, playing games, watching TV)	My patch has not made it hard to do things	0
	My patch has made it a bit hard to do things	1
	My patch has made it very hard to do things	2
Worried	My patch has not made me feel worried	0
	My patch has made me feel a bit worried	1
	My patch has made me feel very worried	2
Playing with my friends	My patch has not stopped me playing with my friends	0

	My patch has stopped me playing with my friends <u>a bit</u>	1
	My patch has stopped me playing with my friends <u>a lot</u>	2

Supplementary Material A

Trust Study ID Number:

Data Collection Form

Patient Initials **Date of Questionnaire**

Date of Birth **Age (yrs)**

Patient's postcode

Ethnicity **Sex** Male/Female

Ophthalmic Diagnosis/Diagnoses

Other Diagnosis/Diagnoses

Details of amblyopia treatment (insert dates)

Glasses _____ Started _____

Ongoing? (please tick)_____ Previously? (please tick)_____

Patching _____ Started _____

Ongoing? (please tick)_____ Previously? (please tick)_____

Atropine _____ Started _____

Ongoing? (please tick)_____ Previously? (please tick)_____

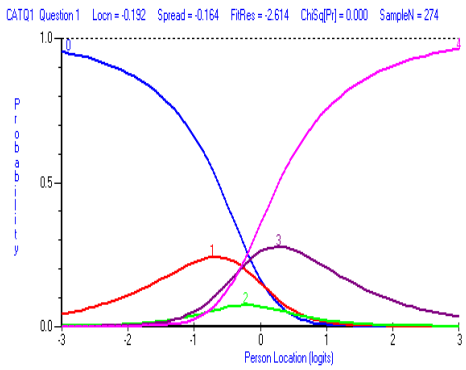
Visual Acuity at time of questionnaire

With/without glasses RE LE Test

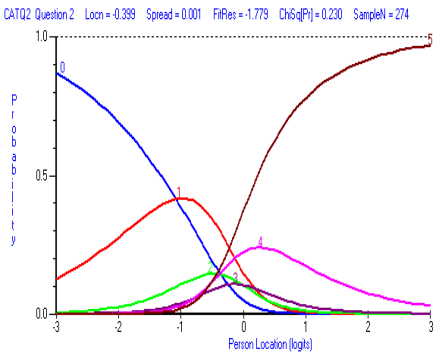
Visual Acuity when treatment first initiated

With/without glasses RE LE Test

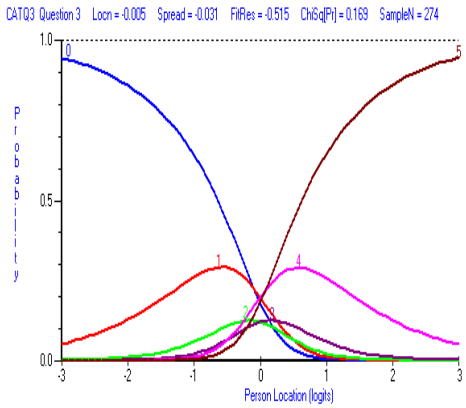
Supplementary Material B: Category probability curves for the 11-item CAT-QoL instrument



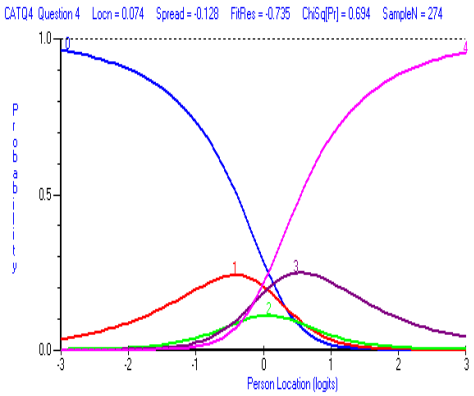
Item 1 (sad)



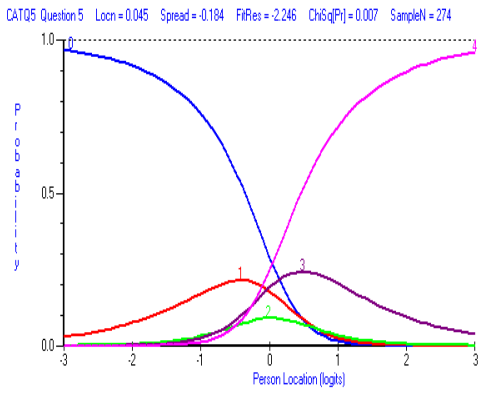
Item 2 (feeling on face)



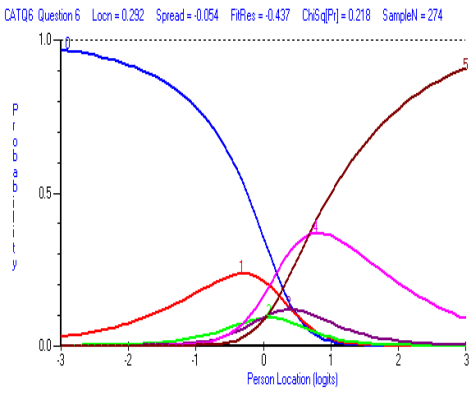
Item 3 (hurt)



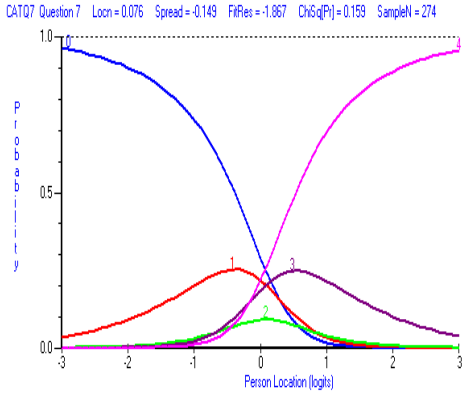
Item 4 (doing schoolwork)



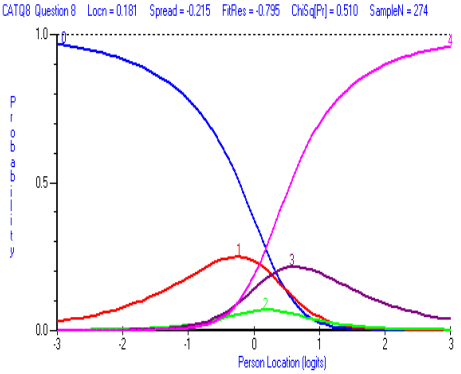
Item 5 (cross)



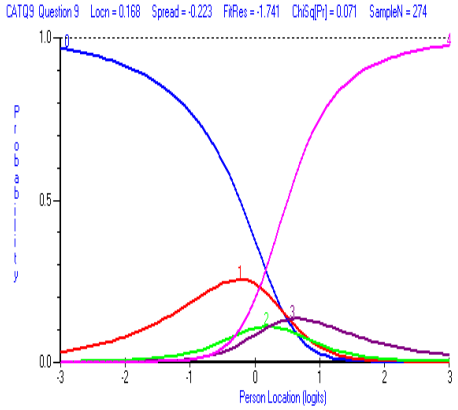
Item 6 (children treating you)



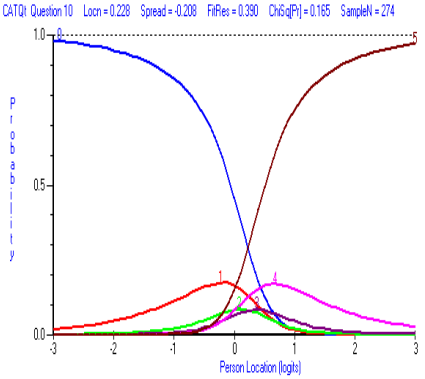
Item 7 (doing things)



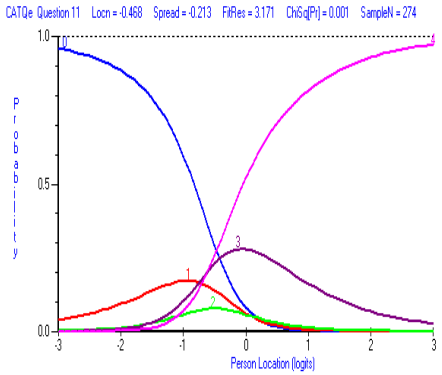
Item 8 (worried)



Item 9 (upset with family)



Item 10 (playing with friends)



Item 11 (happy)