Communications for Statistical Applications and Methods
2016, Vol. 23, No. 6, 531–541

http://dx.doi.org/10.5351/CSAM.2016.23.6.531
Print ISSN 2287-7843 / Online ISSN 2383-4757

The effects of scanning position on evaluation of cerebral atrophy level: assessed by item response theory

Md Mahsin[1,a], Yinshan Zhao[b]

[a]Department of Mathematics and Statistics, University of Calgary, Canada;
[b]MS/MRI Research Group, Department of Medicine, University of British Columbia, Canada

Abstract

Cerebral atrophy affects the brain and is a common feature of patients with mild cognitive impairment or Alzheimer's diseases. It is evaluated by the radiologist or reader based on patient's history, age and the space between the brain and the skull as indicated by magnetic resonance (MR) images. A total of 70 patients were scanned in the supine and prone positions before three radiologist assessed their atrophy level. This study examined the radiologist's assessment of the cerebral atrophy level using a graded response model of item response theory (IRT). A graded response model (GRM) is fitted to our data and then item-fit and person-fit statistics are evaluated to assess the fitted model. Our analysis found that the cerebral atrophy level is better discriminated by readers in the prone position because all item slopes were greater than 2 at this position, versus the supine position where all the slope parameters were less than 1. However, the thresholds are very similar for the first reader and are quite different for the second and third readers because the scanning position affects readers differently as the category threshold estimates vary considerably between the readers..

Keywords: cerebral atrophy, IRT, GRM, supine and prone, readers

1. Introduction

Cerebral atrophy is an aging process caused by the death of brain tissue. The presence of atrophy is normally associated with other diseases, and is used to support further diagnoses. The evaluation of cerebral atrophy is made by radiologists based on patient history, age and the space between the brain and the skull as indicated by magnetic resonance (MR) images that may be used to clinically support diagnoses such as Alzheimer's disease. Quantitative morphometric tools are alternative methods to assess the degree of cortical atrophy, but are generally more applicable in a research setting than in clinical use due to the time-consuming nature of such tools (Karas *et al.*, 2007).

Patients are usually scanned in the supine position; however, the brain might move downward in this position increasing the anterior space, and subsequently affecting the evaluation. For this reason, the prone position is also considered for an MR scan. A number of published reports have examined the effect of prone versus supine imaging on the movement of the spinal cord, which does demonstrates an anteroposterior movement with the effect of gravity (McCullough *et al.*, 1990; Witkamp *et al.*, 2001). This study determines if the radiologist's evaluation of cerebral atrophy is affected by scanning positions using the graded response model (GRM) of item response theory (IRT). As far as our knowledge, this is the first time application of IRT to the effects of scanning position on evaluation of cerebral atrophy level. However, IRT-based models have become recently popular in health

[1] Corresponding author: Department of Mathematics and Statistics, University of Calgary, 612 Campus Place N.W., Calgary, AB, T2N 1N4 Canada. E-mail: md.mahsin@ucalgary.ca

Published 30 November 2016 / journal homepage: http://csam.or.kr

outcomes, quality-of-life research, and clinical research (Adams *et al.*, 2005; Coleman *et al.*, 2002; Edelen and Reeve, 2007; Hays *et al.*, 2000, 2007; Holman *et al.*, 2003; Reise and Waller, 2009; Reise, 2016).

2. Methodology

2.1. Data

The subjects were patients scheduled for magnetic resonance imaging (MRI) of the head at the University of British Columbia (UBC) Hospital, Canada. Patients who were too weak to be scanned in the prone position were excluded from the study. This suggests that those individuals with severe atrophy were probably excluded because they were more likely to be in poor health. Each scan produced a set of 20 selected axial images. In total, 70 patients were scanned in both supine and prone positions in order to create two sets of images for each patient. Each set of images was randomly assigned a unique identifying number. After all the scans were collected, the images were read by three radiologists, independently and blinded to the position, to decide presence or absence of frontal and temporal atrophy. If atrophy was present, patients were further classified into three mutually exclusive categories (mild, moderate and severe cerebral atrophy) with the level of atrophy rated on a 4-point scale: 0 = no, 1 = mild, 2 = medium, 3 = severe.

2.2. Item response theory

IRT models have become increasingly popular in the past 35 years (Rasch, 1960, 1961; Samejima, 1969, 1972). These models are often used to evaluate a test or survey questionnaire that consists of questions with exhaustive and mutually exclusive response choices. The term *item* is generic in IRT that covers the test/survey questions.

IRT often concerned with the measurement of a hypothetical construct that is latent and can only be measured indirectly through the measurement of observable variables. This hypothetical construct often represents the ability, skill, or a latent person characteristic that the items measure. IRT models describe the association between a subject's underlying level on a latent trait and the probability of a particular item response (Reise *et al.*, 1993). In this study, the latent variable represents the actual atrophy level of patients and the item responses are the atrophy assessments by radiologists in a supine or prone position.

There are two key assumptions of IRT models for the data to which the models are applied: appropriate dimensionality and local independence. The first assumption means that the number of latent traits measured by the items correspond to the number of trait variables assumed in the IRT model. We assume that the construct being measured here (atrophy level) is unidimensional and that the covariance among the readings or assessments can be explained by a single underlying latent variable. This assumption can be evaluated by an exploratory factor analysis; including eigenvalues, scree plots, and the magnitude of item loadings on the first factor (Cattell, 1966, 1978; Loehlin, 2004). The second assumption is local independence which means that the items are uncorrelated when latent trait or traits have been controlled for (McDonald, 1981).

Under local independence, the probability of a vector of item responses, $\mathbf{u}$, for a single individual with trait level θ is the product of the probabilities of the individual responses, u_i, to the items on a

test consisting of I items.

$$P(\mathbf{U} = \mathbf{u}|\theta) = \prod_{i=1}^{I} P(u_i|\theta) = P(u_1|\theta)P(u_2|\theta)\cdots p(u_I|\theta), \tag{2.1}$$

where $P(\mathbf{U} = \mathbf{u}|\theta)$ is the probability that the vector of observed item scores for a person with trait level θ has the pattern $\mathbf{u}$, and $P(u_i|\theta)$ is the probability that a person with trait level θ obtains a score of u_i on item i.

Similarly, the probability of the responses to a single item, i , by n individuals with abilities in the vector $\boldsymbol{\theta}$ is given by

$$P(\mathbf{U_i} = \mathbf{u_i}|\boldsymbol{\theta}) = \prod_{j=1}^{n} P(u_{ij}|\theta_j) = P(u_{i1}|\theta_1)P(u_{i2}|\theta_2)\cdots P(u_{in}|\theta_n), \tag{2.2}$$

where $\mathbf{U_i}$ is the vector of responses to item i for persons with abilities in the $\boldsymbol{\theta}$-vector, u_{ij} is the response on item i by person j, and θ_j is the trait level for person j. Thus, the probability of the full matrix of responses of n individuals to I items is given by

$$P(\mathbf{U} = \mathbf{u}|\boldsymbol{\theta}) = \prod_{j=1}^{n}\prod_{i=1}^{I} P(u_{ij}|\theta_j). \tag{2.3}$$

2.3. Graded response model

One of the most basic aspects of the application of parametric IRT models involves choosing the right model. Several parametric unidimensional IRT models are described in Thissen and Steinberg (1986) to analyze the different types of item responses. The first consideration when choosing the right model involves the number of item response categories. For polytomous items, variations of the Partial Credit Model (Masters, 1982); Rating Scale Model (Andrich, 1978); Generalized Partial Credit Model (Muraki, 1992, 1997) as well as the GRM (Samejima 1969, 1997) are available for ordered responses, and the Nominal Model (Bock, 1972) is appropriate for items with a non-specified response order.

We consider the GRM (Samejima, 1997) because it is appropriate to use when item responses are ordered categorical responses. In this model, the cumulative probabilities of an ordinal response scale are modeled directly as a function of the latent trait variable. The probability of scoring in a specific category is provided by the probability of responding in (or above) this category minus the probability of responding in (or above) the next category. Let K_i denote the number of response categories of item i, then the GRM is given by

$$\begin{aligned} P\left(U_{ij} = k|\theta_j, a_i, \mathbf{b_i}\right) &= P\left(U_{ij} \geq k|\theta_j, \mathbf{b_i}\right) - P\left(U_{ij} \geq k+1|\theta_j, \mathbf{b_i}\right) \\ &= \int_{b_{i,k}}^{\infty} \psi\left(z; a_i, \theta_j\right) dz - \int_{b_{i,k+1}}^{\infty} \psi\left(z; a_i, \theta_j\right) dz \\ &= \Psi\left(a_i(\theta_j - b_{i,k+1})\right) - \Psi\left(a_i(\theta_j - b_{i,k})\right) \\ &= \frac{\exp\left(a_i(\theta_j - b_{i,k+1})\right)}{1 + \exp\left(a_i(\theta_j - b_{i,k+1})\right)} - \frac{\exp\left(a_i(\theta_j - b_{i,k})\right)}{1 + \exp\left(a_i(\theta_j - b_{i,k})\right)}, \end{aligned} \tag{2.4}$$

Table 1: Descriptive statistics, estimated item parameters (standard errors) for the graded response model and item-fit statistics

Reader position	Mean (SD)	a	b_1	b_2	b_3	Zh	$S-X^2$	df	p
Reader 1-S	0.557 (0.651)	0.968 (0.353)	0.138 (0.165)	2.837 (0.997)		0.34	5.11	3	0.16
Reader 1-P	0.514 (0.608)	2.636 (0.841)	0.105 (0.171)	1.953 (0.621)		1.41	4.73	2	0.09
Reader 2-S	0.743 (0.630)	0.917 (0.343)	−0.779 (0.422)	2.716 (1.110)		0.31	8.23	5	0.14
Reader 2-P	0.843 (0.629)	3.686 (1.012)	−0.686 (0.371)	1.457 (0.672)	2.484 (1.115)	1.55	1.13	2	0.57
Reader 3-S	1.500 (0.847)	0.875 (0.330)	−3.273 (1.431)	0.386 (0.250)	2.086 (0.949)	0.41	5.40	5	0.37
Reader 3-P	1.486 (0.959)	2.930 (0.876)	−1.047 (0.442)	−0.097 (0.157)	1.268 (0.492)	2.09	11.80	4	0.02

S = supine; P = prone; df = degrees of freedom.

where ψ and Ψ are the logistic density and logistic cumulative distribution function, respectively. In addition, k is the score on the item, a_i is an item slope or discrimination parameter, and $b_{i,k}$ is a category threshold or difficulty parameter for the k^{th} step of the item. The probability of scoring in or above the lowest category is one and the probability of scoring above the highest category is zero. The slope parameter (a_i) indicates how well an item can discriminate between contiguous trait levels. A large value of a_i indicates a better discriminant ability of the item. The threshold parameter ($b_{i,k}$) represents the trait level necessary to have a score of k or higher with a probability of 0.50. Ideally, the range of threshold parameter values for each item indicates the range of the latent trait that the item measures.

The analyses were conducted using the `mirt` package (Chalmers, 2012) in the R Statistical Computing Environment (R Development Core Team, 2011). In our dataset, there is one missing assessment. This missing value was imputed under the missing at random (MAR) assumption using a built-in function in the package that employs a Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010).

3. Results

3.1. GRM item parameter estimates

Table 1 displays item descriptive statistics, item parameter estimates and standard errors from the GRM as well as item-fit statistics. Item 2 (Reader 1, Prone) has the lowest average atrophy score (0.514) with a SD of 0.608 while item 5 (Reader 3, Supine) has the highest average (1.500) with a SD of 0.847. First three of the six items have three categories with no patient rated with severe atrophy; therefore, only two threshold parameters were estimated for each of these items. The slope parameter (a_i) estimates range from 0.875 to 3.686, indicating considerable variation across items in terms of the discrimination ability. However, the slope parameters range from 0.875 to 0.968 for the supine position whereas for prone position range from 2.636 to 3.686 indicating relative consistency for the same position and considerable difference between the two positions. The threshold parameters ($b_{i,k}$) for the 6 items range from −3.273 to 2.837 and reflects the underlying atrophy level. The smallest threshold estimates corresponding to item 5 (Reader 3, Supine) between 'absence of atrophy' and 'mild atrophy' (i.e., fewest patients were rated as absence of atrophy by this item) whereas the largest threshold estimates related to item 1 (Reader 1, Supine) between 'mild atrophy' and 'medium atrophy' (i.e., fewest patients were rated as medium or severe atrophy by this item).

3.2. Assessment of model fit

The fit of the GRM can be examined through a comparison of model predictions and observed data in various ways. In addition to examining the overall fit of the model to the data, it is also possible

Table 2: Two subjects with large values of the person-fit index Z_L

Subject	Reader 1-S	Reader 1-P	Reader 2-S	Reader 2-P	Reader 3-S	Reader 3-P	Z_L
1	2	1	2	0	3	2	−4.1
2	2	0	2	0	3	0	−4.5

S = supine; P = prone.

to examine the fit for each item. The item goodness-of-fit statistic is one diagnostic tool to evaluate the degree of model-data misfit. The number of item fit indices for examining the appropriateness of GRM model are: Bock's χ^2 (Bock, 1972), Yen's Q statistic (Yen, 1981), standardized appropriateness index, Zh (Drasgow *et al.*, 1985), and Kang and Chen (2007) $S - X^2$ statistic which is a different χ^2 statistics based on raw sum-score conditioning. This report presents the Zh values, and the $S - X^2$ statistics to examine the fit of individual items. These statistics indicated that 1 of the 6 items was not well represented by the estimated GRM item parameters at the significance level of 0.05. The chi-squared statistics were added across items and resulted in a value of 36.4 on 21 degrees of freedom ($p = 0.02$), which also indicated that overall model does not fit well. This is essentially due to one of the readings (Reader 3, Prone) being quite different from others.

It is also possible to examine model-data fit at the individual level with person fit indices that evaluates the consistency of individual response patterns with the proposed model (Embretson and Reise, 2000). Several person fit indexes (Reise, 1990; Reise and Flannery, 1996; Reise *et al.*, 1993) have been developed for this purpose. The standardized fit index (Z_L) is one such index where large negative Z_L values ($Z_L \leq -2.0$) indicate misfit or lack of fit. Large positive Z_L values indicate response patterns that are higher in likelihood than the model predicts because a person responds more consistently than expected under the GRM. By examining the person fit statistics, two subjects have been found to depart substantially from the fitted model (Table 2). In both cases, the atrophy levels evaluated at the supine position are different from those in the prone positions; however, the assessments of the three readers under the same position tend to mutually agree. Meanwhile, the values of Z_L fir quite well and range from −1.3 to 1.4 for the remaining subjects.

3.3. IRT assumptions

3.3.1. Unidimensionality

Unidimensionality can be evaluated by conducting an exploratory factor analysis (EFA). The eigenvalues from the EFA was in favor of a single factor, with the first value substantially larger than others (2.98, 1.38, 0.58). In the single-factor solution, item factor loadings were all positive, ranging from 0.426 to 0.791. This factor accounted for 39% of the total variance for the atrophy level. The promax-rotated two-factor solution extracted the 3 prone items in the first factor and the factor loadings were 0.638, 0.854 and 0.847. All three supine items corresponds to the second factor and the factor loadings were 0.848, 0.623 and 0.799. The cumulative proportion of variance explained by these two factors was 60%. The EFA for atrophy level moderately supported the unidimensionality assumption with some indication of the existence of a second factor. This further suggests that the readings based on the two positions might differ.

3.3.2. Local independence

Local independence means that there should be no association among the item responses if the latent atrophy level (θ) is held constant. Violations of this assumption may result in biased parameter estimates and lead to erroneous decisions when selecting items for scale construction. It has been suggested that local independence is achieved in an EFA model if a single factor accounts for the

correlations between items, so that residual correlations do not differ significantly from zero (Embretson and Reise, 2000). The local independence assumption is not violated because the item-to-item residual correlations from our one-factor model are approximately zero.

3.4. Graphical presentations of the fitted model

Based on a GRM, category response curves can be generated that represent the probability of being rated in a particular category conditional on the subject's latent atrophy level. Collectively, category response curves are referred to as item characteristic curves (Embretson and Reise, 2000). Item information curves can also be computed with the fitted GRM which tells us how precisely the latent trait can be estimated by a particular item. Summarizing the item information curves forms a test information curve (Embretson and Reise, 2000) that represents the range of atrophy level over which the latent trait is most useful for differentiating study subjects.

Figure 1 indicates the item characteristic curves for the six items. In item 1 (Reader 1, Supine), the probability of being rated as absence of atrophy is a decreasing function of θ, the latent atrophy variable, and crosses the 0.5-probability line at the value of 0.138 (the red line). At the value of 2.837, the line corresponding to the 'medium' category (the magenta line) crosses the 0.5-probability line. The item slopes were reflective of the estimates (Table 1); a higher slope estimates results in steeper and sharper item characteristic curves. It is evident that the curves for the supine position are wider compared to those for the prone position. In general, as the a_i-parameter increases, the probability to obtain a particular score in the prone item changes more quickly with a change in the θ-value. For instance, for Reader 1 at the supine position, a subject at $\theta = 2$ is most likely to be rated as 'mild atrophy' (55% probability) and has a relatively lower probability to be rated as 'medium atrophy' (30% probability) or 'absent atrophy' (15% probability). For Reader 1 at the prone position, the same subject has a similar chance to be rated as 'mild' (45% probability) or 'moderate' (50% probability), and is unlikely to be rated as 'absent' (2% probability). Similar patterns are seen for Readers 2 and 3. However, an increase in the b_{ik}-parameters shifts the category response curves to the right. This means that a patient needs a higher level of θ to have a similar probability of receiving a score at the same level.

The expected score of a reading as a function of the latent atrophy scale (θ) is given by

$$E\left(U_{ij}|\theta_j\right) = \sum_{k=0}^{m} kP\left(U_{ij} = k|\theta_j\right).$$

Figure 2(a) presents the expected score curves for all the items. The slope of the curve changes from supine to prone position along the range of θ depending on the distance between the b_{ik}-parameters. This figure suggests that the prone position allow radiologists two better discriminate patients with different levels of atrophy than in the supine position as it has steeper slopes that corresponding to the latent atrophy level.

The item information for the GRM, derived by Samejima (Samejima, 1969), reflects how well the latent trait level can be estimated by an item; subsequently, the information reflects reliability and the precision of the item. Figure 2(b) shows the item information curves for individual items that are useful to identify items that perform well or poorly. The plot shows the clear effect of the increase in the a-parameter and the shift in the b_{ik}-parameters. The prone position items (readings) have higher slopes and reach higher information (i.e., precision or reliability) levels than the supine position items. The performance of the prone items is therefore fairly good, while it is poorer for the supine items. Therefore, the three items corresponding to the supine position are less precise based on the fitted

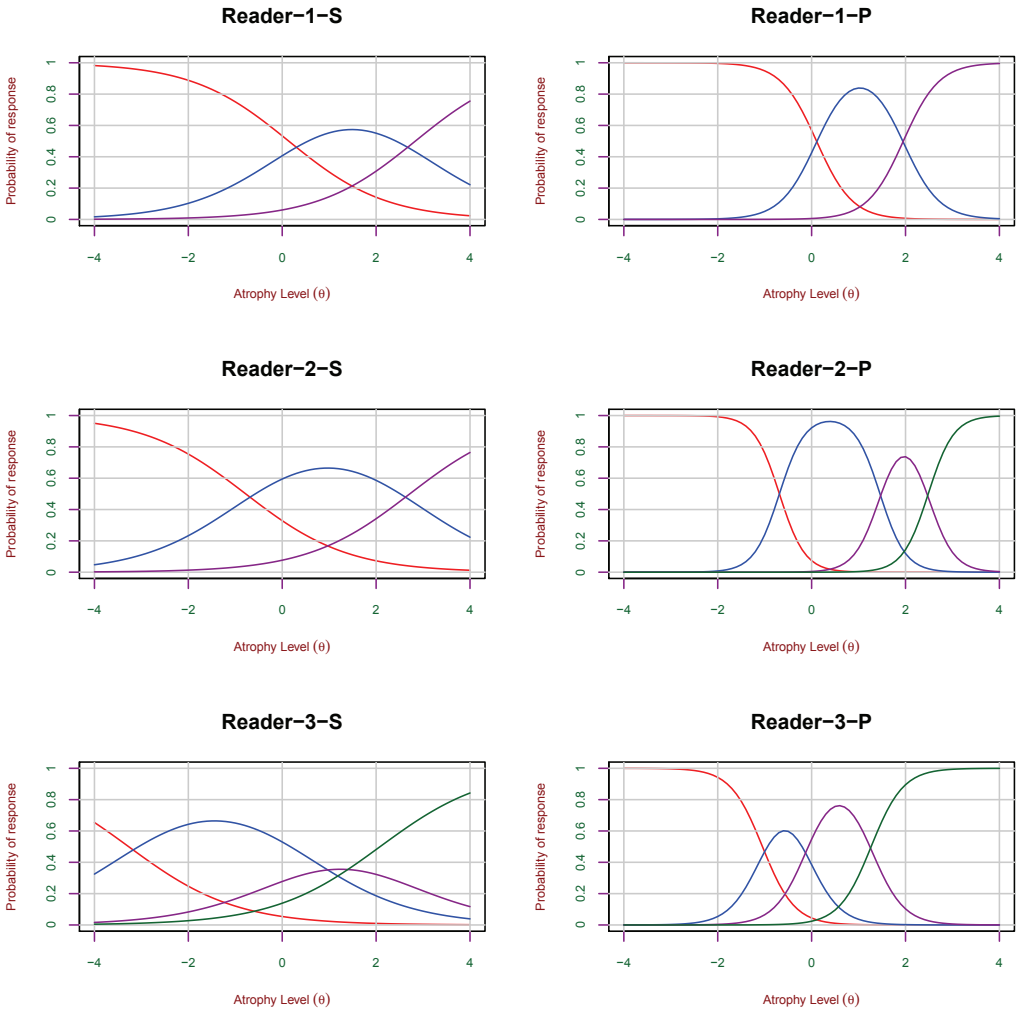
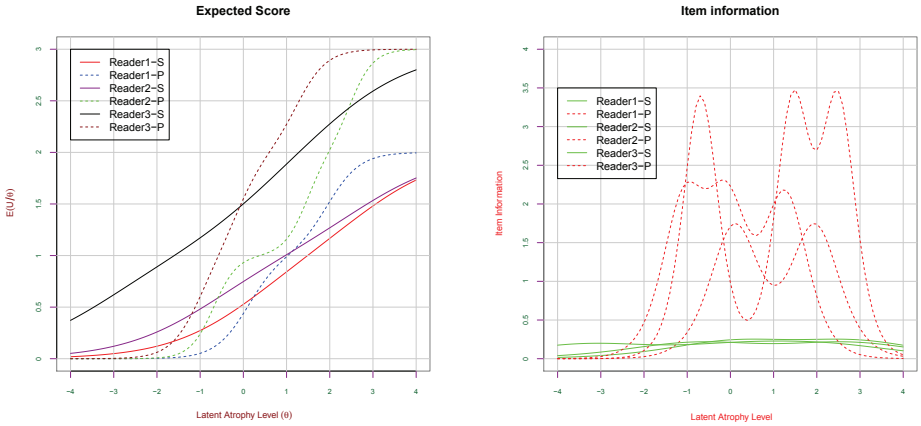


Figure 1: *Category response curves for all 6 items under the fitted Graded Response Model. The red, blue, magenta and green lines represent the "absence of atrophy", "mild", "medium", and "severe" levels of atrophy, respectively (S = supine; P = prone).*

GRM.

The scale (or test) information curve (or function) indicates the level of information (i.e., reliability) provided by the scale (atrophy rating in our case) over the range of the construct continuum (the latent true atrophy level). Figure 3 presents the scale information function and the associated reliability ($r = 1/\text{information}$) for all the items, three supine items as well as three prone items of the atrophy evaluation. The scale is reliable ($r \geq 0.75$) to measure atrophy level across mild and moderate categories for all items as well as only for prone items. The function peaks at the middle of the scale to indicate that mild and moderate atrophy level is measured with most precision. Reliability decreases when measuring the absence of atrophy as well as the severe level of atrophy. By considering only the supine items, it gives the least reliability as well as least consistency in terms of rating the atrophy



(a) Expected score curves for all items under the Graded Response Model

(b) Item information curve for all items modeled by the GRM

Figure 2: *Expected score and information function curves for all items.*

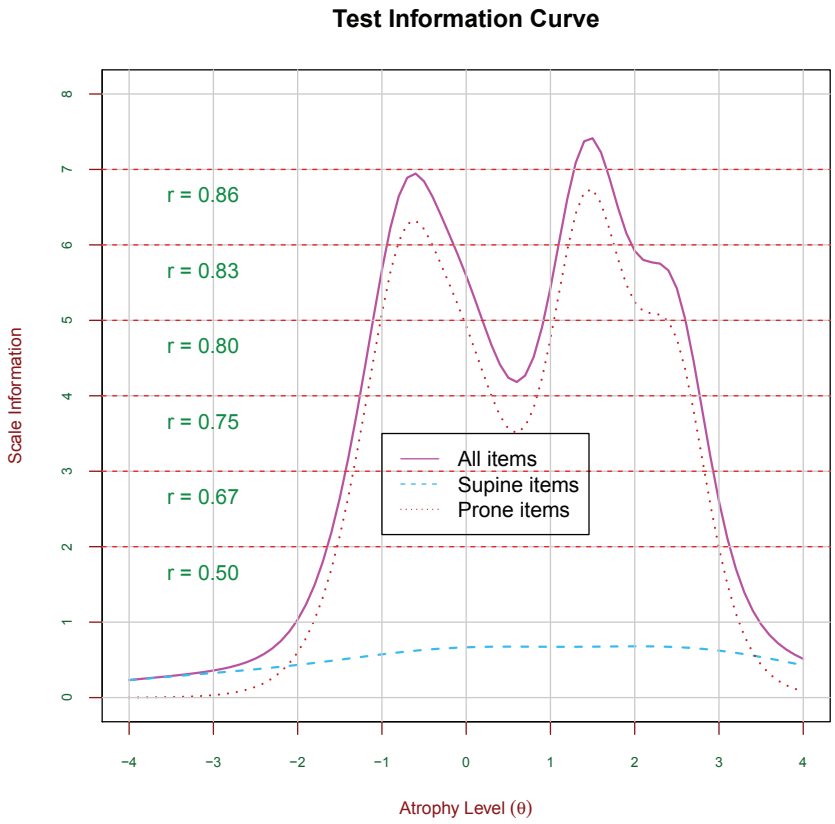


Figure 3: *Scale information curves for all items, 3 supine items and 3 prone items for the atrophy scale. Horizontal dashed lines indicate the approximate level of reliability associated with different information magnitudes.*

level by the three readers.

4. Conclusion

This study evaluated cerebral atrophy for both supine and prone positions by IRT. A GRM, determined by the item threshold and slope parameters, was fitted to our data that was followed by an evaluation of item-fit and person-fit statistics to assess model fit.

The item slope estimates from the GRM varied substantially; however, they were consistent for the same position. All item slopes were greater than 2 for the readings in the prone positions; however, all slope parameters were less than 1 for the supine position. This indicates that the readers were able to better differentiate patients with different atrophy levels in the prone position and less able to differentiate the patients with different levels of atrophy and the supine position.

The category threshold estimates vary considerably between readers. Within readers, the thresholds are similar for the first reader; however, quite different for the second and third readers. In case of the second reader, b_2: the threshold between 'mild' and 'medium', is especially lower for the supine position due to a higher proportion of patients being rated as moderate or severe at this position. For the third reader, b_1: the threshold between 'absence of atrophy' and 'mild', is lower for supine position corresponding to a lower proportion of patients rated as absent of atrophy. This suggests that the scanning position affects the readers differently.

By examining the item information curves, which depend on both the size of the slope and the category thresholds, we learned that the radiologist's evaluation has more precision based on the prone position. After examining the person-fit statistics, we noticed two individuals with considerable inconsistency between the supine and prone positions.

Based on the exploratory factor analysis, the unidimensional assumption is not fully satisfied as the primary factor accounts for only 39% of the total variation while the factors yielded from the two-factor analysis explain 60% of the total variation with the first factor depending on the prone position ratings and the second factor depending on supine position ratings. Therefore, it could be interesting to consider a multidimensional GRM.

In conclusion, we did not observe consistently a higher level of cerebral atrophy evaluation based on the supine position, however, our analysis indicates that the cerebral atrophy level was better discriminated in the prone position with relatively better measurements of mild and moderate levels.

Acknowledgements

We are grateful to the MS/MRI Research Group, Department of Medicine, University of British Columbia, Vancouver, Canada for providing us their MRI data.

References

Adams R, Rosier M, Campbell D, and Ruffin R (2005). Assessment of an asthma quality of life scale using item-response theory, *Respirology*, **10**, 587–593.

Andrich D (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers, *Applied Psychological Measurement*, **2**, 581–594.

Bock RD (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika*, **37**, 29–51.

Cai L (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm, *Psychometrika*, **75**, 33–57.

Cattell RB (1966). The screen test for the number of factors, *Multivariate Behavioral Research*, **1**, 245–276.

Cattell RB (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*, Plenum, New York.

Chalmers RP (2012). mirt: a multidimensional item response theory package for the R environment, *Journal of Statistical Software*, **48**, 1–29.

Coleman MJ, Cook S, Matthysse S, Barnard J, Lo Y, Levy DL, Rubin DB, and Holzman PS (2002). Spatial and object working memory impairments in schizophrenia patients: a Bayesian item-response theory analysis, *Journal of Abnormal Psychology*, **111**, 425–435.

Drasgow F, Levine MV, and Williams EA (1985). Appropriateness measurement with polychotomous item response models and standardized indices, *British Journal of Mathematical and Statistical Psychology*, **38**, 67–86.

Edelen MO and Reeve BB (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement, *Quality of Life Research*, **16**, 5–18.

Embretson, SE and Reise SP (2000). *Item Response Theoryfor Psychologists*, Lawrence Erlbaum Associates, Mahwah, NJ.

Hays RD, Liu H, Spritzer K, and Cella D (2007). Item response theory analyses of physical functioning items in the medical outcomes study, *Medical Care*, **45**, S32–S38.

Hays RD, Morales LS, and Reise SP (2000). Item response theory and health outcomes measurement in the 21st century, *Medical Care*, **38**, II28–II42.

Holman R, Glas CA, and de Haan RJ (2003). Power analysis in randomized clinical trials based on item response theory, *Controlled Clinical Trials*, **24**, 390–410.

Kang T and Chen TT (2007). *An Investigation of the Performance of the Generalized $S - X^2$ Item-Fit Index for Polytomous IRT Models*, ACT Research, Columbus, IN.

Karas G, Scheltens P, Rombouts S, van Schijndel R, Klein M, Jones B, van der Flier W, Vrenken H, and Barkhof F (2007). Precuneus atrophy in early-onset Alzheimer's disease: a morphometric structural MRI study, *Neuroradiology*, **49**, 967–976.

Loehlin JC (2004). *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis* (4th ed), Lawrence Erlbaum Associates, Mahwah, NJ.

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

McCullough D, Levy L, DiChiro G, and Johnson D (1990). Toward the prediction of neurological injury from tethered spinal cord: investigation of cord motion with magnetic resonance, *Pediatric Neurosurgery*, **16**, 3–7.

McDonald RP (1981). The dimensionality of tests and items, *British Journal of Mathematical and Statistical Psychology*, **34**, 100–117.

Muraki E (1992). A generalized partial credit model: application of an em algorithm, *Applied Psychological Measurement*, **16**, 159–176.

Muraki E (1997). A generalized partial credit model. In van der Linden WJ and Hambleton RK (Eds), *Handbook of Modern Item Response Theory* (pp. 153–164), Springer, New York.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Achievement Tests*, Danish Institute for Educational Research, Copenhagen.

Rasch G (1961). On general laws and the meaning of measurement in psychology, *In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 321–333.

Reise SP (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT,

Applied Psychological Measurement, **14**, 127–137.
Reise SP (2016). The emergence of item response theory models and the patient reported outcomes measurement information systems, *Austrian Journal of Statistics*, **38**, 211–220.
Reise SP and Flannery P (1996). Assessing person-fit on measures of typical performance, *Applied Measurement in Education*, **9**, 9–26.
Reise SP and Waller NG (2009). Item response theory and clinical measurement, *Annual Review of Clinical Psychology*, **5**, 27–48.
Reise SP, Widaman KF, and Pugh RH (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance, *Psychological Bulletin*, **114**, 552–566.
Samejima F (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, **34**, 100.
Samejima F (1972). A general model for free-response data, *Psychometrika Monograph Supplement*, **37**, 68.
Samejima F (1997). Graded response model. In van der Linden WJ and Hambleton RK (Eds), *Handbook of Modern Item Response Theory* (pp. 85–100), Springer, New York.
Thissen D and Steinberg L (1986). A taxonomy of item response models, *Psychometrika*, **51**, 567–577.
Witkamp TD, Vandertop WP, Beek FJ, Notermans NC, Gooskens RH, and van Waes PF (2001). Medullary cone movement in subjects with a normal spinal cord and in patients with a tethered spinal cord 1, *Radiology*, **220**, 208–212.
Yen WM (1981). Using simulation results to choose a latent trait model, *Applied Psychological Measurement*, **5**, 245–262.

Received July 20, 2016; Revised November 3, 2016; Accepted November 3, 2016