Universiteit Leiden

Psychologie
Faculteit der Sociale Wetenschappen

Influence of imputation methods on the psychometric properties of the Visual Assessment Scale for non-structural and structural missing values.

M.A.J. Luijten

Leiden University

On behalf of:

Visio

Master's Thesis Methodology and Statistics Master

Methodology and Statistics unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University.

Name: Michiel Adrianus Jozef Luijten
Date: June 18th 2018
Student number: S1742493
Supervisors internal: E. Dusseldorp & J. van Ginkel
Supervisors external: M. Wallroth & M. Steendam

Abstract

When assessing the psychometric qualities of questionnaires, performance tests or observational instruments, missing values are a common problem. In the presence of structural and non-structural missing data, the problem becomes more complex and several methods of handling the missing data can be applied. In this thesis we considered the following five methods within a Rasch framework: a) treating missing values as fails (MAF); b) treating non-structural missing values by full information maximum likelihood (FIML) and structural missing values as fails (FIML-MAF); c) treating all missing values by FIML (FIML); d) treating non-structural missing values by plausible value multiple imputation (PVMI) and structural missing values as fails (PVMI-MAF), and e) treating missing values by PVMI (PVMI). To get insight into the impact of these methods on the assessment of psychometric properties of an instrument, we applied them to binary (pass/fail) data gathered in children with cerebral visual impairment (CVI) with the Visual Assessment Scale (VAS). Children with CVI often are often unable to follow instructions, resulting in a large amount of non-structural missing values for the VAS The VAS items are divided across six levels of visual ability and items in a higher level of visual ability are assumed to have a higher difficulty (based on theoretical background). The structural missing values of the VAS are a result of raters no longer rating items once a patient does not pass the majority of items in a certain level of visual functioning. Patients who cannot pass items in certain level of visual functioning, should not be able to pass items in a higher level of visual functioning. The difficulty parameters, item, person and model fit and internal consistency were compared to assess the psychometric properties of the VAS under the five different methods for handling missing data. The theoretical framework of the VAS was used to compare item difficulty misfit.

The results indicate that treating non-structural missing values as fails leads to worse item, person and model fit than treating these missing values with a model-based imputation method such as PVMI or FIML. For structural missing values on items that were completed by only few patients, the item difficulties were substantially lower when applying a model-based imputation method (FIML/PVMI) than when replacing these missing values with fails (MAF/FIML-MAF/PVMI-MAF). This resulted in item difficulties that differed from the theoretically assigned difficulty (i.e. they required less visual ability than assumed), when applying PVMI and FIML methods for handling missing data. However, we do not know what the true difficulty parameters are. This means that we cannot say that replacing structural missing values with fails improves the difficulty parameter estimation, unless the a priori assumptions we make about the increasing item difficulty holds. If this assumption does hold (i.e. the true difficulty parameters are known and increase in difficulty), then treating structural missing values as fail will be a solution for treating missing data. The choice of which method should be used thus depends largely on the assumptions that are made about the questionnaire/instrument prior to assessing the psychometric qualities of the instrument.

Table of Contents

1. Introduction

Missing data are a common problem in analyzing data of performance tests, questionnaires, or observational instruments. A missing value occurs when a question (or item) is not filled in by a participant or not scored by the observer. Reasons behind missing values in the data obtained by psychometric instruments can vary. In this study we are interested in two types of missing data (Adèr, Mellenbergh & Hand, 2011); non-structural item skip and structural missingness. Item skip refers to a rater skipping one or several items in a non-structural manner. There are numerous reasons why a rater may skip an item: it could have been an accidental skip (e.g., missing an item at the bottom of the observation form) or it could be due to the content of an item, such as observing the ability of a child at building with blocks, without blocks being present at the location. In case of structural missingness, the missingness of responses has a underlying assumption or mechanism defined by the researcher, that explains the missingness. For example, a researcher might provide young children with a different subset of items than older children (either due to expected differences in ability level, or due to the formulation of items). In this case all items do apply to both groups, but not all items are administered to both groups. This results in structural missing values.

Little and Rubin (2002) distinguished three types of missing-data mechanisms; Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). When the missing data is independent of all observed variables and the unobserved values, the missing data are MCAR. This implies that the cause of the missing data is unrelated to the data itself. When the data is MAR, the missing data depends on one or more observed variables, but is independent of unobserved data. An example of MAR is when the observed variable gender is associated with a higher percentage of missing data on questions or items relating to anxiety for women, in which case we can use the variable gender as covariate to help us explain the missing data. In case of NMAR the missing data are dependent on the missing data itself (e.g. the ability we are trying to measure) or on another, unobserved variable. Suppose that gender was not observed in the earlier example, then the missing data are NMAR. This can cause several problems, as the cause of the missing data is unknown or unmeasured. This is why it is important to take multiple variables into account by adding them as covariates to improve the feasibility of the assumption of MAR.

This study focuses on the differences between handling structural missing data and non-structural missing data. The main interest is comparing several methods of dealing with these missing values and their impact on exploratory psychometric analyses of an observational instrument. This introduction chapter will present the choice of observational instrument, the methods for dealing with missing data and the analyses we will perform to assess the psychometric properties of the chosen observational instrument.

1.1. Visual Assessment Scale

As a motivating example, we chose an observational instrument that includes both structural and non-structural missing data. This resulted in the choice of the Visual Assessment Scale (VAS; see Appendix A). This is an observational instrument containing 45 dichotomous (fail/pass) items aimed to measure the visual functioning of patients with visual impairment caused by brain damage during (or shortly after) birth. This is also known as cerebral visual impairment (CVI; Frebel, 2006). CVI patients are often affected by profound intellectual and multiple disabilities (PIMD), including cognitive and physical disabilities (e.g. quadriplegia, intellectual disabilities, psychomotor disabilities, epilepsy). This makes assessing their visual functioning more difficult than for other patients. Patients with CVI are often non-verbal and unable to follow instructions, which often results in non-structural missing data. Additionally, the VAS has a predefined clustering of items (based on theoretical background) into six levels of visual functioning, which increase in difficulty (e.g. items that belong to the first level are easier to answer than items of the second level) (see Appendix A). Raters assign a level of visual functioning to the patient, based on the responses of the patient to items on that level (e.g. if most items that belong to level one of visual functioning are a pass, then the patient has reached level one of visual functioning). Once a level of visual functioning is not reached, the remaining items in higher levels of visual functioning are assumed to be fails as well, but they are never observed and thus missing. In other words, the missing values on items of higher levels is in this case structural and forced by the questionnaire design. We cannot assume that the structural missing values of the VAS are MAR, although they are dependent on an observed variable. This is because the raters applied a forced cut-off after which no other items were answered by any of the patients with a similar score. For example a patient that has reached level one of visual functioning and fails items in level two of visual functioning, will never have data available on level three/four/five/six of visual functioning. The chance that this patient has structural missing data on items in level three or higher is 100%. This indicates that missing data mechanisms such as NMAR/MAR no longer apply as the structural missing data are deterministic. For the non-structural missing data, covariates are available to make the assumption of MAR more feasible.

1.2. Psychometric Evaluation

To investigate the psychometric quality of the VAS, the reliability (i.e. Cronbach's alpha; Cronbach, 1951) and construct validity will be assessed. Construct validity can be assessed by applying an item response theory (IRT) model. In IRT models responses reflect the underlying ability that we are attempting to measure. This underlying ability is also known as the latent trait. For dichotomous items and small sample sizes, the Rasch model (Rasch, 1960) is the recommended choice (Chen et al., 2013; Fischer & Molenaar, 1995). In a Rasch model, the difficulty of an item (β) is modeled as a function of a latent trait (Rasch, 1960). The latent trait levels of respondents are reflected by a dimension called theta (θ). In the Rasch model the probability of a patient passing an item is influenced by the trait level of the patient as well as the difficulty of the item (Furr & Bacharach, 2014). A common formulation of the response function of the Rasch model is;

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}, \quad (1)$$

where $P(X_{is} = 1|\theta_s, \beta_i)$ is the probability that response $X = 1$ (which in this case is "Yes") on item i by patient s, given the trait level of the patient (θ_s) and the difficulty of the item (β_i). The specific IRT parameters are estimated from the observed data. Besides these parameter estimates, fit indices can be estimated as well, which are explained below.

First we would like to look at the difficulty parameter of items. The difficulty of the item represents the amount of latent ability required to have a probability (P) of 0.50 to pass the item (X = 1). For the VAS we expect that items that belong to a higher level of visual functioning, require more visual functioning (a higher latent trait) to be passed. To assess how well items measure the latent trait we can look at a fit index known as item fit. This index compares the observed response with the expected response given the difficulty of the item and the θ of the patient. If this difference between observed and expected response on one item is large, the item does not fit well and might not measure the same latent trait we are intending to measure (e.g. an item that people with a high ability fail, but people with a low ability pass).

A similar index can be calculated for individual patients. This index is known as the person fit index. A person that passes only items with high difficulty, but fails items with a low difficulty is indicated as a person misfit (i.e. the observed response pattern of this patient does not match the model expected response pattern). While item fit and person fit are good indices for specific items or individuals, we would also like to know how well the Rasch model fits the data. This is done by estimating the maximum likelihood of the parameters given the observed data, also known as the

model fit. In section two of this thesis a more detailed explanation and calculation of the Rasch model, the difficulty parameter and fit indices, are given

1.3. Methods for handling missing data

To investigate the influence of the non-structural and structural missing values on these parameters, three different methods for handling missing data will be used. The first method is treating all missing values as fail (MAF). This method is the instructed method for handling missing data in the VAS. This method relies on the assumption that if an observation is missing, it was not observed, so it is scored as a fail. If a patient fails items of a lower level of visual functioning, we assume that more difficult items are fails as well. However, non-structural missing values do not adhere to this assumption since the cause of the missing value is not related to the ability of the patient. Treating these non-structural missing values as fail can lead to less accurate, biased parameters in IRT models (He & Wolfe, 2012) than ignoring or imputing them. Treating missing values as fails results in higher difficulty parameters for items with many missing values, and overall underestimation of patients' ability. As this is the default method for how missing data are currently handled according to the VAS instructions, this method can be used as a baseline for comparing the other methods with.

The second method is known as full information maximum likelihood (FIML; Ferro, 2014; Peyre, Leplège, & Coste, 2011). It is the most common method for handling missing values in IRT and provides unbiased parameters and unbiased confidence intervals (Finch, 2008; Forero & Maydeu-Olivares, 2009). This method is based on defining a Rasch model for the observed data, using the available responses and response patterns by means of maximum likelihood (ML). The process of calculating maximum likelihood will be elaborated in detail in section two of this thesis.

The third method involves using multiple imputation (MI; Rubin, 1987). This is a method that replaces (imputes) the missing values with multiple plausible values. This results in multiple plausible complete versions of the incomplete dataset. These plausible complete datasets are analyzed separately and the results are combined into one overall analysis, using specific combination rules, defined by Rubin (1987). Rubin provides the following rule to calculate the mean of pooled parameters: Let $\bar{Q}$ be the pooled parameter, *M* the total amount of imputed datasets and $\hat{Q}_m$ the parameter estimate of each dataset, then the mean pooled parameter is given by (Rubin, 1987);

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m \,. \tag{2}$$

To test the parameter $\bar{Q}$, we require an associated standard error. The standard error of the pooled parameter can be calculated by taking the square root of the total variance. To calculate the total

variance, we need to combine the within-imputation and between-imputation variance. The within-imputation variance of the parameters estimates is calculated as

$$\bar{U} = \frac{1}{m}\sum_{m=1}^{m} U_m \ , \tag{3}$$

where U_m is the variance of the parameters in imputed dataset *m*. The between-imputation variance is calculated by

$$B = \left(\frac{1}{m-1}\right)\sum_{m=1}^{m}(Q_i - \bar{Q})^2. \tag{4}$$

Finally, the total variance is computed as

$$\bar{T} = \bar{U} + B + B/m\,. \tag{5}$$

MI in Rasch is often done by first estimating θ for a model with missing data. To include the uncertainty of our θ estimates of the Rasch model, we randomly draw multiple plausible θ values from a distribution of θ values for each patient, instead of using the point-estimate θ value. Drawing plausible θ values from an IRT model and using this as base to perform multiple imputation is known as plausible value multiple imputation (PVMI). The distribution of the plausible θ approximates the *sample* θ distribution ($F(\theta)$), with associated likelihoods for each θ value. It can be described mathematically by the response pattern *x* (a vector of passes and fails) and θ of the patient forming the item response probability of $f(x|\theta)$ and the given sample θ distribution $F(\theta)$. It can then be shown that the posterior distribution $h(\theta|x)$ is given by

$$h(\theta|x) = \frac{f(x|\theta)F(\theta)}{\int f(x|\theta)F(\theta)d\theta}\,. \tag{6}$$

This implies that if a patient has response pattern *x* the posterior distribution of the patient is given by $h(\theta|x)$. The plausible values are random draws from the probability distribution with the density of $h(\theta|x)$. Subsequently, the response data are imputed based on the plausible θ values and FIML-estimated parameters. This is done by forming the probability of selecting a particular response category given the plausible θ value of a patient and randomly sampling the responses given the probability weights (Chalmers, 2012).

MI is known to provide similar parameter estimates as FIML (Ferro, 2014). An advantage of using MI as a method for handling missing data is that it allows investigating multiple scenarios that are created due to the uncertainty of the model parameters. An advantage of multiple imputation over

FIML is that it can include covariates in the process of handling the missing data. If the missing data depend on observed covariates, they can be included in the imputation model. This prevents covariates from being part of the analyses model, which is not possible for the FIML-method. Including covariates that provide information of the missing data, makes the assumption of MAR more likely and may provide less biased estimates than when covariates are ignored, especially for items with a large amount of missing data.

In the current study we considered the a-priori estimated level of visual functioning as a possible covariate for explaining the missing values. This level of visual functioning might be able to explain part of the structural missing values, as patients with higher a-priori levels of visual functioning should have fewer missing values than patients with low a-priori levels of visual functioning. The CVI criteria are also used as a covariate, since it is hypothesized that more criteria is related to a lower level of visual functioning.

1.4. Research Question

The goal of this study is to compare the impact of three different methods for handling missing data (treating missing values as fails, full information maximum likelihood and plausible value multiple imputation) on the psychometric properties of an observational instrument that contains both structural and non-structural missing values.

The three methods are compared on the difficulty parameters of the items and four indices that are often used to describe the psychometric properties of an instrument: Cronbach's alpha, item fit, person fit and model fit. By comparing the differences in these values between methods, we can observe how big the influence is of the missing data (and the way they are handled). We can assess if the assumptions the VAS instructions implicitly make about treating the missing values as fail were correct. If the different methods for handling missing data give substantially different results with respect to the psychometric properties, then the missing data contain important information that cannot be ignored and the cause of the missing data needs to be investigated. The observational instrument VAS provides us with the opportunity to explore the impact of different methods for dealing with missing data on the difficulty parameter of items, because the instrument was developed with predefined clusters of items with similar difficulty (on theoretical grounds). The presence of structural and non-structural missing values allows us to assess the different methods for handling missing data on two types of missing values.

It is hypothesized that scoring missing values as fails will lead to higher difficulty parameters, more person and item misfit and a worse overall model fit, especially for items and patients that have a large amount of missing values. FIML and PVMI are hypothesized to provide similar difficulty parameters and represent the theoretically based clustering of item difficulties better than treating

missings as fail. The PVMI method allows adding covariates to the model, which might be able to partially explain non-structural missing values. If this is the case, the inclusion of the covariate in the model may improve the estimates of the difficulty parameters of items with few response patterns. As a result, the item fit and person fit would also improve.

In the next section we will elaborate further on the Rasch model and the associated parameter and fit indices. In the method section we will elaborate on the procedure to compare the different methods of dealing with missing data and provide more detailed information about the data used to perform this study. The result section will describe the results of using different methods for handling missing data on the VAS data. Finally the conclusions, limitations and practical implications of the study will be described in the discussion section.

2. The Rasch Model

By fitting the Rasch model to the observed data the difficulty parameters (β_i) of items are estimated using a method called marginal maximum likelihood (MML; Wright & Masters, 1982). Marginal maximum likelihood maximizes the likelihood of model parameters given the observed responses. The person abilities, θ, are modeled as a sample from a normal distribution, $F(\theta)$(with a mean of 0 and a standard deviation of 1), for the purpose of estimating the item parameters. The maximum likelihood of the model parameters can be estimated using the Expectation-Maximization (EM; Dempster, Laird & Rubin, 1977) algorithm. The EM algorithm performs two steps: 1) the Expectation (E) step, which calculates the log-likelihood for the current parameter estimates, followed by 2) the Maximization (M) step, which maximizes the expected log-likelihood from step E by computing new parameters. These two steps are repeated until successive iterations do not improve the log-likelihood of the parameters anymore. This results in a maximized likelihood function and associated model parameters. The maximized likelihood function allows us to estimate θ of individuals using the Expected a Posteriori (EAP) estimator (Bock & Mislevy, 1982). The EAP estimate of the ability of a person ($\hat{\theta}s$) is approximated by (Bock & Mislevy, 1982):

$$\hat{\theta}s = \sum_{k=1}^{q} X_k L_s(X_k) W(X_k) \; / \; \sum_{k=1}^{q} L_s(X_k) W(X_k) \, , \tag{7}$$

where X_k is one of the q quadrature points and $W(X_k)$ is the weight associated with that quadrature point (based on the density of the prior distribution $F(\theta)$) and L_s is the likelihood function at this quadrature point of this patient.

Using the estimates of the difficulty parameters and latent trait levels of patients we can calculate the expected response on an item, using the response function formula from Equation 1. To obtain item residuals we subtract the expected response (e_{ni}) from the observed response (x_{ni}). These item residuals can be used to calculate item fit statistics. Item fit statistics represent how well an item fits the observed data. For Rasch models there are residual-based outfit and infit statistics (Hohensinn & Kubinger, 2011). These outfit and infit statistics can be used to determine which items do not fit the Rasch model adequately. Using standardized residuals ($Z_{ni} = (x_{ni} - e_{ni}) \; / \; \sqrt{\mathrm{Var}(x_{ni})}$) outfit mean-squared error (MSQ) and infit MSQ can be calculated. The outfit MSQ is the averaged sum of squared residuals of an item;

$$\mathrm{o}_i = \sum_{n=1}^{N} \mathrm{Z}_{ni}^2 \; / \, N \, . \tag{8}$$

While the outfit MSQ does not account for the amount of variance of the item responses, the infit MSQ does. The infit MSQ weighs the mean-squared error according to the variance of the response ($\mathrm{Var}(X_{ni})$);

$$\mathrm{i}_i = \sum_{n=1}^{N} \mathrm{Var}(X_{ni}) * \mathrm{Z}_{ni}^2 \; / \; \sum_{n=1}^{N} \mathrm{Var}(X_{ni}) \,. \tag{9}$$

Bond and Fox (2007) suggested that fit values < 0.75 indicate overfit and fit values > 1.3 indicate underfit. A mean-square of 1.3 indicates that there is 30% more randomness in the data than the Rasch model expects. A mean-square of 0.75 indicates a 25% deficiency in Rasch-model-predicted randomness. This implies that the item discriminates better than expected by the probabilistic Rasch model, which could be cause for alarm. The probabilities of passing the item are then no longer based on the Rasch model, but solely on the estimated theta ($\hat{\theta}$). In this case the difficulty parameter of the item could be substantially different for people with low $\hat{\theta}$ than for people with a high $\hat{\theta}$. This is also known as differential item functioning. Outfit statistics are dominated by unexpected outlying, low-information responses and is outlier-sensitive. Infit statistics are less influenced by single extreme outlying cases, because they are weighted by item variance. Item variance is higher near the mean difficulty and lower at the extremes.

In addition to item fit we can also inspect person fit by using person fit indices. Levine and Rubin (1979) defined the person fit statistic l_0 as;

$$l_0(\theta_s) = \sum_{i=1}^{n} [u_{is} \ln P_i(\hat{\theta}_s) + (1 - u_{is}) \ln Q_i(\hat{\theta}_s)] \,. \tag{10}$$

Here the likelihood (l_0) of patient *s* with ability θ responding *u* (pass or fail) to item *i* is calculated, where $P_i(\hat{\theta}_s)$ is the probability of giving that response to that item (P_i) given the estimated theta of patient ($\hat{\theta}_s$). However the l_0 statistic is conditionally dependent on the $\hat{\theta}$. To counter this dependence l_0 was standardized. The standardized person fit index l_z, (Drasgow, Levine & Williams, 1985) is given by

$$l_z = \frac{l_0 - \mathrm{E}(l_0)}{[\mathrm{Var}(l_0)]^{1/2}} \,, \tag{11}$$

where

$$E(l_0) = \sum_{i=1}^{n} P_i(\hat{\theta}_s) ln\, P_i(\hat{\theta}_s) + [1 - P_i(\hat{\theta}_s)] \ln[1 - P_i(\hat{\theta}_s)] \,, \tag{12}$$

$$\mathrm{Var}(l_0) = \sum_{i=1}^{n} P_i(\hat{\theta}_s)[1 - P_i(\hat{\theta}_s)] \left\{ \ln \left[\frac{P_i(\hat{\theta}_s)}{1 - P_i(\hat{\theta}_s)} \right] \right\}^2 . \quad (13)$$

This is calculated for each item, summed across all items, and then standardized to get the l_z statistic. The l_z-statistic can be used to determine person misfit. Person misfit represents a person that has an unlikely response pattern (e.g., passing difficult items that require a high visual functioning, while failing items that require a lower visual functioning).

Additionally, we can assess the model fit to the data as a whole by using likelihood based indices. Instead of maximizing the likelihood of a model, we choose to minimize the negative of the natural logarithm of the likelihood function as it is more convenient (as this logarithm monotonically increases). This is called the log-likelihood. A lower log-likelihood represents a better fitting model. The log-likelihood does not take the amount of parameters into account, nor does it provide a test for comparing the model fit of two models with the same amount of parameters. Consequently, we have to include additional model fit indices.

The Akaike Information Criterion (AIC, Akaike, 1974) is based on the log-likelihood, but takes the amount of parameters of the data into account. It is possible to compare models with the AIC due to the addition of correction for amount of parameters in the model. Suppose that the *k* is the amount of parameters $\hat{L}$ is the maximum value of the likelihood function, then the AIC is calculated as follows:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) . \quad (14)$$

A similar information criterion is the Bayesian Information Criterion (BIC; Schwarz, 1978). The BIC has a larger penalty for adding more model parameters. Given the number of observations, *n*, the BIC is calculated as;

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}) . \quad (15)$$

For both the AIC and the BIC a lower value indicates better model fit.

By calculating the difficulty parameter and fit indices for all the methods for handling missing data, we can compare the effects of these methods on the psychometric analyses of the VAS.

3. Importance of the VAS validation

Currently CVI is the number one cause of visual impairment in the western world (Khan, O'Keefe, Kenny & Nolan, 2007). Due to improvements in medical research new treatments have been developed for optical visual impairments, while it also increased the survival rate of children with CVI. Optical visual impairments have standardized measurement techniques to determine a patients' functional vision. CVI patients are often affected by PIMD, which makes assessing their visual functioning more difficult. These patients are often non-verbal and unable to follow instructions. It is important to measure the visual functioning in patients with CVI, to allow professionals to provide better services for their patients. By measuring visual functioning in CVI patients professional can discriminate between patients with only cortical visual impairments and patients with both cortical visual impairments as well as optical visual impairments. Weinstein et al. (2012) mention motion processing as one of the distinctive CVI features that separate CVI patients from non-neurological patients. Nakken and Vlaskamp (2007) emphasize the importance of standardized assessments for patients with PIMD, which includes CVI patients.

Several tools have been developed for assessing the visual functioning of CVI patients. However, none of them have been validated in a clinical sample. One of the first tools developed is the Individualized Systematic Assessment of Visual Efficiency, ISAVE (Langley, 1998). The ISAVE contains screening of a patients' visual functioning, divided into separate areas such as acuity, visual field and attention testing. The ISAVE also includes a CVI assessment protocol to determine the presence of CVI (Langley, 1998). However, the reliability and validity of the ISAVE has never been assessed. Roman-Lantzy (2007) developed the CVI Range, a tool specifically designed for patients with CVI. The CVI Range is based on previous literature and descriptions of distinctive behavioral traits of CVI patients. The CVI Range includes an observational form, a parent/guardian interview and direct assessment. The reliability of the CVI Range has been assessed by Newcomb (2010); the internal consistency and test-retest reliability were good. Assessment of the validity of the CVI Range however, was to our knowledge, never conducted. A different study by Ortibus et al. (2011) developed a closed-ended questionnaire to screen for CVI. This questionnaire was completed by the

parents/guardian of the patient prior to neuropsychological assessment. This questionnaire has a good discriminate validity, but ocular impairment is assessed separately with neuro-ophthalmological evaluation.

The VAS is, to our knowledge, the first measurement instrument intended for CVI patients that will be validated using modern psychometric techniques. The importance of the development of the VAS is connected to the importance of the way missing data are handled, because VAS data often contains a large amount of missing values. This is due to children with CVI often suffering from PIMD. This makes it difficult to score all items, which often results in missing data.

4. Method

4.1. Empirical Data Application

Patients with CVI of the Koninklijke Visio clinic in Den Haag (N = 73) were retrospectively assessed on their visual functioning using the VAS. The VAS was completed by counselors of the Koninklijke Visio, based on documentation (progress reports, diagnostics and logs) and observations made of the patient, during a period spanning one or more years. Patients often suffered from multiple disabilities including mental retardation and physical disabilities. The age of the patients ranged from six months to 22 years (M = 9.3, SD = 5.39). The VAS is a scale that is intended to measure visual functioning in patients with CVI. The 45 items of the VAS are divided into six different levels of visual functioning (at a developmental age of 24 months) and are administered from lowest to highest level. These six levels are subsequently described; Blind/fully visually impaired (1), functionally blind/severely visually impaired (2), passive visual attention/badly visually impaired (3), basal perception/moderately visually impaired (4), expansive visual recognition/slightly visually impaired (5) and normal visual functioning/no visual impairment (6). Structural missing values are introduced into the VAS data when raters assign a level of visual ability to the patient and do no longer rate items above this level of visual ability. Non-structural missing values are often introduced by the fact that patients with CVI often have PIMD, which causes observational items to be difficult to score, especially when the patient is unable to follow instructions. Rating children as observer with the VAS requires experience with children with CVI, as well as practical training on recognizing the characteristics/traits that are included on the observational form. In addition to the VAS data, our data also contain a list of nine CVI criteria (dichotomous) to assess whether or not the patient has CVI.

4.2. Data Preparation

To prepare the data for IRT modelling the questions have to be aligned so that a fail on any item would represent a lower level of visual functioning and a pass would indicate a higher level of visual functioning. Negatively worded items were recoded into the correct direction, such as the first item of the VAS; “Shows no sign of visual reactions, even in visual stimulation chamber.”. Passing this item would indicate *worse* visual functioning, so the item had to be recoded. Another item (item 3.3a) has a follow-up item associated with it (item 3.3b), which requires a different recoding scheme. The first item (“Shows fixated visual functioning *during daylight*, especially with strong visual stimuli”) requires a higher level of visual functioning than the follow-up item (“Only sees these visual stimuli when they are offered within the visual field of the patient.”), but the second item is dependent on the first item to be answered. If the first item is answered with a pass, this implies that the patient can fixate on strong visual stimuli during daylight, regardless of whether it is offered within the visual field. However, if a patient can fixate on visual stimuli outside of the visual field (as is implied by the

first item), he/she can also fixate on stimuli offered within the visual field. The word "only" causes a problem for IRT as the item is theoretically easier than the first item, but the patient fails this item if he/she can fixate on stimuli outside of the visual field. The item pair was recoded in such a way that if a patient had a pass on both items, he/she could *only* fixate on stimuli when they were offered in the visual field, resulting in a pass for the item: "Only sees these visual stimuli when they are offered within the visual field of the patient." and a fail on the other item. Patients that can fixate on visual stimuli outside of their visual field can also fixate on stimuli within their visual field, which resulted in recoding a pass on the first item and a fail on the second item to a pass on both items. Fails on both items remained as fails on both items.

4.3. Design and Procedure

From the original dataset, two datasets were created: One in which all missing values were coded as missing and one in which the non-structural missing values were coded as missing and the structural missing values were coded as fails. To distinguish the non-structural missing values from the structural missing values, raters were asked to only use the response category "no information available" for non-structural missing values. For structural missing values raters simply stopped rating items (blanks). In total, three methods were used to deal with the missing data: scoring missing values as fails (MAF), full information maximum likelihood (FIML), and plausible value multiple imputations (PVMI) with covariate (m = 10). The included covariate is the number of CVI criteria (on a scale of one to nine) present in the patient. The covariate a-priori level of visual functioning was not used in the analyses, as there was insufficient overlap between different level of visual functioning. All three methods were applied to the two different versions of the dataset (non-structural missing as missings, non-structural missing as fail), which results in the combinations shown in Table 1

Table 1. Design matrix with types of missing data and methods for handling those missing data.

	Combinations of Methods				
Type of missing values	1	2	3	4	5
Non-structural missing data	MAF	FIML	FIML	PVMI	PVMI
Structural missing data	MAF	MAF	FIML	MAF	PVMI

Note; MAF, missing values as fails; FIML, full information maximum likelihood; PVMI, plausible value multiple imputation.

There are two assumptions for fitting a Rasch model (Yang & Kao, 2014; Wright, 1995).The first assumption is that the observational form represents one latent trait (θ). This is known as the unidimensionality assumption. Unidimensionality was assessed using the Martin-Löf test of unidimensionality (Martin-Löf, 1973) as implemented in the R-package "eRm" (Mair & Hatzinger, 2007). This test splits the data into two subsets (with i_1 and i_2 items respectively) and calculates the maximum likelihood associated with the two subsets. The null-hypothesis is that both subsets tap into

the same dimension and the product of maximum likelihood of the two subsets approximately equals the maximum likelihood when calculated on both sets together. The likelihood-ratio test that is performed to test this approximates a chi-square distribution with $i_1\ i_2 - 1$ degrees of freedom. If the Martin-Löf test yields a *p*-value > .05, the hypothesis of unidimensionality cannot be rejected. The second assumptions is that a patients' responses to the items are not statistically related to each other; the difference in the responses should be explained solely by differences in the latent trait. This assumption is called the local independence assumption and it is checked by inspecting the residual correlation between items. If an item pair violates local independence we could decide to delete one of the items, after looking at the item content. A residual correlation above 0.20 is a strong indication of local dependence.

Once the assumptions were checked a unidimensional Rasch model was fitted to each dataset using MML. The Rasch model was fitted using the mirt package (Chalmers, 2012) in R (R Core Team, 2016). To assess and compare the five methods for handling missing data with each other the model fit, item fit, and difficulty parameters were estimated using the mirt-package as well. Additionally the person fit was estimated using the PerFit package in R (Tendeiro, Meijer & Niessen, 2016).

Model fit was assessed using the log-likelihood, BIC and AIC. The lower the log-likelihood, the AIC and the BIC the better the model fits. This way we can rank the models based on their model fit. For a comparison of models, the AIC was used in accordance with the following formula of Burnham & Anderson (2002):

$$\Delta AIC = AIC_{m1} - AIC_{m2}\,, \qquad (16)$$

where AIC_{m1} stands for the AIC of model 1 and AIC_{m2} is the AIC of model 2. A ΔAIC higher than 10 is considered a substantial difference in models (Burnham & Anderson, 2002).

Item fit was assessed using infit and outfit mean-squared error statistics. To judge the infit and outfit mean-squared error statistics the amount of underfit (>1.3) and overfit (<0.75) items between methods and within methods (Hohensinn & Kubinger, 2011) was compared. The cause of infit and outfit was assessed by ordering patients' responses by their estimated θ (Linacre & Wright, 1994).

The l_z statistic was estimated (Drasgow et al., 1985) to assess person fit. A value of l_z = -1.645 is normally used as a theoretical cut-off score for person misfit (Seo & Weiss, 2013). To assess the effect of different methods for dealing with missing data, the amount of patients that misfit the data and the severity of the misfit (e.g. lower number indicates a stronger misfit) were compared.

The difficulty parameters of each item were estimated using MML. The difficulty parameters were used to determine the extent to which the theoretical increase in difficulty of items across the

VAS could also be found empirically. As the VAS consists of items divided into six levels of visual functioning, we expected six blocks of clustered item difficulties. The items can vary in difficulty within each block, but should be more difficult than any item from the previous block.

Additionally, Cronbach's alpha was calculated for each method for handling missing data. The discrepancy between alpha coefficients among methods for handling missing data, were tested using a *t*-test statistic for dependent samples. Given two alpha coefficients from two dependent samples with *S* amount of subjects, α_1 and α_2, and the squared correlation of total test scores ρ^2 the *t*-statistic is calculated as (Feldt, 1980);

$$t = \frac{(\alpha_1 - \alpha_2)(S-2)^{1/2}}{[4(1-\alpha_1)(1-\alpha_2)(1-\rho^2)]^{1/2}}, \quad (17)$$

with DF = $S - 2$. If there is a significant discrepancy between alpha coefficients this indicates that one of the models (e.g. one of the methods for handling missing values) gives a stronger internal consistency than the other model.

4.4. Practical Implication of Results

The present study provides us information about which items of the VAS perform poorly, by assessing the difficulty parameters and item fits. As we have predefined clusters of items, within which we expect the item difficulties to be similar, we may consider moving certain items into a lower or higher cluster of visual functioning. Item fit allows us to check if the item contributes additional information to measuring the visual functioning latent trait. If an item has a poor item fit, this indicates that it might warrant removing as it does not contribute (positively) to the measuring of visual functioning.

Apart from changing, moving or removing items this study can also be used to develop a new scale of visual functioning of patients, using the estimated theta values ($\hat{\theta}$). We can first check if the theta values accurately represent the level of visual functioning by correlating the θ estimates of patients with their assigned level of visual functioning using Spearman's Rho. Subsequently, we can transform the $\hat{\theta}$ to form a new interval scale of visual functioning, which could provide more detailed information about patients than the ordinal levels of visual functioning, as the scale is not limited to six levels.

For the practical implication of the VAS inter-rater reliability was also assessed (for a subsample of forty patients) for the overall VAS scale (e.g. the assigned level of visual functioning) and the number of CVI criteria. Inter-rater agreement is assessed using Cohen's κ, which represents the agreement between the scoring of all patient between observers. Cohen (1960) suggested the following cut-off scores for κ: ≤ 0 represents no agreement, 0.01-0.20 is none to slight, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is substantial and 0.81-1.00 is almost perfect agreement.

This study will hopefully contribute to both improving the VAS and its psychometric properties and give more insight into differences between methods for handling missing data in the presence of structural and non-structural missing values.

5. Results

The assumptions for a Rasch model were checked for the default method of handling missing data as fails. The assumption of unidimensionality was not rejected, χ^2 (360) = 86.53, p = .99. The criteria for local independence were met; no item pairs displayed residual correlations higher than .2. Two items were removed from further analyses because they did not have any variation in answers: item 6.3 ("Understands part/whole relations (e.g. recognizes a bike by only the handlebars)") and item 6.7. ("Interest in details (including richly illustrated pictures). Can easily find something within this picture. (good selective attention/visual scanning)"). These two items only contained fails, causing problems calculating the likelihood of the Rasch model. For each method for handling missing data a Rasch model was fit to the data. Model fit indices, internal consistency, item fit indices and person fit indices were calculated. The results for these indices will be described next.

5.1. Model Fit and Internal Consistency

The log likelihood, AIC and BIC of the five methods for handling missing are given in Table 2.

Table 2. Model fit criterion for different methods of dealing with missing data.

	MAF	FIML-MAF	FIML	PVMI-MAF	PVMI
Log Likelihood	-814.0	-669.3	-665.1	-713.3	-747
BIC	1816.8	1527.4	1519.1	1615.3	1682.9
AIC	1716	1426.6	1418.3	1514.5	1582.1
Cronbach's α	.957	.975	.864	.964	.963

Note: BIC, Bayesian information criterion; AIC, Akaike information criterion

As expected, the model fit was best for the method where all missing values (structural and non-structural) were handled by FIML. FIML maximizes the likelihood given the obtained response patterns, which results in a better fit, when there are fewer varying response patterns. PVMI-based methods had a better model fit than scoring items as fails. This could indicate that the missing values or the included covariate offer information about the respondents, which we do not receive when we simply treat every missing value as a fail. For both the full PVMI model (F(1, 71) = 81.03, p < .001) and the PVMI-MAF model (F(1, 71) = 78.41, p < .001) the covariate of total number of CVI indicators was influential on the $\hat{\theta}$.

Model fits between methods (FIML vs. PVMI) differed significantly as the ΔAIC between these models was higher than 10. The FIML methods outperformed the PVMI methods in terms of model fit (ΔAIC > 10). This was also seen when rank ordering the log likelihood and the BIC. Both the FIML and PVMI methods had a better model fit than the MAF method (ΔAIC > 10).

For the FIML method, no difference was found in model fit between handling non-structural missing values and handling both structural and non-structural missing values, For the PVMI methods the model fit was worse when structurally missing values were imputed as well. This could be an indication that these values should not be imputed.

All internal consistency values are high (>.85). For FIML-based methods the internal consistency was calculated for a database with missing data, using a pairwise deletion method. The FIML-MAF method has a significantly higher Cronbach's α than FIML ($t(71) = 56.84$, $p < .01$) and MAF ($t(71) = 16.40$, $p < .01$). A possible reason for this is that non-structural missing values (i.e. accidentally skipped items or items where no information is available) were not imputed for the FIML-MAF method. These non-structural missing values are independent of the latent trait of the patients, which means treating them as a fail results in a covariance matrix with lower values. For the FIML-MAF method, non-structural missing values do not contribute to the covariance matrix, leading to a higher α than the MAF method. However, when comparing full FIML method to the other methods, we can see that the full FIML method shows substantially lower Cronbach's α (all *t*-tests with a *p*-value < .01) than the other models. This is due to the high amount of structurally missing values (items not being administered due to being judged too difficult for certain patients). Treating structural missing values as fails has a positive effect on the Cronbach's α, as it increases the strength of covariances for items that previously had little information or few response patterns available. Treating non-structural values as fails in the FIML-MAF method has a negative effect on the Cronbach's α, compared to FIML. A possible explanation for this is that these non-structural values were often on items with a low difficulty parameter (high proportion of corrects). For these items replacing missing values with fails lowers the correlations, resulting in a lower Cronbach's α. The PVMI and PVMI-MAF methods did not differ significantly from each other, $t(71) = .08$, $p = .42$. PVMI and PVMI-MAF differed significantly from MAF ($t(71) = 4.49$, $p = < .01$, $t(71) = 5.31$, $p = < .01$), FIML ($t(71) = 41.68$, $p = < .01$, $t(71) = 42.68$, $p = < .01$) and FIML-MAF ($t(71) = 11.78$, $p = < .01$, $t(71) = 10.95$, $p = < .01$). The *t*-statistic is calculated with the correlations between test scores, which are extremely high for all methods (> .99). This resulted in small differences being statistically significant while Δα was less than .10.

5.2. Difficulty Parameters

Difficulty parameters for all five methods for handling missing data can be observed in Appendix B. The lowest and highest difficulty parameter for items within a (theoretical) cluster were used to describe the range of difficulty parameters. Table 3 shows the range of difficulty parameters for each cluster.

Table 3. Range of difficulty parameters by theoretical VAS clusters.

VAS Level	*N* items	Range β_{MAF}		Range $\beta_{FIML\text{-}MAF}$		Range β_{FIML}		Range $\beta_{PVMI\text{-}MAF}$		Rangeβ_{PVMI}	
1	1		-7.44		-7.69		-7.61		-7.70		-7.70
2	4	-7.44,	-4.90	-7.69,	-5.22	-7.61,	-5.13	-7.70,	-5.14	-7.65,	-5.19
3	9	-3.71,	-1.02	-4.32,	-1.26	-4.02,	-1.28	-4.34,	-1.28	-4.08,	-1.37
4	11	-0.88,	2.27	-1.11,	0.88	-1.10,	0.84	-1.14,	0.82	-1.16,	0.85
5	12	1.17,	4.67	0.93,	5.07	0.61,	4.91	0.91,	5.08	0.57,	4.98
6	7	4.43,	7.32	4.56,	7.52	3.86,	7.27	4.54,	7.46	3.94,	7.54

Note: Range is displayed as minβ, maxβ for all items within the theoretical cluster.

As the clusters are used in practice to differentiate between levels of visual ability, we expect no overlap of item difficulties between clusters. Using this method, items that do not fit the cluster they were theoretically assigned to can be easily identified as they will overlap with a higher or lower cluster. To visually demonstrate the difficulty parameters we use a Wright map. A Wright map shows the difficulty of items across the range and distribution of the latent trait. The Wright maps for each method can be seen in Figures 1 to 5. The items that displayed a difficulty parameter misfit to the cluster they were assigned to can are marked in red.

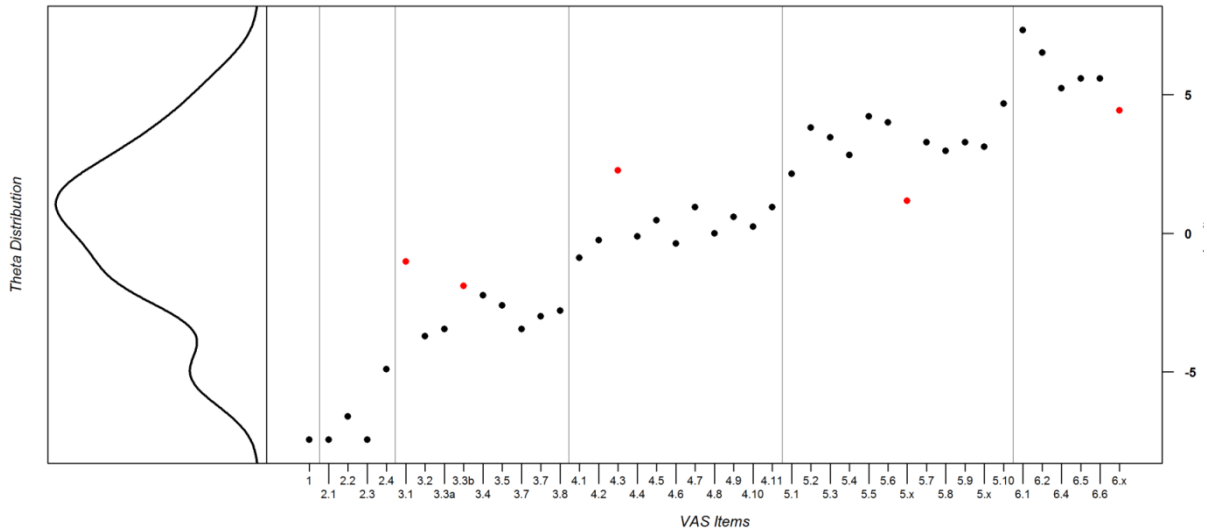

Figure 1. A Wright map of the VAS under the MAF method.

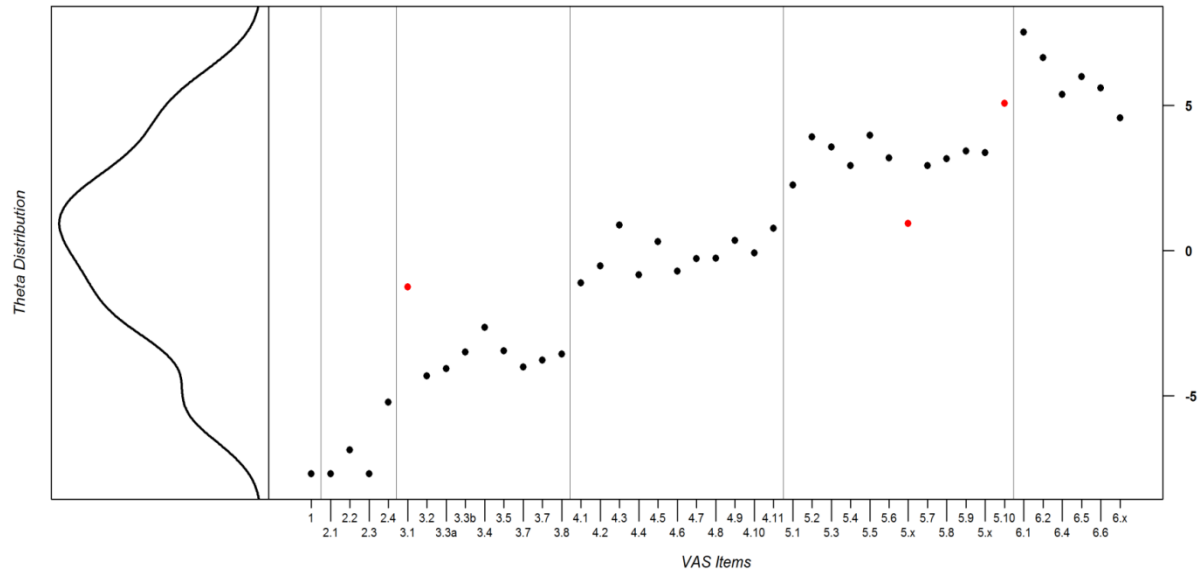

Figure 2. A Wright map of the VAS under the FIML-MAF method.

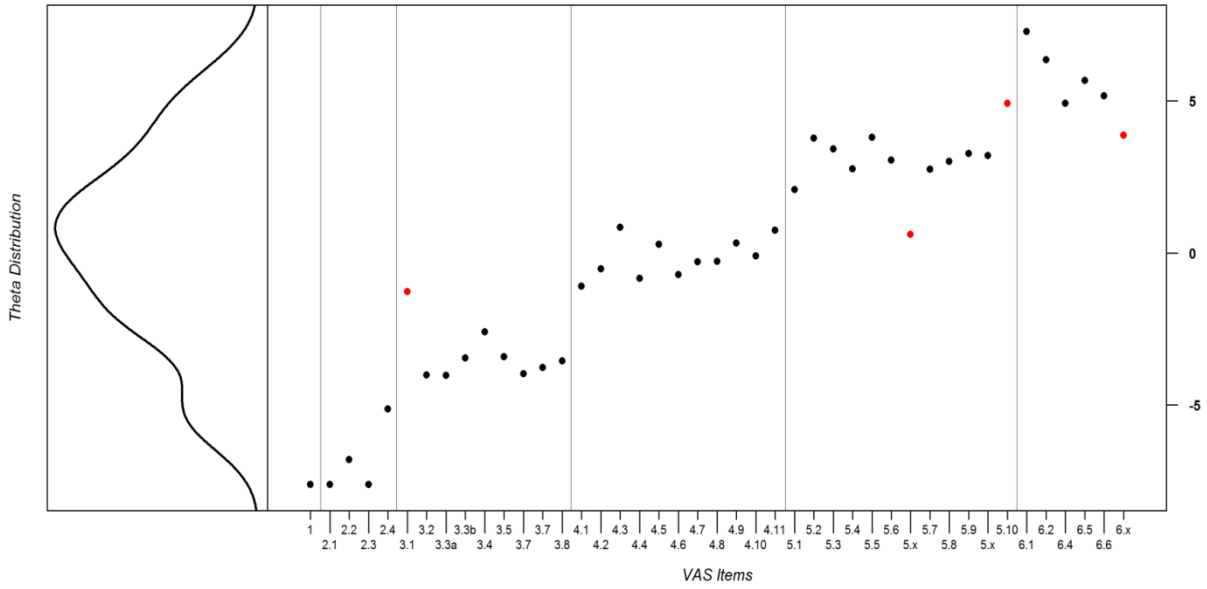

Figure 3. A Wright map of the VAS under the FIML method.

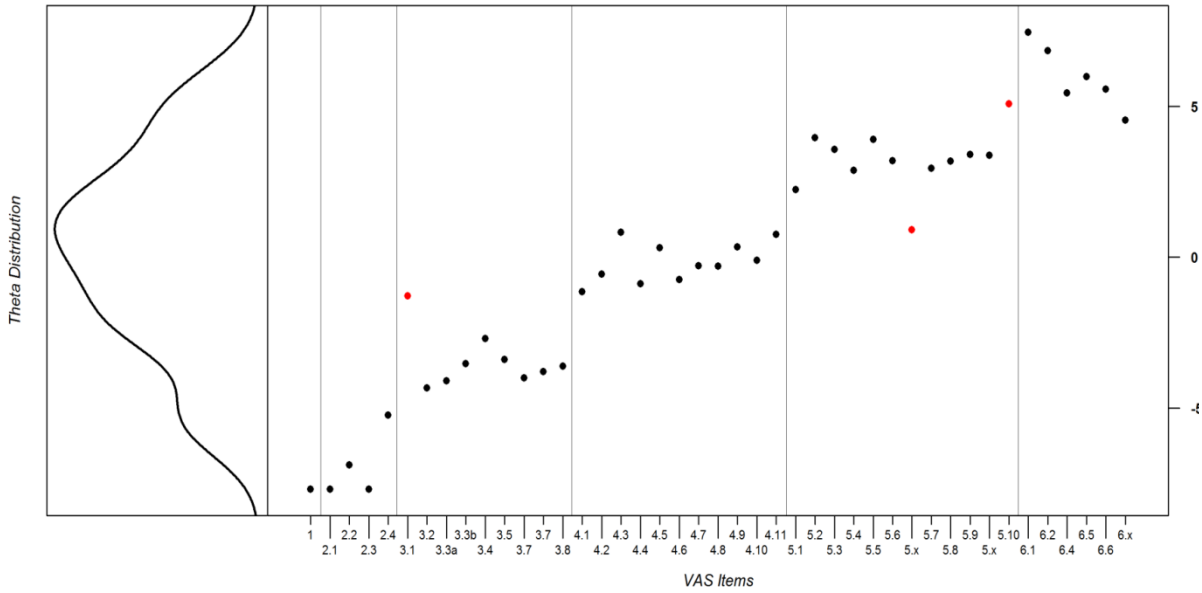

Figure 4. A Wright map of the VAS under PVMI-MAF method.

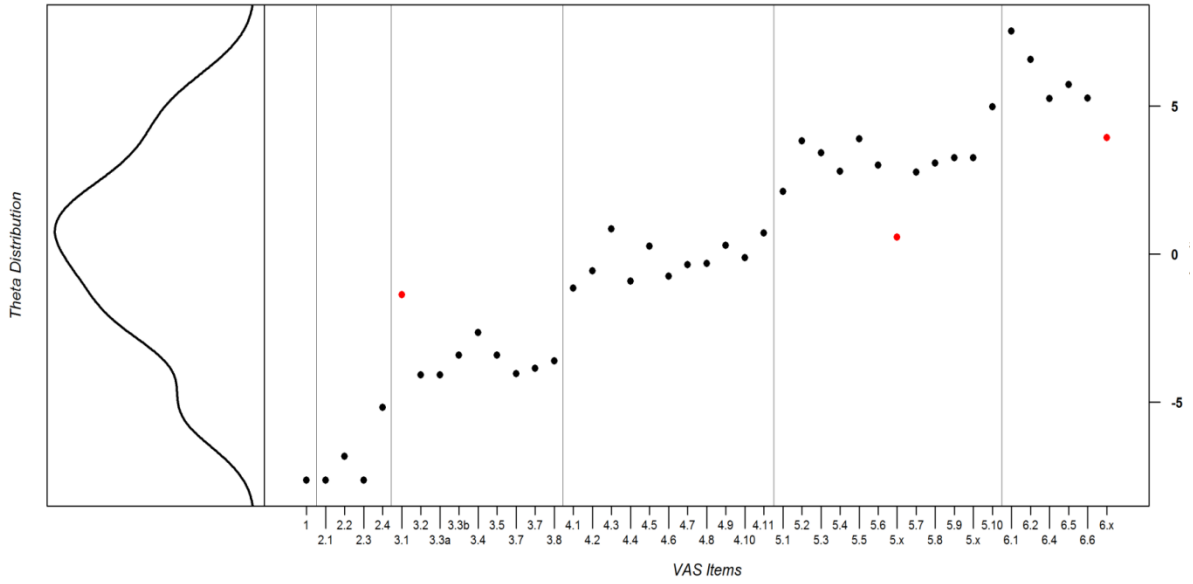

Figure 5. A Wright map of the VAS under the PVMI method.

5.2.1. Missing as fails

Using the MAF method the clusters often overlapped due to items with a high amount of (non-structural) missing values. These items have increased difficulty parameters when these missing values are replaced with fails. For the third level of visual ability a good example is question 3.3b; "Has visual attention mainly by auditory stimuli *during daylight*.". This item has 12.3% missing data, which results in a substantially higher difficulty parameter (β = -1.90) when replaced by fails, compared to the other methods (range β = -3.42 – -3.54). Another item from the third level of visual ability is 3.1 ("Shows fixated visual functioning *during daylight*, especially with strong visual stimuli."). This item is a clear outlier when compared to the other items in the cluster. However, this item only has 2.7% missing data. This indicates that the item difficulty of this item is barely influenced by the missing data, but is actually more difficult than expected. If we look at the response categories we can see that this is indeed the case, since 31.5% of the patients fail this item, compared to a range of 12.3% – 20.5% of all other items that belong to the third level of visual functioning as well. In the fourth level of visual functioning the item difficulty of 4.3 "Tracks toy that falls onto the floor (object permanence)." overlaps with item difficulties of the fifth level of visual functioning. Similar to item 3.3b, this item has many non-structural missing values (30.1%), resulting in a higher difficulty parameter (β = 2.27) when replaced with fails, but not when any of the other methods for handling missing data are applied (range: 0.82 – 0.88). This item has many non-structural missing values because information was not available about this specific behavior (e.g. there was no toy that could fall onto the ground present). The item "Recognizes familiars/family members visually (without voice)." has a much lower difficulty parameter (β = 1.17) than the difficulties of the other items in this cluster (range β = 2.14 – 4.67). This item seems much easier than expected, as the success rate of this item is 41.1%, compared to the 9.6% - 30.1% success rates in all the other items in this cluster. Finally, one item in the highest level of visual functioning has a lower difficulty parameter than expected, based on the cluster. The item "Can orient himself well in familiar surroundings." has a difficulty parameter of 4.43, while the range of difficulty parameters in this cluster is 5.23 – 7.32. The success rate of this item is much higher (11%) than the other items in this cluster (range 1.4%-6.8%).

Generally, the item difficulty parameters of items were inflated by the MAF method if items had many non-structural missing values. Structural missing values seemed to have a smaller impact, when compared to the difficulty parameters of the FIML and PVMI methods. Another noticeable difference is that the range of difficulty parameters is more limited for the MAF method (-7.44 – 7.32) than for the other methods (-7.70 – 7.50).

5.2.2. FIML and PVMI

While quite large differences between methods can be seen between MAF and any other methods, PVMI and FIML differences are small. Especially FIML-MAF and PVMI-MAF are nearly identical with respect to item parameters. This makes sense since the imputations of PVMI are based on the FIML model parameters and $\hat{\theta}$ (albeit randomly drawn from a posterior distribution).

There is a noticeable difference at the high end of the scale when non-structural and structural missing values are both handled by FIML or PVMI. All items that belong to level six of visual functioning have higher difficulty parameters for the PVMI method than for the FIML method. Item 6.4. ("Displays joint attention. Makes eye contact, points at an object or brings an object to show it."), has a lower difficulty parameter when all missing values are handled by FIML For the FIML-method this causes an overlap of item 5.10 (which also shows cluster misfit for FIML-MAF and PVMI-MAF methods) with item 6.4. This is due to that fact that there are more available response patterns at the higher end of the scale when missing values are imputed with PVMI. This can also be seen by comparing the highest difficulty parameters ($\beta_{\text{FIML}} = 7.27$, $\beta_{\text{PVMI}} = 7.54$). Another effect of imputation as well as treating missing values as fails, is that more response patterns become available at the high end of the scale. Item 6.4. only has a cluster misfit for the FIML method. The FIML method uses the few response patterns that are available for this item to base the difficulty parameter on. The response patterns for item 6.4. are only response patterns from patients with high (level five or six) levels of visual functioning. When patients from level five also pass a level six item, this lowers the difficulty parameter substantially, as for these items no other response patterns are available when applying FIML.

5.2.3. Structural vs. non-structural missingness

For the PVMI and FIML methods estimates of item difficulty were similar. However, differences were found between methods that only handle non-structural missing values (PVMI/FIML) and methods that handle both non-structural and structural missing values (PVMI-MAF/FIML-MAF).

The most noticeable differences are at the higher ends of the scale, because they contain more structural missing values. For the mixed methods these missing values were replaced by fails, which means that the rater did not consider the patient to be able to pass the item. For the full methods structural missing values were either not used (for FIML) or imputed (for PVMI). This causes a difference in the proportion of passes in items at high levels of visual functioning (where a high amount of structural missing values are present), which in turn results in lower item difficulties for the items that belong to a high level of visual functioning. The biggest impact between treating structural missing values as fails can be seen in item 6.x. "Can orient himself well in familiar surroundings.", where the difficulty parameters is considerably lower ($\Delta\beta = .60$) for FIML and PVMI than for

FIML-MAF and PVMI-MAF. This is caused by a combination of the differences in proportion of patients that pass the item and the amount of response patterns available in the higher end of the scale. The percentage of passes is lower for items at the higher end of the scale when structural missing values are treated as fails and more response patterns become available for the Rasch model as now all patients have complete data.

5.3. Item Fit

The outfit and infit statistics can be seen in Table 4.The outfit and infit statistics are residuals of the model, calculated as the difference between the expected value and the observed value. For each item we expect fails for patients with a low $\hat{\theta}$ and passes for patients with a high $\hat{\theta}$. In numbers we can display the pattern of the responses ranked on the $\hat{\theta}$ of patients for each item. For example, item 1 has only one fail, for the patient with the lowest $\hat{\theta}$. This means that the item discriminates perfectly and has a low infit (range 0.69-0.81) and outfit (range 0.08-0.11) statistic. We expect that items discriminate between low and high ability patients reasonably well (high/perfect discrimination leads to overfit). A pattern for a low difficulty item should only have fails for patients with a low $\hat{\theta}$. A high difficulty item on the other hand should have fails for most patients, except the ones with a high $\hat{\theta}$. As example of item misfit we can investigate item 3.5. “Can show indication of preference for stimuli, without indication of *recognition*.”, which has a high outfit statistic for all methods. This item has a response pattern with one large outlying value from a patient with a $\hat{\theta}$ (range 4.31 – 4.62) but with a fail on this item. One noticeable thing about this item is that the outfit statistic is lower for the MAF method than for the other methods. This is due to the $\hat{\theta}$ of one of the patients that has a fail on this item being higher for the other methods (due to non-structural missing values), making this patient a stronger outlier in the PVMI and FIML methods. As we are mainly interested in the difference between methods, we will focus on infit and outfit statistics that differ between methods. Similar to the difficulty parameters, the MAF method increases item misfit, as it replaces missing values with fails, regardless of the patients’ $\hat{\theta}$. This impacts both infit and outfit statistics, depending on where the missing values are located and the $\hat{\theta}$ of the patient (e.g. a patient with a high $\hat{\theta}$ with missing values on easy items will contribute more to outfit than infit and vice versa). For examples, see items 3.3b, 4.7 and 4.8. There are two cases in which FIML handling both structural and non-structural missing values causes item misfit for the outfit statistic, namely item 5.9. “Uses visual communication (responds to the other person’s mimics and gestures.” and item 6.4. “Displays joint attention. Makes eye contact, points at an object or brings an object to show it.”. This is caused by the fact that there are only few responses on these items and that consequently a low amount of observations caused all of the misfit. For item 6.4. only few responses were used in calculation of the item fit statistic. This resulted in a single outlying case that was responsible for the underfit of the item.

Overfit was present in all methods where non-structural missing values were replaced with fails (MAF, FIML-MAF, PVMI-MAF) for items with a high difficulty parameter. The FIML and PVMI methods did not have this (extreme) overfit, as they either follow the available data (FIML) or estimate responses in accordance with the model (PVMI). We have seen however that certain items, for example item 6.x. (“Can orient himself well in familiar surroundings.”), had a lower difficulty parameter than was expected from a theoretical point of view. This has an impact on the infit and outfit statistics.

Table 4. Item infit and outfit statistics of the VAS items with different methods for handling missing data.

Item	$Infit_{MAF}$	$Outfit_{MAF}$	$Infit_{FIML:-MAF}$*	$Outfit_{FIML-MAF}$*	$Infit_{FIML}$*	$Outfit_{FIML}$*	$Infit_{PVMI-MAF}$	$Outfit_{PVMI-MAF}$	$Infit_{PVMI}$	$Outfit_{PVMI}$
1	0.78	0.08	0.81	0.11	0.69	0.08	0.76	0.11	0.76	0.10
2.1	0.78	0.08	0.81	0.11	0.69	0.08	0.76	0.11	0.76	0.10
2.2	0.82	0.11	0.85	0.13	0.73	0.11	0.79	0.12	0.78	0.12
2.3	0.78	0.08	0.81	0.11	0.69	0.08	0.76	0.11	0.76	0.10
2.4	0.59	0.11	0.63	0.13	0.70	0.15	0.63	0.14	0.62	0.14
3.1	0.47	0.24	0.54	0.25	0.63	0.30	0.61	0.28	0.69	0.40
3.2	0.43	0.27	0.36	0.13	0.53	**3.12**	0.39	0.16	0.63	**3.64**
3.3a	0.50	0.22	0.48	0.19	0.44	0.19	0.45	0.15	0.51	0.25
3.3b	**1.36**	**1.43**	0.79	0.61	0.86	0.82	0.80	0.51	0.86	0.49
3.4	0.84	0.48	0.94	0.72	0.96	0.49	0.95	0.63	0.93	0.57
3.5	1.24	**6.88**	1.09	**11.29**	1.08	**10.01**	1.11	**11.70**	1.11	**11.23**
3.6	0.55	**4.43**	0.59	**4.45**	0.56	**4.42**	0.63	**4.46**	0.63	**4.47**
3.7	0.70	0.68	0.44	0.15	0.49	0.19	0.52	0.22	0.64	0.54
3.8	0.83	0.72	0.59	0.20	0.74	0.29	0.56	0.22	0.60	0.23
4.1	0.79	0.43	0.89	0.48	0.81	0.43	0.97	0.52	0.95	0.60
4.2	0.73	**4.67**	0.82	**4.68**	0.95	**5.16**	0.86	**4.72**	0.92	**4.84**
4.3	1.26	0.82	1.13	1.01	1.07	0.67	1.02	0.78	1.08	0.77
4.4	0.97	**3.98**	0.89	**4.77**	0.96	**4.74**	0.91	**4.77**	0.99	**4.99**
4.5	0.78	0.47	0.95	0.67	0.78	1.01	0.90	0.57	0.99	0.68
4.6	0.60	0.91	0.58	0.30	0.70	0.40	0.65	0.34	0.67	0.53
4.7	**1.35**	1.14	1.11	0.77	0.97	0.61	1.07	1.00	1.14	1.12
4.8	0.80	1.44	0.71	0.52	0.78	0.91	0.79	0.50	0.75	0.57
4.9	0.83	0.98	0.82	0.46	0.83	0.48	0.84	0.51	0.80	0.44
4.10	0.81	0.48	0.74	0.38	0.75	0.44	0.79	0.44	0.78	0.42
4.11	0.90	0.86	0.97	0.62	0.82	0.48	0.88	0.50	0.88	0.49
5.1	0.67	0.35	0.74	0.35	0.62	0.32	0.70	0.33	0.76	0.35
5.2	0.75	0.37	0.85	0.41	0.88	0.68	0.96	0.50	0.92	0.47
5.3	0.70	0.30	0.83	0.41	0.78	0.37	0.85	0.39	0.78	0.36
5.4	0.52	0.26	0.51	0.23	0.61	0.33	0.64	0.37	0.66	0.81
5.5	0.92	0.31	0.77	0.30	0.99	0.83	0.85	0.29	0.87	1.17
5.6	**1.55**	**1.75**	**1.30**	**1.04**	**1.24**	**1.34**	**1.29**	**1.23**	**1.32**	**1.42**
5.x	0.80	0.57	0.73	0.38	0.85	0.81	0.73	0.39	0.98	1.12
5.7	0.72	0.41	0.84	0.43	0.82	0.66	0.87	0.46	0.91	1.03
5.8	0.76	0.42	0.79	0.42	0.84	0.51	0.78	0.42	0.81	0.68
5.9	0.94	0.50	0.92	0.47	1.04	**1.35**	0.89	0.46	0.94	0.57
5.x	0.72	0.34	0.73	0.37	0.77	0.32	0.79	0.35	0.80	0.42
5.10	0.66	0.20	0.60	0.16	0.63	0.17	0.62	0.15	0.53	0.13
6.1	0.83	0.09	0.83	0.14	0.85	0.62	0.70	0.08	0.84	0.18
6.2	1.00	0.19	1.00	0.26	1.19	1.00	1.12	0.28	1.01	0.39
6.4	0.50	0.11	0.72	0.18	0.84	**1.67**	0.63	0.14	0.73	0.48
6.5	0.68	0.13	0.72	0.14	0.91	0.47	0.70	0.12	0.62	0.54
6.6	0.64	0.13	0.77	0.18	0.92	0.32	0.72	0.16	0.95	0.59
6.x	0.47	0.14	0.66	0.20	0.82	0.80	0.55	0.15	0.76	0.87

Note: $_{MAF}$, missing data is scored as fail; $_{FIML-MAF}$, structural missing values are scored as fail and non-structural missing values are handled by FIML; $_{FIML,}$ missing data is handled by FIML; $_{PVMI-MAF}$, structural missing values are scored as fail and non-structural missing values are handled by PVMI; $_{PVMI,}$ missing data is handled by PVMI;* indicates use of multiple imputation (n = 5) to calculate the infit and outfit statistics. Underfit is in bold.

5.4. Person Fit

Before comparing the person fit statistic, it is useful to look at the distribution of $\hat{\theta}$ under different models. The density of $\hat{\theta}$ distribution is plotted for each method in Figure 6. $\hat{\theta}$ of all models can be seen in Appendix C.

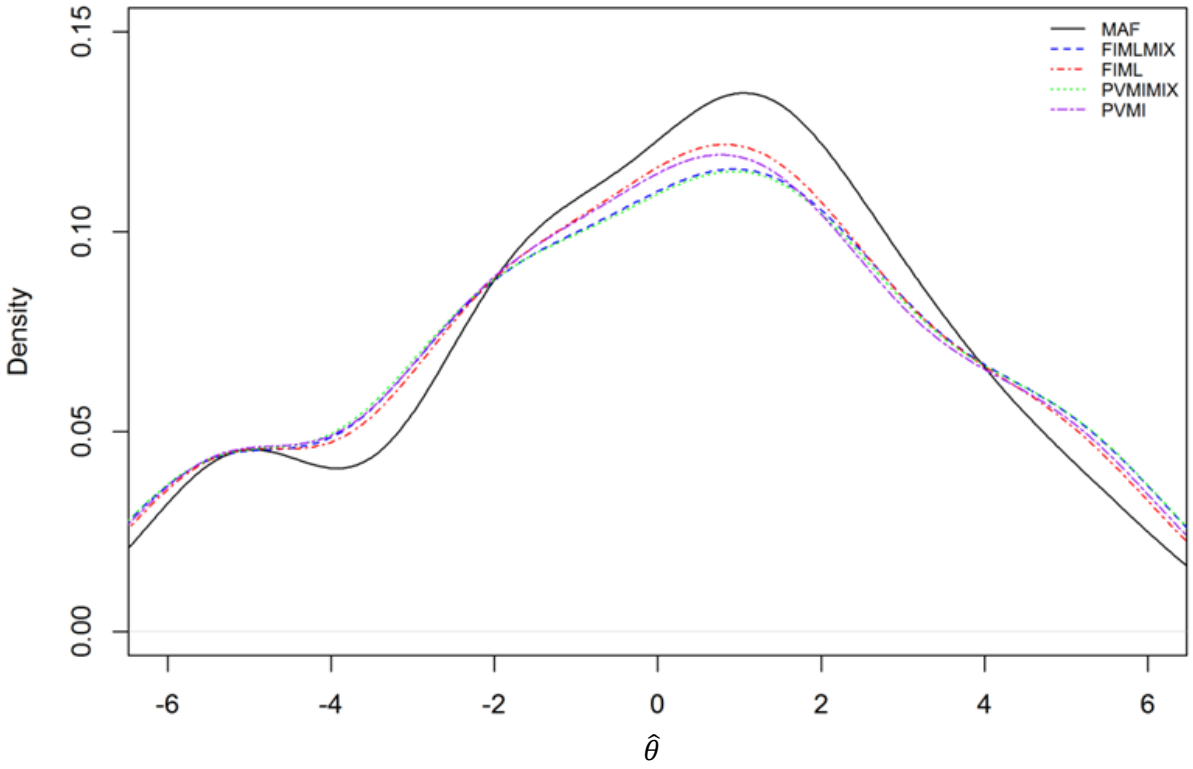

Figure 6. Density of theta estimates distribution per method.

One noticeable thing about the $\hat{\theta}$ is that the EAP estimator attempts to standardize the $\hat{\theta}$ in such a way that the $\hat{\theta}$ of all patients in the sample follow a normal distribution with a mean of 0 and a SD of 1. This results in a cropped range of theta values. This is especially noticeable in the blind patient, that still has a $\hat{\theta}$ of -5.8, while he/she should have an estimate closer to the lowest difficulty parameter of ~-7.6.

$\hat{\theta}$ values were similar across all methods for handling missing data. When using MAF as method of handling missing data, high $\hat{\theta}$'s were, on average, lower than for other methods. This is because at higher $\hat{\theta}$ missing values on items with lower difficulties (the non-structural missing values) are expected to be answered correctly, which MAF does not account for. This results in lower theta values for high $\hat{\theta}$ with missing values in lower difficulty items. This does not apply for patients that

initially had low $\hat{\theta}$. The probability that patients with a low $\hat{\theta}$ passed these items were lower, meaning that replacing the missing value with a fail was less influential on the final $\hat{\theta}$ of these patients than for patients with a high $\hat{\theta}$.

Person fit statistics can also be seen in Appendix C. For all methods for handling missing data it can be seen that only three respondents have consistent person misfit (range l_z = -1.84 - -7.35). These three patients all have high $\hat{\theta}$ and missing values or fails in items with a lower difficulty. However, these person-fit statistics do not tell us anything about the influence of the methods for handling missing values on the person fit indices. To assess this influence we have to look at the patients that only show misfit under some, but not all, methods. Some patients only show person misfit when the missing data are handled using the MAF-method. These patients also have lower $\hat{\theta}$ when the MAF-method is used. The explanation of person misfit is simple in this case, as missing values in lower difficulty items are replaced with fails, while the expected response is a pass. A single patient only had person misfit for the PVMI methods. This patient had no missing data and a $\hat{\theta}$ with range of 1.00 – 1.40 across all methods for handling missing data. This patient also has low l_z-statistics for the other methods (range l_z = -1.47 – -1.64). The imputation of the data of other patients with (structural or non-structural) missing values modifies the difficulty parameters such that this patient no longer fits the model when missing values are handled using PVMI. This patient has one of the most varying response patterns in the higher end of the scale (e.g. cluster 4, 5 and 6). In these clusters there is only little data available, due to structural missing values. If these structural missing values are imputed, more information is available at the high end of the scale and the l_z-statistic of this patient decreases. Finally, for one patient person misfit only arises when structural missing values are handled using FIML or PVMI. This patient has trouble with focus-related items, which influences the l_z- statistic as the Rasch model does not differentiate between items as it measures a unidimensional scale of *visual ability*. The response pattern of this patient contains many fails on items that have a lower difficulty parameter when all structural values are handled by PVMI/FIML, which contributes to a stronger misfit.

Generally, methods that add more information to the Rasch model, such as MAF, PVMI and PVMI-MAF can increase or decrease person misfit. It depends on whether the response pattern of the patient adheres to the method the missing data are handled. If the response pattern of a patient contradicts the model, person misfit increases. The model is (partially) defined by the method of handling the non-structural and structural missing values and thus influences the person fit statistics.

6. Results from a practical perspective

Apart from studying the effects of different methods for handling missing data, another goal of this thesis was to assess the psychometric properties of the VAS and the way it can be implemented as observational instrument in clinical practice. Unlike the previous ordinal scale of visual ability (1-6) the Rasch model allowed us to estimate theta values on a continuous scale. We can use these theta values to discriminate the visual ability of different patients more precisely. However, we can still assign an ordinal level to each theta value using the difficulty parameters. Note that a difficulty parameter in a Rasch model represents the theta value at which the probability of a correct answer is fifty percent. People with a theta value between -4 and -2 would be assigned a visual ability of -3, as items within the theoretical visual ability cluster have a difficulty parameter between these two values. To create a continuous scale that is easier to interpret we use a linear transformation of theta scores to get *T*-scores;

$$T\text{-score} = \theta * 10 + 50.$$

T-scores can be used to define clinical cut-off points, if the assumption of a normal distribution in the population holds If the population *T*-scores are normally distributed, 68% of all patients have a *T*-score between 40 and 60 and 96% have a *T*-score between 30 and 70.

Based on the results from all analyses the VAS could benefit from restructuring. First some specific items were either too difficult or too easy when comparing the difficulty with the assigned level of visual ability of the item based on theory. The item “Shows fixated visual functioning *during daylight*, especially with strong visual stimuli” requires more visual ability than any other item within the same theoretical cluster. This is, however, one of the items that was reformulated because it had a follow-up item that did not adhere to the requirements of the Rasch model. The reformulating procedure might have changed the initial meaning of the item, but revision of the item is recommended. The item “Recognizes familiars/family members visually (without voice).” requires less visual ability than expected by the theoretical cluster of items. This could be due to the ambiguous formulation of the item. Recognition is not defined properly; does the child need to wave, name the

person or perform a specific action? Visually (without voice) is also open to interpretation. For example, the child could recognize the stature or clothing of a family member. It could also be that the child has a reaction to anyone entering the room, which does not necessarily indicate recognition. These results are based on retrospective observation using documentations and personal experience. This may also have an influence on the validity of the results. Using the VAS as an observatory instrument might lead to different results than the results based on retrospective information. Based on item content and item difficulty parameters, several items seem to represent an ordinal scale. For example, the items "Limited visual tracking", ($\beta = -3.8$) "Able to visually track" ($\beta = 0.3$) and "Fixate, tracking and moving gaze well-developed, possible start of scanning." ($\beta = 5.0$) could be turned into a single ordinal item with several levels of visual tracking. Another set of items that could be used as an ordinal scale instead of binary items were the items; "Viewing distance to about arm length" and ($\beta = -4.0$), "Viewing distance to about 1 meter, walking persons are tracked up to 2-3 meter" ($\beta = -0.3$), "Viewing distance enhanced to a minimum of a few meters, provided that visual acuity allows this." ($\beta = 3.1$) and "Sees an object in the distance that is being pointed out" ($\beta = 6.6$). These items all measure viewing distance and could thus be reformulated into one ordinal item related to viewing distance.

Because the VAS is an observational instrument judged by raters it is important that the items have the same meaning for different raters. Based on the results of the inter-rater reliability the overall VAS level (on a scale of one to six) had a moderate agreement between raters ($\kappa = .658$, $p < .001$). The overall amount of CVI criteria had a slight agreement ($\kappa = .196$, $p < .001$). However, the individual item κ (which can be seen in Table 5), varied from no agreement to near perfect agreement.

The low inter-rater agreement could be caused by several problems within this study. First, the VAS was not used as an observational instrument, but filled out retrospectively by both raters, using documentation and/or recalling the patient. Some patients were more involved with one of the two raters than the other. This means that one of the raters would have more information about the patient than the other, resulting in differences between responses of the two raters. It could also be that the item descriptions are too ambiguous and thus differently interpreted by the raters. Finally, it could be

that the items have not been properly defined to be measurable or contain too vague criteria for a fail/pass to be consistent.

Table 5. Cohen's κ of the individual items of the CVI Criteria (n = 73).

Item	κ	P-value
1. "No visual curiosity."	.435	.004
2. "Looking away when reaching or handling."	.570	< .001
3. "Cursory looks and short visual behavior."	.843	< .001
4. "Varying visual behavior."	1.00	-
5. "Cannot use vision simultaneously to other senses, like hearing or touching."	**.137**	**.476**
6. "Looking is tiring."	.284	.046
7. "Familiarity gives better visual behavior and/or recognition."	**.245**	**.099**
8. "Prefers listening above looking."	.402	.014
9. "Staring into lightsources."	.426	.005

Note: Missing data has been pairwise deleted. Bold indicates non-significant agreement.

7. Discussion

This exploratory study focused on investigating the effects of handling missing data with different methods on important psychometric properties of the VAS. The psychometric quality of the VAS was assessed by fitting a Rasch model to the data. The model fit, difficulty parameters, item fit and person fit statistics were compared across five methods for handling missing data. Additionally, the inter-rated reliability of the VAS was assessed to inspect the effectiveness of the VAS in clinical practice. Suggestions were made to improve the psychometric qualities of the VAS.

In terms of difficulty parameter cluster misfit, item misfit, person misfit and model fit our results showed that treating all missing values as fail performed poorest. A possible cause is that treating all missing values as fails creates inconsistent response patterns for anyone with any missing value on items that are easy for their estimated ability level. The differences between FIML and PVMI were small. This is likely due to PVMI using parameter $\hat{\theta}$ provided by the FIML method for the imputation process. The covariate included in the PVMI method does affect the plausible theta value draws, but does not alter the difficulty parameters initially estimated by FIML. After imputation the PVMI method provides new, complete datasets that will provide different parameters and fit indices, but these are heavily influenced by the initial FIML estimation of the Rasch model. There were differences, however, between FIML/PVMI and FIML-MAF/PVMI-MAF methods.

In general, treating non-structural missing values as fails resulted in more difficulty parameter cluster misfit, item misfit, person misfit and worse model fit than any of the other methods. This became especially apparent for items with a large amount of non-structural missing values. A good example of this is item 3.3b “Has visual attention mainly by auditory stimuli *during daylight*”. This item serves as an example that non-structural missing values should not be replaced with fails, because it creates an unrealistic difficulty parameter and causes item misfit.

Treating structural missing values (i.e. items that were not scored by the observer, because they were considered too hard for the patient) as fails resulted in overall decrease of difficulty parameters cluster misfit, item misfit and person misfit. The FIML-MAF and PVMI-MAF methods provided more consistent response patterns by replacing the structural missing values as fails, resulting in less item and person misfit. The FIML-MAF and PVMI-MAF methods that treated structural missing values as fails enforces the theoretical background of the VAS that items in a higher level of visual functioning require a higher ability of visual functioning and cannot be passed by people that have a lower ability of visual functioning. The model-based method FIML displayed more difficulty cluster misfit in higher levels of visual functioning, where few response patterns were available. Due to the small sample size the response patterns at higher levels of visual functioning were unstable. This might have influenced the FIML difficulty parameter estimations to be lower than expected by the

theoretical background of the VAS. For PVMI this was not the case. A possible explanation for this is that the included covariate resulted in different possible value draws and in more consistent response patterns.

The quality of item, person, and model fit could be compared between methods for handling missing data, because they contain cut-off scores that define better or worse fit. This way we could quantify the amount of misfit and compare methods for handling missing data. However, difficulty parameters could not be so easily compared. This thesis was an exploratory study using real data, for which, unlike simulated data, the true parameters were not available for comparison. Instead, we attempted to address not knowing the true difficulty parameters of items by following the theoretical clustering of items created by the developmental team of the VAS. Each item that is at a higher level of visual functioning should, theoretically, require a higher latent trait ability. The method of rating the VAS makes the a priori assumption that items in a higher level of visual functioning require a higher visual ability. For some items such as "Viewing distance to about arm length" and "Viewing distance to about 1 meter", it is logical that this assumption holds. If a patient has a viewing distance of around one meter, he/she a has a viewing distance of less than one meter as well. For some items, however, it is less clear to which clusters they belong. We found several items where all methods displayed a theoretical cluster misfit, indicating that the item might not require as much visual ability as was initially expected. This can be caused by certain characteristics of the patients (e.g. it could be caused by the sample having an overall high visual ability), or due to the item not tapping into the same latent trait (e.g. due to multidimensionality) or simply due to the item not being placed in the correct level of visual functioning by the developers of the instrument. In addition to the item, person and model fit, we also compared this difficulty parameter cluster misfit between methods for handling missing data.

Assessing the psychometric quality of the VAS has led to new insights which can be used to improve the quality of the VAS. By inspecting the item difficulty parameters and dividing the parameters into theoretical clusters, we found that several items did not fit the cluster they were theoretically assigned to. After discussing the results with the developers of the VAS, the items that displayed cluster misfit, were items that they had already considered (re)moving.

A major limitation of this study is that the data were collected retrospectively and the instrument was not used the way it would be used in clinical practice. The idea behind the VAS is that it is used as a checklist. Each item has to be filled out by observing the patient's behavior. For this study the observations were based on previous documentation and/or recollecting consultation with patients. This limitation was especially noticeable for Cohen's Kappa. The first rater (the patients' optician) answered with her own experiences and documentation, whereas the second rater (a colleague) responded with only documented information. A second limitation is that for the data collection phase of this study, raters were requested to continue filling in items for one additional level

of visual functioning above the assigned level of visual functioning of the patient. The problem here is that it could enforce cluster forming of item difficulties for the items in higher levels of visual functioning. These items only contain responses of patients with similar $\hat{\theta}$ and no responses of patients with low $\hat{\theta}$ (unless these values are treated as incorrect). If these structural missing values in high levels of visual functioning are instead handled by FIML (where they are ignored) or by PVMI (where they are imputed in accordance to the available response patterns) this can result in lower difficulty parameters than when these values are replaced with what they represent, namely fails. This is especially the case for items with few response patterns at high levels of visual functioning, and with a relatively high success rate of observed responses. Raters were requested to continue filling in items for one additional level of visual functioning above the assigned level of visual functioning of the patient. A possible issue with this is that we artificially created clusters of item difficulty parameters for items in higher levels of visual functioning. These items only contain responses of patients with similar $\hat{\theta}$. If these structural missing values in high levels of visual functioning are then handled by FIML (where they are ignored) or by PVMI (where they are imputed in accordance with the available response patterns) this can result in lower difficulty parameters. This is particularly the case for items with few response patterns at high levels of visual functioning, and with a relatively high success rate.

The limitation of the data collection is amplified by another limitation, namely the small sample size. A Rasch model can be reliably performed on small sample sizes of fifty respondents, if there are enough items without misfit (more than 30) to provide reliable person measurements (Linacre, 1994). For item estimation a rule of thumb for minimum requirements are eight passes and eight fails to provide a stable (within one logit) estimation of item difficulties (Linacre, 1994). Due to skewness of the sample some items in the VAS do not meet the minimum requirements for a stable item difficulty calibration. Although these are the recommended minimum requirements, the estimated parameters become more robust and precise with larger sample sizes (Chen et al., 2014, Khan, 2014). Additional respondents can have a strong influence on the parameter estimations and patients with high person misfit have a bigger impact on estimated parameters when the sample size is small. All known cases of CVI are treated at the Koninklijke Visio, so a larger sample would not have been possible for the Dutch CVI population. The FIML method would also perform better if the sample size was larger and more response patterns were available. As the overlap between items increases, parameter estimation of FIML will become more stable.

If an instrument or questionnaire is developed with a specific theoretical framework, thought has to be put into the method to handle missing data beforehand. In case of the VAS, difficulty parameters information was available a priori, because items were intended to increase in difficulty. This can also be said about power tests, where items become increasingly difficult and not each respondent finishes all items. In such tests, every item that has not been completed is considered a fail.

The same applies to the VAS, but only if the a priori assumption of item difficulty hold. Analyzing complete data of patients is advised to confirm the theoretical framework on which the VAS was developed, before deciding on which method of handling missing data to apply. If the a priori assumptions about difficulty parameters do hold, we would suggest that structural missing values should be treated as incorrect. Non-structural missing values can be handled by either FIML or PVMI, as they provide only small differences. In clinical practice it might be more appropriate to use PVMI to get a full profile of the patient. Finally, we would advise additional data collection where the VAS is used as an observational, rather than as a retrospective, instrument. This could both improve the low inter-rater reliability found in this study as provide us with data with fewer non-structural missing values (as items can be purposely observed).

In this thesis we found that the model-based methods, FIML(-MAF) and PVMI(-MAF), outperform MAF for non-structural missing values when assessing the psychometric properties of the VAS. For structural missing values, FIML provided items with lower difficulty parameters than we expected based on the theoretical background of the VAS. This was especially the case for items belonging to a high level of visual functioning, where few response patterns were available. Treating structural missing values as fails (FIML-MAF/PVMI-MAF) resulted in less item difficulty misfit and less person misfit than using the model-based methods FIML and PVMI. However, differences in item difficulty depend strongly on the assumption made a priori about the item difficulties of the VAS. If this assumption holds, we would recommend using a model-based method (such as FIML or PVMI) to handle non-structural missing values and handle structural missing values by replacing them as fail (MAF). The results indicate that a combination of imputation methods (FIML-MAF/PVMI-MAF) outperform MAF, FIML and PVMI when assessing the psychometric properties of the VAS (or other instruments with a similar mechanism) with a Rasch model.

References

Adér, H.J., Mellenbergh, G.J., & Hand, D.J. (2011). *Advising on research methods: A consultant's companion.* Huizen, Johannes van Kessel Publishing.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723

Bock, R. & Mislevy, R. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model.* Mahwah, NJ: Lawrence Erlbaum.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York, NY: Springer-Verlag.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1-29.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485-493.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement, 20*(1), 37-46.

Cronbach, J. L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Drasgow, F., Levin, M. V., & Williams, E. A. (1985). Appropiateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Ferro, M. (2014). Missing data in longitudinal studies: Cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of Epidemiology, 24*(1), 75-77.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models: foundations, recent developments and applications*. New York: Springer-Verlag.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological methods, 14*(3), 275.

Frebel, H. (2006). CVI?! How to define and what terminology to use: Cerebral, cortical or cognitive visual impairment. *British Journal of Visual Impairment, 24*(3), 117-120.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. Thousand Oaks, CA: Sage.

He, W. & Wolfe, E. W. (2012). Treatment of not-administered items on individually administered intelligence tests. *Educational and Psychological Measurement, 72*(5), 808-826.

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling, 53*(3), 380-393.

Khan, M. I. (2014). Recovery and stability of item parameter and model fit across vrying sample sizes and test lengths in Rasch analysis with small sample. *Social Science International, 30*(1), 43.

Khan, R. I., O'Keefe, M., Kenny, D., & Nolan, L. (2007). Changing pattern of childhood blindness. *Irish Medical Journal, 100*(5), 458-461.

Langley, M. B. (1998). *ISAVE: Individualized Systematic Assessment of Visual Efficiency for the developmentally young and individuals with multihandicapping conditions.* Louisville, KY: American Printing House for the Blind.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Linacre, J.M., & Wright, B.D. (1994), Chi-square fit statistics. *Rasch Measurement Transactions, 8*(3), 360.

Linacre, J.M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data (2nd ed.).* New York, NY: Wiley.

Martin-Löf, P. (1973). Statistiska modeller [Statistical models.] Anteckningar från seminarier lasåret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt ändrat nytryck, October 1973. Stockholm: Institutet för Försäkringsmatemetik och Matematisk Statistisk vid Stockholms Universitet.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Nakken, H., & Vlaskamp, C. (2007). A need for a taxonomy for profound intellectual and multiple disabilities. *Journal of Policy and Practice in Intellectual Disabilities, 4*(2), 83-87.

Newcomb, S. (2010). The reliability of the CVI Range: A functional vision assessment for children with cortical visual impairment. *Journal of Visual Impairment & Blindness, 104*, 637-647.

Ortibus, E., Laenen, A., Verhoeven, J., De Cock, P., Casteels, I., Schoolmeesters, B., Buyck, A., & Lagae, L. (2011). Screening for cerebral visual impairment: value of a CVI questionnaire. *Neuropediatrics, 42*(4), 138-147.

Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research, 20*(2), 287-300.

R Core Team (2016). *R: A language and environment for statistical computing*. R. Foundation for Statistical Computing, Vienna, Austria.

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207-230.

Roman-Lantzy, C. (2007). *Cortical visual impairment: An approach to assessment and intervention.* New York, NY: American Foundation for the Blind.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NJ: Wiley.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaludating the fit of Structural Equation Models: Tests of significance and descriptie goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schwarz, G. E. (1978). Estimating the dimension of a model, *Annals of Statistics, 6*(2), 461-464.

Seo, D. G. & Weiss, D. J. (2013). l_z Person-Fit Index to Identify Misfit Students With Achievement Test Data. *Educational and Psychological Measurement, 73*(6), 994-1016.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software, 74*(5), 1-27.

Yang, F. M. & Kao, S.T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry, 26*(3), 171-177.

Weinstein, J. M., Gilmore, R.O., Shaikh, S., Fesi, J., Lauren, T., & Cheung, A. (2010). Local and global motion processing in premature children with cerebral visual impairment (CVI). *Journal of AAPOS, 14*(10). DOI: 10.1016/j.jaapos.2009.12.124.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement,* Chicago, IL: MESA Press.

Wright, B.D. (1995). Scores, reliabilities and assumptions. *Rasch Measurement Transactions, 5*(3), 157-158.

Appendix A – VAS

Observatieschaal CVI-ZEVMB

Cerebrale visuele stoornissen bij kinderen met een zeer ernstige verstandelijke en meervoudige beperking

Schaal voor visuele waarneming tbv. dossieronderzoek

ID nummer kind	
Locatienummer	
Observator	

Niveau 1 -Totale visuele beperking/ Blind	**ja**	**nee**	**geen info**	**anders/ nvt**
• Laat geen enkele visuele reactie zien, ook niet in visuele stimulatieruimte				

Niveau 2- Zeer ernstige visuele beperking /Functioneel blind	**ja**	**nee**	**geen infor-matie**	**anders/ nvt**
• Reageert alleen bij lichtprikkels in verduisterde ruimte met gericht kijken				
• Kan in normaal verlichte ruimtes reageren op sterke visuele prikkels, maar niet met gericht kijken (denk aan verstillen, in de richting kijken van)				
Algemeen kijkgedrag en visuele vaardigheden:				
• Zeer korte fixatie				
• Soms minimale volgbeweging				

Niveau 3 -Ernstige beperking/ Passief visueel aandachtsysteem	**ja**	**nee**	**geen info**	**anders/ nvt**
• Laat gericht kijkgedrag zien *bij daglicht*, vooral bij sterke visuele prikkels				
• Kan visuele prikkel alleen waarnemen als die in blikrichting wordt aangeboden/ zoekt niet actief visuele prikkels op				

Heeft *bij daglicht* vooral visuele aandacht als deze getriggerd wordt door: • bewegende objecten of personen				
• een auditieve prikkel				
• Maakt slechts incidenteel oogcontact				
• Kan voorkeur laten blijken voor bepaalde prikkels, zonder duidelijke indicatie van *herkennen*				
Algemeen kijkgedrag en visuele vaardigheden:				
• Kijkt voornamelijk dichtbij, tot armlengte				
• Kortdurende fixatie				
• Beperkt visueel volgen				
Niveau 4 - Matige beperking/ Basale waarneming	**ja**	**nee**	**geen info**	**anders/ nvt**
• Actief visueel aandachtsysteem, zoekt actief (interessante) visuele prikkels op				
• Kijkt met interesse naar voorwerpen uit dagelijks leven, zoals speelgoed, nog geen of nauwelijks aandacht voor details				
• Volgt speeltje wat op de grond valt (objectpermanentie)				
Herkenning:				
• Herkent 5 tot 10 dagelijkse voorwerpen, bv. drinkbeker, doekje of lepel en reageert adequaat, zonder auditieve input				
• Herkent gezicht van bekenden zonder auditieve input, (verhogen alertheid, lachen, reiken etc.)				
• Herkent (favoriet) speelgoed, zonder auditieve input en reageert adequaat				
• Komt min of meer toevallig in bepaalde (hoeken van de) ruimtes en herkent deze (basale ruimtelijke oriëntatie)				
Algemeen kijkgedrag en visuele vaardigheden:				
• Kijkafstand tot ong. 1 meter, lopende personen worden tot 2-3 meter gevolgd				
• Kan visueel volgen				
• Blik verplaatsen is mogelijk				
• Maakt regelmatig oogcontact				

Niveau 5 - Lichte beperking/Uitgebreide visuele herkenning	**ja**	**nee**	**geen info**	**anders/ nvt**
• Visueel alert: houdt actief omgeving in de gaten				
• Enige aandacht voor details/ziet bijvoorbeeld hagelslag op tafel				
• Zoekt oogcontact op grotere afstand				
Visuele herkenning /selectieve aandacht				
• Herkent meer dan 10 voorwerpen				
• Herkent voorwerpen en familie/bekenden op niet te drukke foto's				
• Zoekt gericht voorwerp uit tussen beperkte hoeveelheid voorwerpen				
• Herkent bekenden/ familieleden visueel (zonder stem)				
• Kan zich oriënteren in bekende omgeving				
Algemeen kijkgedrag en visuele vaardigheden				
• Kijkafstand vergroot tot minimaal enkele meters, mits gezichtsscherpte dit toe laat				
• Gebruikt visus bij communicatie (reageert op mimiek van de ander en op gebaren)				
• Kijkt actief ruimte rond, probeert overzicht te krijgen				
• Fixeren, volgen en blik verplaatsen goed ontwikkeld, mogelijk start van scannen				

Niveau 6 -Geen beperking /Normaal visueel functioneren (voor ontwikkelingsleeftijd van 24 maanden)	**ja**	**nee**	**geen infor-matie**	**anders/ nvt**
Visuele herkenning/selectieve aandacht				
• Gaat op zoek naar favoriete speeltjes die niet zichtbaar zijn (geeft blijk van visueel geheugen)				
• Ziet in de verte een voorwerp dat wordt aangewezen				
• Begrip deel/geheel relaties (herkent bv. fiets alleen aan het stuur)				
• Geeft blijk van joint attention. Maakt oogcontact en laat de ander een speeltje zien. Wijst iets aan of brengt iets om het te laten zien.				
• Imiteert gedrag (bv zwaaien, glimlachen, neusoptrekken)				

• Begrijpt voorwerpen/personen/handelingen op pictogrammen (PECs/PCS)				
• Kan zich in bekende omgeving goed oriënteren				
• Interesse in details (o.a bij rijk geïllustreerde plaatjes). Kan daarin vlot iets opzoeken.				

Marjolein Wallroth, GZ-psycholoog VVB-afdeling Amsterdam

Marieke Steendam, Ergotherapeut VVB-team Leiden

Appendix B – VAS Item difficulty parameters across different methods for handling missing data.

Item		β_{MAF}	$\beta_{FIML\text{-}MAF}$	β_{FIML}	$\beta_{PVMI\text{-}MAF}$	Var	β_{PVMI}	Var
1: "Laat een visuele reactie zien (gehercodeerd)."	**1**	-7,44	-7,69	-7,61	-7,70	1,14	-7,65	1,14
2.1: "Reageert bij lichtprikkels in verduisterde ruimtes met gericht kijken."	**2**	-7,44	-7,69	-7,61	-7,70	1,14	-7,65	1,14
2.2: "Kan in normaal verlichte ruimtes reageren op sterke prikkels (niet gericht)"	**3**	-6,61	-6,87	-6,79	-6,89	0,65	-6,84	0,65
2.3: "Kan kort fixeren (minimaal zeer korte fixatie)."	**4**	-7,44	-7,69	-7,61	-7,70	1,14	-7,65	1,14
2.4: "Soms minimale volgbeweging"	**5**	-4,90	-5,22	-5,13	-5,24	0,35	-5,19	0,36
3.1: "Laat gericht kijkgedrag zien *bij daglicht*, vooral bij sterke prikkels"	**6**	**-1,02**	**-1,26**	**-1,28**	**-1,28**	0,21	**-1,37**	0,22
3.2: "Kan visuele prikkel waarnemen als deze in blikrichting wordt aangeboden"	**7**	-3,71	-4,32	-4,01	-4,34	0,32	-4,08	0,32
3.3a: "Heeft *bij daglicht* vooral visuele aandacht voor bewegende objecten/pers."	**8**	-3,46	-4,06	-4,02	-4,10	0,30	-4,08	0,32
3.3b: "Heeft *bij daglicht* vooral visuele aandacht door auditieve prikkel"	**9**	-1,90	-3,50	-3,45	-3,54	0,31	-3,42	0,32
3.4: "Maakt slechts incidenteel oogcontact."	**10**	-2,23	-2,65	-2,59	-2,70	0,26	-2,65	0,26
3.5: "Kan voorkeur laten blijken voor prikkels, zonder indicatie van *herkennen*."	**11**	-2,60	-3,46	-3,41	-3,39	0,30	-3,42	0,31
3.6: "Kan voornamelijk dichtbij, tot armlengte kijken."	**12**	-3,46	-4,01	-3,97	-4,00	0,32	-4,04	0,33
3.7: "Kortdurende fixatie."	**13**	-3,00	-3,77	-3,77	-3,80	0,30	-3,86	0,35
3.8: "Beperkt visueel volgen."	**14**	-2,79	-3,57	-3,55	-3,61	0,28	-3,62	0,30
4.1: "Actief visueel aandachtsysteem, zoekt actief (interessante) visuele prikkels op."	**15**	-0,88	-1,11	-1,10	-1,14	0,21	-1,16	0,21
4.2: "Kijkt met interesse naar voorwerpen uit dagelijks leven, weinig/geen aandacht voor details."	**16**	-0,25	-0,53	-0,53	-0,57	0,20	-0,57	0,20
4.3: "Volgt speeltje wat op de grond valt (objectpermanentie)."	**17**	**2,27**	0,88	0,84	0,82	0,20	0,85	0,20
4.4: "Herkent 1 tot 10 dagelijkse voorwerpen en reageert adequaat, zonder auditieve input."	**18**	-0,12	-0,83	-0,84	-0,88	0,22	-0,91	0,23
4.5: "Herkent gezicht van bekenden zonder auditieve input."	**19**	0,47	0,31	0,28	0,31	0,20	0,26	0,21
4.6: "Herkent (favoriet) speelgoed, zonder auditieve input en reageert adequaat."	**20**	-0,37	-0,71	-0,71	-0,74	0,21	-0,75	0,21
4.7: "Komt min of meer toevallig in bepaalde ruimtes en herkent deze (basale ruimtelijke oriëntatie)."	**21**	0,94	-0,28	-0,29	-0,28	0,21	-0,36	0,25
4.8: "Kijkafstand tot ong. 1 meter, lopende personen worden tot 2-3 meter gevolgd."	**22**	0,00	-0,27	-0,28	-0,30	0,20	-0,32	0,20
4.9: "Kan visueel volgen."	**23**	0,59	0,34	0,32	0,34	0,20	0,29	0,19
4.10: "Blik verplaatsen is mogelijk."	**24**	0,24	-0,09	-0,10	-0,11	0,21	-0,12	0,20
4.11: "Maakt regelmatig oogcontact."	**25**	0,94	0,77	0,74	0,76	0,19	0,71	0,19
5.1: "Visueel alert: houdt actief omgeving/mensen in de gaten."	**26**	2,14	2,25	2,08	2,24	0,21	2,12	0,22

5.2: "Enige aandacht voor details/ziet bijvoorbeeld hagelslag op tafel."	**27**	3,81	3,91	3,77	3,96	0,28	3,83	0,27
5.3: "Zoekt oogcontact op grotere afstand ."	**28**	3,45	3,56	3,41	3,57	0,26	3,42	0,25
5.4: "Herkent meer dan 10 voorwerpen."	**29**	2,82	2,92	2,76	2,87	0,23	2,79	0,24
5.5: "Herkent voorwerpen en familie/bekenden op niet te drukke foto's."	**30**	4,21	3,96	3,79	3,90	0,27	3,89	0,31
5.6: "Zoekt gericht voorwerp uit tussen beperkte hoeveelheid voorwerpen."	**31**	4,00	3,18	3,04	3,20	0,28	3,01	0,25
5.x: "Herkent bekenden/familieleden visueel (zonder stem)	**32**	1,17	**0,93**	**0,61**	**0,91**	0,20	**0,57**	0,22
5.7: "Kan zich oriënteren in bekende omgeving."	**33**	3,28	2,92	2,74	2,94	0,30	2,76	0,24
5.8: "Kijkafstand vergroot tot minimaal enkele meters, mits gezichtsscherpte dit toe laat."	**34**	2,97	3,16	3,00	3,18	0,23	3,07	0,24
5.9: "Gebruikt visus bij communicatie (reageert op mimiek van de ander en op gebaren)."	**35**	3,28	3,42	3,27	3,40	0,24	3,26	0,24
5.x: "Kijkt actief ruimte rond, probeert overzicht te krijgen."	**36**	3,12	3,37	3,20	3,37	0,24	3,26	0,24
5.10: "Fixeren, volgen en blik verplaatsen goed ontwikkeld, mogelijk start van scannen."	**37**	4,67	**5,07**	**4,91**	**5,08**	0,31	4,98	0,32
6.1: "Gaat op zoek naar favoriete speeltjes die niet zichtbaar zijn (geeft blijk van visueel geheugen)."	**38**	7,32	7,52	7,27	7,46	1,14	7,54	1,16
6.2: "Ziet in de verte een voorwerp dat wordt aangewezen."	**39**	6,51	6,64	6,34	6,85	0,65	6,57	0,71
6.4: "Geeft blijk van joint attention. Maakt oogcontact, wijst iets aan of brengt iets om het te laten zien."	**41**	5,23	5,37	4,91	5,44	0,36	5,26	0,38
6.5: "Imiteert gedrag (bv zwaaien, glimlachen, neusoptrekken)."	**42**	5,58	5,98	5,66	5,99	0,40	5,73	0,43
6.6: "Begrijpt voorwerpen/personen/handelingen op pictogrammen (PECs/PCS)."	**43**	5,58	5,60	5,16	5,57	0,43	5,27	0,40
6.x "Kan zich in bekende omgeving goed oriënteren."	**44**	4,43	4,56	**3,86**	4,54	0,30	**3,94**	0,38

Note: β_{MAF}, difficulty parameter of the item when all missing data are scored as fails; $\beta_{FIML\text{-}MAF}$, difficulty parameter of the item when structural missings are scored as fails and non-structural missing values are handled by FIML; $\beta_{FIML,}$ difficulty parameter of the item when missing data is handled by FIML; $\beta_{PVMI\text{-}MAF}$, difficulty parameter when structural missing values are scored as fails and non-structural missings are imputed by multiple plausible values; β_{PVMI}, difficulty parameter of the item when missing data is imputed by multiples plausible values; Var, the variance of multiple imputations.

Appendix C - VAS theta estimates and person-fit statistics

Rank	$\hat{\theta}_{MAF}$	$\hat{\theta}_{FIML\text{-}MAF}$	$\hat{\theta}_{FIML}$	$\hat{\theta}_{PVMI\text{-}MAF}$	$\hat{\theta}_{PVMI}$	l_{zMAF}	$l_{zFIML\text{-}MAF}$	l_{zFIML}	$l_{zPVMI\text{-}MAF}$	l_{zPVMI}
1	-5,82	-5,86	-5,78	-5,86	-5,79	-1,24	-1,36	**-2,21**	-1,11	**-1,83**
2	-5,13	-5,43	-5,35	-5,44	-5,42	0,57	0,75	0,61	0,70	0,59
3	-5,13	-5,43	-5,35	-5,44	-5,39	1,27	1,44	1,31	1,41	1,28
4	-5,13	-5,43	-5,23	-5,44	-5,26	1,27	1,44	1,06	1,42	1,07
5	-5,13	-5,43	-5,35	-5,44	-5,36	1,27	1,44	1,30	1,41	1,28
6	-5,13	-5,43	-5,35	-5,44	-5,42	1,27	1,44	1,30	1,37	1,25
7	-5,13	-5,43	-5,35	-5,44	-5,42	1,27	1,44	1,30	1,42	1,28
8	-4,73	-5,10	-5,02	-5,14	-5,11	1,35	1,36	1,35	1,32	1,29
9	-4,73	-5,16	-5,05	-5,17	-5,11	1,35	1,45	1,28	1,31	1,15
10	-2,63	-3,24	-3,16	-3,24	-3,23	1,34	1,10	1,03	0,45	0,32
11	-2,63	-2,74	-2,66	-2,84	-2,80	0,72	0,96	0,89	0,57	0,45
12	-2,29	-2,63	-2,55	-2,77	-2,73	0,41	0,30	0,20	-0,25	-0,42
13	-2,29	-2,71	-2,63	-2,66	-2,73	1,28	1,29	1,19	0,90	0,74
14	-2,29	-2,89	-2,82	-2,91	-2,90	1,28	1,35	1,27	0,45	0,29
15	-1,95	-2,53	-2,46	-2,55	-2,55	1,58	1,63	1,54	0,75	0,60
16	-1,95	-2,23	-2,20	-2,24	-2,25	0,60	1,03	1,06	0,65	0,58
17	-1,64	-2,18	-2,11	-2,20	-2,18	1,80	1,76	1,63	1,55	1,42
18	-1,64	-2,09	-2,06	-2,14	-2,08	1,80	1,64	1,65	1,44	1,41
19	-1,64	-2,17	-2,09	-2,20	-2,11	0,74	0,95	0,84	0,51	0,37
20	-1,64	-1,91	-1,83	-1,97	-1,82	0,97	1,35	1,21	1,11	0,93
21	-1,64	-2,10	-1,88	-2,14	-1,92	1,16	0,58	0,38	0,31	0,16
22	-1,33	-1,19	-1,08	-1,19	-1,17	-0,48	0,71	0,46	0,08	-0,12
23	-1,33	-1,82	-1,74	-1,83	-1,82	1,73	1,55	1,40	1,21	1,07
24	-1,33	-1,84	-1,77	-1,87	-1,82	1,68	1,55	1,39	0,91	0,73
25	-1,03	-1,22	-1,21	-1,26	-1,26	0,58	1,34	1,30	-0,22	-0,42
26	-0,74	-1,04	-0,90	-1,16	-1,02	0,50	-0,30	-0,57	-0,89	-1,05
27	-0,46	-0,47	-0,49	-0,51	-0,50	-0,66	0,17	0,20	-0,24	-0,27
28	-0,46	-0,62	-0,47	-0,63	-0,70	-0,78	-0,35	-0,71	-0,94	-1,25
29	-0,46	-0,55	-0,33	-0,54	-0,36	1,19	0,87	0,41	0,35	0,04
30	-0,18	-0,55	-0,57	-0,63	-0,64	1,63	1,45	1,44	0,77	0,63
31	0,09	0,16	0,52	0,17	0,47	0,09	0,52	-0,18	0,03	-0,36
32	0,09	-0,32	-0,33	-0,34	-0,36	-0,65	-0,34	-0,34	-0,85	-0,96
33	0,35	0,05	0,03	0,03	-0,06	1,16	1,14	1,06	0,77	0,64
34	0,35	-0,04	-0,08	-0,06	-0,11	0,46	0,48	0,64	-0,21	-0,20
35	0,35	0,05	0,01	0,00	-0,06	1,24	0,88	0,97	0,41	0,39
36	0,35	-0,04	-0,07	-0,06	-0,09	0,20	0,22	0,24	-0,07	-0,14
37	0,61	0,94	0,88	0,91	0,89	0,10	0,40	0,44	-0,10	-0,06
38	0,61	0,80	0,75	0,79	0,66	-1,18	1,11	1,04	0,82	0,69
39	0,88	0,55	0,51	0,54	0,47	0,86	1,07	1,02	0,59	0,47
40	0,88	0,53	0,48	0,51	0,44	1,08	1,42	1,36	1,28	1,13
41	0,88	1,14	1,08	1,17	1,08	-1,64	-0,15	-0,11	-0,41	-0,42
42	0,88	0,65	0,60	0,68	0,61	1,38	1,25	1,19	0,97	0,82
43	1,14	1,36	1,30	1,35	1,23	-1,54	-0,68	-0,59	-1,05	-1,00
44	1,40	1,68	1,62	1,67	1,57	1,02	1,67	1,56	1,53	1,41
45	1,40	1,27	1,21	1,38	1,20	1,86	1,68	1,55	1,48	1,33
46	1,40	1,27	1,16	1,29	1,11	1,11	0,84	0,98	0,63	0,65

47	1,40	1,10	1,00	1,08	1,00	1,45	1,62	1,63	1,14	1,04
48	1,40	2,24	2,19	2,18	2,11	0,17	1,13	0,98	0,85	0,71
49	**1,40**	**1,10**	**1,00**	**1,08**	**1,00**	-1,27	-1,64	-1,47	**-1,96**	**-1,94**
50	1,66	2,09	2,06	1,97	2,05	1,61	1,43	1,28	1,20	1,06
51	1,66	1,79	1,39	1,67	1,45	-0,07	-0,77	**-2,39**	-1,10	**-2,92**
52	1,66	3,03	3,13	3,22	3,16	**-2,63**	-1,27	-1,57	-1,49	**-1,75**
53	1,93	1,75	1,62	1,70	1,57	1,10	0,65	0,76	0,50	0,51
54	2,19	1,98	1,92	1,97	1,88	1,00	0,74	0,64	0,39	0,36
55	2,19	2,04	1,98	2,06	1,97	1,88	1,72	1,59	1,39	1,25
56	2,19	2,57	2,40	2,56	2,45	-0,38	-0,46	-0,28	-0,75	-0,70
57	2,45	2,27	2,22	2,26	2,17	1,69	1,48	1,33	1,19	1,04
58	2,45	2,27	2,12	2,26	2,14	1,00	1,12	1,19	0,70	0,69
59	2,71	2,81	2,63	2,82	2,71	0,41	0,35	0,41	0,14	0,09
60	2,97	2,85	2,67	2,85	2,71	1,41	1,07	1,14	0,60	0,50
61	3,23	4,82	4,66	4,83	4,62	0,10	1,13	1,16	0,89	0,87
62	3,23	4,65	4,44	4,61	4,50	-0,01	0,91	0,92	0,65	0,64
63	3,49	3,75	3,52	3,72	3,56	1,29	1,14	1,16	0,90	0,84
64	3,49	3,43	3,50	3,43	3,56	1,28	1,16	0,73	0,81	0,44
65	3,49	4,46	4,19	4,51	4,47	-0,79	1,11	1,04	1,01	0,87
66	3,76	4,22	3,97	4,29	4,09	0,33	1,16	1,34	0,93	1,04
67	4,31	5,48	5,34	5,50	5,43	**-3,29**	1,24	1,26	1,09	1,05
68	4,31	4,32	4,07	4,32	4,15	1,26	0,95	1,13	0,74	0,82
69	**4,31**	**4,62**	**4,36**	**4,64**	**4,47**	**-6,87**	**-7,20**	**-6,69**	**-7,35**	**-6,94**
70	**5,39**	**5,66**	**5,58**	**5,68**	**5,58**	**-1,84**	**-2,13**	**-1,98**	**-2,27**	**-2,14**
71	5,56	5,60	5,48	5,60	5,53	1,70	1,61	1,72	1,56	1,53
72	5,56	5,73	5,66	5,72	5,69	0,99	1,30	1,48	1,22	1,32
73	**5,56**	**5,70**	**5,61**	**5,72**	**5,67**	**-2,34**	**-2,30**	**-2,12**	**-2,34**	**-2,39**
				Total amount of misfit:		**5**	**3**	**5**	**4**	**7**

Note: $\hat{\theta}$, theta estimate; l_z, person-fit statistics; $_{MAF}$, missing values treated as fails, $_{FIML\text{-}MAF}$, non-structural missing values handled by full information maximum likelihood; $_{FIML}$, all missing values handled by full maximum likelihood; $_{PVMI\text{-}MAF}$, non-structural missing values handled by plausible multiple value imputation; $_{PVMI}$, plausible value multiple imputation. Bold values indicate person misfit.