

Chapter 11

Application of Polytomous Rasch Models to the Subtest Written Language of the Aachen Aphasia Test (AAT)

Klaus Willmes

Institute for Psychology & Neurological Clinic, Technical University Aachen

1. Introduction

Aphasias are acquired language disorders in adults subsequent to acquired (focal) brain damage predominantly due to a cerebrovascular lesion (about 80%), brain trauma, tumor or inflammatory as well as degenerative disease of the brain affecting the cerebral hemisphere dominant for language, which is the left one in the great majority of cases (LaPointe, 1990).

Aphasias can be described linguistically as impairments in the different components of the language system: phonology, lexicon, semantics, and syntax. In general, aphasic disorders are multimodal or even supramodal, i.e., they affect several or all expressive and receptive language modalities: oral language production and auditory comprehension as well as reading and writing to a similar degree and in a similar way as far as the quality of linguistic impairments is concerned.

Despite more recent theoretical developments in cognitive neurolinguistics in terms of explicit language processing models leading to more theory-oriented experimental assessment procedures, diagnostic examinations of aphasic language impairments are still required from a clinical and practical perspective, which calls for a rather comprehensive assessment of the most relevant basic language functions within reasonable time limits of about 60-90 minutes. This aim has been accomplished internationally by some test batteries (for an overview see Willmes, 1993).

In neuropsychology and aphasiology, in particular, psychometric properties of (sub-)tests and items have almost exclusively been evaluated by means of the classical test theory model. One practical reason is that probabilistic test models require sizeable samples for proper parameter estimation and tests of model fit, but collecting data on hundreds of neurologically and neuroradiologically well documented cases of aphasia is very time consuming. Likewise, considerations of construct validity on the subtest or even item level have not been paid much attention. The only application of the (dichotomous) Rasch model in aphasiology has been for the Token Test, a test requiring auditory comprehension (De Renzi & Vignolo, 1962), which is considered the most adequate global measure of the overall level of aphasic impairment. It could be demonstrated for the German version (Willmes, 1981) that the first four parts of the test as well as the final fifth part alone are in agreement with the Rasch model for dichotomous scores. Additional analyses with the linear logistic model revealed that basic parameters related to processing information about the shape, size, and colour of the tokens to be pointed to on verbal command could serve to reproduce the item difficulties. Recently, Butter, De Boeck, and Verhelst (1994) used German and Dutch Token Test data for a successful application of their Rasch model with

internal restrictions on the item difficulty parameters, which also allows for heterogeneous item discrimination.

For clinical and practical diagnostic purposes standardized test information on the overall degree of impairments and the quality of errors observed is usually considered sufficient. Nevertheless, when enough data about a comparatively well constructed aphasia test are available it is reasonable to carry out a more detailed analysis of the psychometric properties of the assessment procedure. Therefore, the aim of this study was to learn more about the item properties of one particular subtest of the Aachen Aphasia Test, the first standardized aphasia test for the German language (AAT; Huber, Poeck, Weniger, & Willmes, 1983). One particular subtest, Written Language, was chosen for presentation not just because of space constraints. It was also expected that because of its constructional properties (see below) there would be a fair chance that a probabilistic test model assuming one quantitative latent trait would hold for each of its three parts. More specifically, the study was undertaken to examine:

- (1) whether the polytomous Rasch model (Partial Credit model) holds, thus indicating that expressive impairments of aphasic patients in written language modalities can be measured each on a latent continuum;
- (2) whether the response categories are ordered in the sense that they can be mapped onto successive intervals of the latent trait;
- (3) whether the difficulty order in terms of linguistic processing complexity (see below) holds;
- (4) whether - in case of reasonable model fit - restricted polytomous Rasch models like the Rating Scale model (assuming identical threshold distances across items, see eq. (17) in chapter 1.4) or a dispersion model (assuming equidistant thresholds per item, see eq. (18) in chapter 1.4) can be assumed to hold; or whether substantial deviations from these restricted models are present that can be identified introducing a skewness parameter (Andrich, 1985) in addition to item dispersion and location.
- (5) Finally, since the AAT is a diagnostic instrument, it was also tempting to look for patients with grossly deviant person fit measures in order to find (post hoc) explanations for these discrepancies.

2. Test instrument

The linguistic subtests of the AAT have been constructed according to explicit linguistic criteria. These four subtests: Repetition, Written Language, Confrontation Naming, and (auditory and written) Language Comprehension tap on the major language modalities. The subtests are made up of three to five parts containing 10 items each. A detailed description of the linguistic structure of all test items is given in Willmes, Poeck, Weniger, and Huber (1983). This linguistic structure can be cast within the facet theory framework of Louis Guttman (Shye & Elizur, 1994) introducing a facet theory oriented definition of language test items: An item belongs to the universe of language items if its domain concerns a (phonological, semantic, syntactic) *regularity* on the (phoneme, word, sentence) level and it calls for a response toward that linguistic *unit* in a (expressive, receptive) *language modality* and its response range is ordered from very correct to very wrong with respect to that linguistic regularity.

Another property of language test items is that the facets regularity and unit cannot be fully crossed. Particular phonemic or graphemic, morphological, semantic, and syntactic regularities are tied to particular linguistic units, i.e., phonemes or graphemes, morphemes, lexemes or sentences. This fact is taken into consideration by introducing one combined facet „linguistic processing complexity“ which can be crossed with the modality facet.

These constructional principles have been followed in particular for the three parts of the subtest Written Language. The linguistic structure of the items is presented in Table 1. Corresponding items in the three parts with identical sequence number are very similar linguistically.

Item	Linguistic Structure	Reading aloud	Putting together on dictation*	Writing to dictation
	<i>One-syllable word</i>			
1	CVC	Wahl	S/A/A/L	Tal
2	CCVCC	Quirl	QU/A/L/M	Quark
3	CCVCC	schlicht	SCH/L/A/CH/T	Schlucht
	<i>Two-syllable word</i>			
4	C(C)V(C)+CC(C)VC	Sportler	P/R/IE/S/T/E/R	Künstler
	<i>Three-syllable word</i>			
5	C(C)V+CV+CV	Blamage	G/A/R/A/G/E	Montage
	<i>Compound noun</i>			
6	Adj-suffix	Eitelkeit	TRAURIG/KEIT	Heiterkeit
7	N/Adj-N-N	Schaumgummipolster	KAMEL/HAAR/ MANTEL	Leichtmetalleiter
	<i>Sentence</i>			
8	Declarative: NP(Pro)-V(intrans.)-NP(poss.)	Sie will mein Auto	ER/HAT/SEIN/ BUCH	Sie wird seine Frau
9	Declarative: NP(Pro)-V-(Adj-Inf+Copula)	Er pflegte eitel zu sein	SIE/SCHEINT/ TRAURIG/ZU/SEIN	Er glaubte heiter zu sein
10	Interrogat: Adv(Pro)-Aux-NP(Pro)-NP(Pro)-NP(Pro)-V	Warum will er es mir geben	WANN/HAT/SIE/ ES/UNS/GESAGT	Wohin wird sie es mir bringen

* Slashes separate segments that are presented in the same 16-choice set of either 16 letters (items 1-5) or morphemes/words (items 6-10)

Table 1: Linguistic structure and stimuli of the AAT subtest Written Language

There is increasing linguistic complexity from item 1 to 10: words precede compound words followed by sentences. Among the words there is an increasing number of syllables. The processing complexity of the one-syllable words gets bigger with respect to initial and final consonant clusters as well as their phonological complexity. The one-syllable words are followed by a bisyllabic multiconsonant word and the three-syllable loan word is characterized by a different phonotactic structure compared to German. The two compound nouns differ in the number of constituents and in whether their components may not (item 6) or may be (item 7) words of their own. The syntactic complexity of the final three sentences again gets higher and in addition their overall complexity is raised because of pronominalization.

In part 1 (Reading aloud) each item is printed on a separate sheet and the patient has to read it aloud. In part 2 (Putting together to dictation) the word or sentence constituents of the auditorily presented word or sentence have to be picked from a 16-choice set of graphemes (items 1-5) or morphemes/words (items 6-10) and have to be placed in the correct order. Part 3 (Writing to dictation) requires the patient to simply write words or sentences to dictation.

Nonmetric multidimensional scaling analyses for the AAT have revealed that there is a clearcut separation between items according to language modalities in line with the three parts of the subtest (Willmes et al., 1983), but no obvious structuring with respect to linguistic complexity. These results also support the decision to carry out the analyses with polytomous test models for each of the three parts of the subtest separately.

Scoring of responses

Scoring aphasic language test performance is not straightforward. A simple pass/fail scoring would not efficiently capture the large range and quality of possible language impairments per item. Rost (1996) provides some general arguments for the utility of ordinal item scorings. Separately for each subtest of the AAT, an ordinal 4-point scale is used for all items; it is presented in Table 2. The graded scores are intended to denote the varying degree of similarity of the patients' responses with the target response which may vary from no or a vastly deviant response (score 0) to a correct response (score 3). The intermediate scores 1 and 2 indicate decreasing degrees of deviation from the target according to fixed linguistic criteria. Incorrect responses scored 1 or 2 can be rather diverse according to the type of aphasia and language processing impairment. Phonemes/graphemes, morphemes or words may be substituted, deleted, permuted or changed phonemically. There may also be augmentations or multiple changes, which are combinations of the former. A score of 2 is also given for correct responses after self-corrections or a requested second presentation of the item since these are very often indicative of minor language processing problems as well.

<i>Score</i>	<i>Description</i>	<i>Score</i>	<i>Description</i>
3	Correct response	1	Little similarity with target - more than 1/3 of stimulus elements changed
2	High similarity with target - no more than 1/3 of stimulus elements changed - correct response after self-correction - correct response after 2nd presentation	0	No similarity with target - no response - neologistic utterance - fragmentary response

Table 2: Item scoring for AAT subtest Written Language

Subjects

For the analyses, a large sample ($n = 1769$) from the computerized AAT data base (Willmes & Ratajczak, 1987) was used. Selection criteria were homogeneity with respect to etiology (only vascular origin), age (adult patients not older than 75 years), and adequate neurological and neuroradiological information. Table 3 lists sample characteristics including the grouping of patients according to major aphasic syndromes.

<i>Type of aphasia</i>	<i>n</i>
Global aphasia	599
Wernicke's aph.	395
Broca's aph.	393
Amnesic aph.	306
Other	76
Total	1769
Sex (female/male)	444/1325
Age (yrs.): median (range)	55 (18-75)
Duration (months)	8 (1-221)

Table 3: Sample characteristics

3. Computations

Analyses for the Partial Credit model were carried out by means of the PC-program QUEST (Adams & Khoo, 1993), which provides unconditional maximum likelihood estimates of ability and threshold parameters (The LPCM program by Fischer and Ponocny (1994) computing conditional maximum likelihood estimates was not available at the time of data analysis). The estimation routines as well as the item- and person-fit statistics and the tests for model comparisons in the QUEST package closely follow the exposition of these methods in the book by Masters and Wright (1982). They introduced several statistics for the assessment of item-fit to the Partial Credit model. The so-called 'infit' weighted t-statistic and the weighted mean square fit statistic express the amount of discrepancy between the observed score and the score predicted by the Partial Credit model. Large positive t-statistic values are indicative of poorly discriminating items, negative values hint at over-discrimination. In the technical notes of the QUEST handbook Adams and Khoo (1993) argue that for large sample sizes the t-statistic may be too sensitive. They suggest accepting compatibility of an item with the Partial Credit model if the weighted mean square residual has a value from the interval 0.77 to 1.30, indicating a maximum of 30% deviation from the expected value = 1 of that statistic when the model holds.

In case of four response categories, an equivalent parametrisation (Andrich 1983, 1985; Rost, 1988) for the threshold parameters can be given in terms of a location parameter (arithmetic mean of the threshold parameters), a dispersion parameter (arithmetic mean of differences between adjacent threshold parameters; θ^* in the notation of Andrich (1985)), and a skewness parameter (arithmetic mean of the difference between the dispersion parameter and the adjacent threshold differences, denoted η^* by Andrich). The location, dispersion, and skewness parameter estimates themselves can easily be computed from the threshold estimates according to the definitions just given in brackets using the threshold parameter point estimates provided by the QUEST program. In order to test for differences in

dispersion and skewness parameters among items and to obtain standard errors of estimates for these parameters, the SKEWLOC version of the DISLOC program (Andrich, De'Ath, Lyne, Hill, & Jennings, 1983) was used.

The dispersion parameters of the Partial Credit model or the Rating Scale model also allow for a binomial trials interpretation (Masters & Wright, 1984). If it seems reasonable to treat the score attained as if it were the sum of successful independent attempts at a task irrespective of the order of successes and failures, then the dispersion parameter ought to have a fixed value depending only on the number of graded scores of an item. For an item with four different scores this value is 1.10 (cf. Table 1 in Masters & Wright, 1984). In case of positive dependence among 'trials' the dispersion parameter becomes substantially less than 1.10. If the dispersion value is much larger than 1.10 this is compatible with independent trials of varying difficulty.

4. Results

As stated before, only separate analyses for the three parts of the subtest Written Language were carried out, since presence of a uni-dimensional latent trait was only expected for each sub-modality of written language processing. In Table 4 the results of analyses with the Partial Credit model are summarized. In the three leftmost columns the three threshold estimates for each item are given. In addition, the next two columns to the right provide differences between adjacent thresholds for a better apprehension of variability in these threshold differences within and across items. The next three columns give the estimates for the item reparametrization in terms of location, dispersion and skewness. Item ordering with respect to location and score thresholds, respectively, is indicated in brackets. Finally, the two rightmost columns present information on item fit with respect to the Partial Credit model.

- (1) Generally, the 10 items of each subtest part fit the Partial Credit model reasonably well. There is one item (no. 9) in part 1 and four items (no. 1, 2, 7, 9) in part 3 with poor fit according to the 30%-criterion for the weighted mean square residual statistic. One way of assessing this lack of fit is to look for differences in item properties across different ability groups. If one divides the raw score range into five intervals to contain a similar number of patients, for item 9 in part 1 the observed frequency of low score 0 or 1 is higher than expected from the model in the two lowest scoring patient groups and that the frequency of item score 2 is too low. A similar effect is found for the difficult items 7 and 9 in part 3. The observed frequency for score 1 is too low in the two lowest scoring groups. Quite to the contrary, for the two most simple items 1 and 2 too many patients with poor performance attain high scores 2 or 3 too often. On the other hand, for item 1 there are too many patients scoring less than 3 in the two best subgroups, possibly indicating some minor problems in becoming familiar with the new task. A similar overshoot of scores 1 and 2 hints at particular problems in correctly writing the initial grapheme /Qu/. This irregular phoneme-to-grapheme conversion seems to be very difficult even for the least impaired patients.
- (2) With two minor exceptions, the graded response categories are mirrored in the ordering of threshold estimates. The irregularities in the easiest item 1 in part 1 and in the very difficult item 9 in part 2 are, however, within one standard error limit around the point

estimates. For item 1 it is about as easy to get a score of 2 instead of 1. For such a short and easily read word it is not very likely that more than one grapheme is wrong if a patient can read it at all in an intelligible way. The ordering for item 9 is violated at the other end of the item scale. With 5 constituents, a score of 2 is attained only if no more than just one constituent is wrong. Therefore, getting all the elements right or failing on just one of them does not make a real difference in ability. The threshold for scoring 2 rather than 3 is even considerably higher than the one for the most difficult item 10 with 6 elements, in which 2 of them may be failed for a score of 2. In summary, ordinal scale properties may be assumed to hold for the item scoring of the subtest Written Language.

- (3) The most straightforward way of examining agreement between the item difficulty ordering and expectations about the linguistic processing complexity of the items is to look at the rank order of the location parameter estimates. Placing ‘confidence-intervals’ of ± 2 times the standard error around the location estimate reveals that there are two major ordering violations in part 1: item 4 and more so item 8 are too easy. Although the noun has two syllables with two consonant clusters in it, it is a highly familiar word with high frequency of occurrence. The first declarative sentence of part 1 is very easy, mainly due to the very low first threshold. The sentence contains the high-frequency content word /Auto/ which is also easy to be read aloud due to its very simple phonotactic structure. Of course it cannot be deduced from a score of 1 which of the four sentence constituents was actually read aloud correctly, but it is known that the remaining three closed class words, a modal verb and two pronouns in particular, are difficult to process for many aphasic patients (LaPointe, 1990). Since a score of 2 is credited only if just one of the four words is read incorrectly, the highest threshold is not so easy at all.

Concerning part 2, one has to remember that the first and the second half of the items have to be looked at separately since the counting unit for errors changes from graphemes to words. For each part separately there are no substantial violations of the ordering. It is also interesting to note that items 6 to 9 are as simple or even more simple than items in the first half since the tokens to be allocated are already whole words.

In part 3, the ordering closely fits the theoretical expectations, the only more moderate exception being again the first simple declarative sentence. The difficulty is reduced because of the first threshold which can be passed already when the high-frequency noun /Frau/ is written correctly. On the other hand (like in part 1), the more complex second compound noun is even as difficult as the second sentence stimulus. This is due mainly because of the very high last threshold which expresses the difficulty to write down (or read aloud) the complete multi-compound noun correctly.

- (4) A quick glance at the threshold differences in Table 4 already reveals that neither the Rating Scale model nor a model with equidistant thresholds per item can hold. This is also obvious if one compares these threshold differences with those computed after an analysis with the Rating Scale model as shown in Table 5. Tests provided by QUEST for the match of threshold values computed under the Partial Credit and the Rating Scale model, as reported in that table, yielded overall fit statistic values (t-values) no less than $t = 5.6$ (for score 2 in part 3, all $p < 0.0001$). The total t-fit statistics per test part were even higher (t-values above 12.0 throughout). In addition, chi-square tests of the identity of dispersion or skewness parameters provided by SKEWLOC clearly revealed that for

each test part there were highly significant differences (all $p < .0001$) between items as regards the dispersion and skewness parameters.

Item	Partial Credit threshold estimates			Differences		Reparametrisation			Item-fit ³	
	1	2	3	2 - 1	3 - 2	location ¹	dispersion ²	skewness ²	t-value	MS
Part 1: Reading aloud ($n'=1421$)										
1	-1.61 (3)	-1.76 (1)	-0.44 (1)	-0.15	1.32	-1.27 (1)	0.43d	0.82*	3.63	1.20
2	-1.18 (5)	-0.12 (7)	1.02 (3)	1.06	1.14	-0.09 (5)	1.10	0.04	3.57	1.16
3	-1.50 (4)	-0.54 (3)	1.18 (4)	0.96	1.72	-0.29 (4)	1.34+	0.38*	-0.01	1.00
4	-1.68 (2)	-1.07 (2)	0.93 (2)	0.61	2.00	-0.61 (2)	1.31+	0.69*	0.25	1.01
5	-1.04 (6)	-0.42 (4)	1.62 (5)	0.62	2.04	0.05 (6)	1.33+	0.71*	-1.28	0.95
6	-0.90 (8)	-0.26 (6)	1.65 (6)	0.64	1.91	0.16 (7)	1.28	0.64*	3.56	1.15
7	-0.83 (9)	-0.03 (8)	3.25 (9)	0.80	3.28	0.80 (8)	2.04+	1.24*	-4.87	0.82
8	-2.60 (1)	-0.35 (5)	1.66 (7)	2.25	2.01	-0.43 (3)	2.13+	-0.12	-1.70	0.94
9	-0.95 (7)	0.45 (10)	2.91 (8)	0.50	2.46	0.80 (9)	1.93+	0.53*	-8.39	0.72
10	-0.66 (10)	0.04 (9)	3.26 (10)	0.70	3.22	0.88 (10)	1.96+	1.26*	-3.73	0.86
Part 2: Putting together on dictation ($n'=1433$)										
1	-1.66 (1)	-0.69 (1)	0.44 (1)	0.97	1.13	-0.64 (1)	1.05	0.08	6.56	1.30
2	-0.89 (4)	-0.08 (2)	1.67 (4)	0.81	1.75	0.23 (2)	1.28	0.47*	1.17	1.05
3	-1.23 (2)	0.87 (5)	1.39 (2)	2.10	0.52	0.34 (4)	1.31+	-0.79*	-0.86	0.97
4	-0.90 (3)	0.26 (4)	1.60 (3)	1.16	1.34	0.32 (3)	1.25	0.09	-4.90	0.82
5	-0.85 (5)	0.25 (3)	1.69 (5)	1.10	1.44	0.36 (5)	1.27	0.17	-1.48	0.94
6	-1.98 (3)	-0.69 (2)	0.43 (1)	1.29	1.12	-0.75 (2)	1.20	-0.09	5.16	1.23
7	-3.07 (1)	-0.93 (1)	0.99 (2)	2.14	1.92	-1.00 (1)	2.03+	-0.11	-1.23	0.95
8	-2.24 (2)	0.26 (3)	1.10 (3)	2.50	0.84	-0.29 (3)	1.67+	-0.83*	-4.72	0.83
9	-1.83 (4)	1.58 (5)	1.48 (4)	3.41	-0.10	0.41 (4)	1.65+	-1.77*	-3.20	0.88
10	-1.09 (5)	1.15 (4)	2.97 (5)	2.24	1.82	1.01 (5)	2.03+	-0.21	-2.17	0.92
Part 3: Writing to dictation ($n'=1282$)										
1	-2.44 (2)	-2.25 (1)	-1.31 (1)	0.19	0.90	-2.01 (1)	0.55d	0.36	10.72	1.68
2	-2.58 (1)	-0.39 (3)	0.49 (3)	2.19	0.88	-0.83 (2)	1.54+	-0.66*	8.22	1.39
3	-1.74 (3)	-0.26 (4)	0.10 (2)	1.48	0.36	-0.63 (3)	0.92	-0.56*	0.08	1.00
4	-1.25 (5)	-0.62 (2)	1.16 (6)	0.63	1.78	-0.24 (4)	1.21	0.58*	-0.70	0.97
5	-0.97 (6)	-0.05 (5)	0.61 (5)	0.92	0.66	-0.14 (5)	0.79d	-0.13	-1.10	0.95
6	-0.68 (7)	-0.02 (6)	0.51 (4)	0.66	0.53	-0.06 (6)	0.60d	-0.07	-3.96	0.82
7	-0.45 (8)	1.06 (8)	2.68 (9)	1.51	1.62	1.10 (9)	1.57+	0.06	-6.52	0.75
8	-1.31 (4)	0.59 (7)	1.42 (7)	1.90	0.83	0.23 (7)	1.37+	-0.54*	-5.36	0.79
9	-0.19 (9)	1.06 (9)	2.34 (8)	1.25	1.28	1.07 (8)	1.27	0.02	-7.91	0.69
10	0.21 (10)	1.39 (10)	2.93 (10)	1.18	1.54	1.51 (10)	1.36+	0.18	-4.95	0.80

n' sample size without patients having extreme scores of 0 or 30

¹ arithmetic mean of item threshold estimates ('item difficulty')

² mean threshold difference among adjacent thresholds (θ^*) and mean deviation from mean dispersion (η^* , see Andrich, 1985)

³ weighted residual-based fit statistics ('infit measures', see Wright & Masters, 1982)

* estimate substantially different from 0, i.e. estimate $\pm 3.0 \times \text{s.e.}$ does not cover 0

d dependency according to binomial trials model, i.e. estimate $+ 3.0 \times \text{s.e.} < 1.10$; see text

+ compatibility with binomial trials model assuming differences in trial difficulties,

i.e. estimate $- 3.0 \times \text{s.e.} > 1.10$

Table 4: Partial Credit analysis for the 3 parts of the subtest Written Language

Nevertheless, there are some general differences among the three test parts in the type of deviations from more restricted polytomous models. In part 1, for the majority of items the threshold distance between scores 1 and 2 is much smaller than between scores 2 and

3. This fact is mirrored by 8 skewness parameters being significantly above 0 (cf. Table 4), which are particularly large for items 7 and 10. Obviously, it is very difficult to read correctly all of the constituents either because many ‘little’ (closed class) words have to be pronounced in the correct sequence (item 10) or because three nouns have to be correctly assembled in a response buffer before being read aloud. Similar arguments

Part	Rating Scale model threshold estimates ¹			Differences		Global fit	Average	
	1	2	3	2-1	3-2	t-statistic	dispersion	skewness
1	-1.33 (11.75)	-0.31 (7.27)	1.64 (13.91)	1.02	1.95	16.29 p<.0001	1.49	0.47
2	-1.54 (12.90)	0.22 (11.62)	1.32 (8.93)	1.76	1.10	16.16 p<.0001	1.43	0.33
3	-1.13 (10.15)	0.13 (5.57)	1.00 (9.54)	1.26	0.87	12.42 p<.0001	1.07	0.30

¹ Test of Rating Scale model to Partial Credit model fit of threshold parameters
(t-values given in brackets; all p<.0001)

Table 5: Summary results of Rating Scale model analysis

could be provided for the other items which show this asymmetry between thresholds to a somewhat lesser degree. Finally, invoking a binomial trials interpretation, it is interesting to note that the dispersion parameter estimates are often larger than 1.10, sometimes significantly so. This lends support to the notion that similar to independent trials, although of unequal difficulty, patients can get partial credit for different parts of the stimuli when uttered correctly. Item 1 violates this general pattern: there is dependence of steps in a way already indicated in (1). If a patient can somehow read the short word in an intelligible way it will be very unlikely to get just one third of it correct; just a small discrepancy with the target is much more likely.

In contrast to part 1, there is almost a reversal in the size of the threshold difference in part 2. Since the stimulus is presented auditorily only once, there is a heavy load on verbal working memory compared to part 1, in which the written stimulus is present all the time. Therefore, if a patient can manage more than two thirds of the constituents not much more of the latent ability is required for a totally correct response (items 3, 8, 9). For item 2 there is a tendency in the opposite direction. Since the initial grapheme /QU/ has an irregular phoneme-grapheme correspondence it seems to be rather difficult to get all the constituents right.

The pattern of threshold differences looks a little bit more regular for part 3. A similar argument like the one for part 2 can be given to account for the three large negative skewness parameter estimates. Item 4 with positive skewness is different in that it is complicated to write down the large middle consonant cluster in a completely correct way. The strong item step dependency for items 1 and 6 has a quite obvious interpretation as well. If they can be written down at all, it is very unlikely that there will be deviations from the target at any position(s) resulting in 2 out of the 3 constituents to be made wrong; e.g., for item 6 it is quite likely to have the two-syllable part /Heiter/ correct and to fail on the suffix /-keit/. In item 5 only the irregular phoneme-to-grapheme transformation for /-ge/ is complicated.

- (5) The QUEST program also provides fit-statistics per patient, comparing the observed and the expected response according to overall performance for that part of the subtest. The percentage of patients with t-fit statistic values outside ± 2 is 8.2%, 7.4%, and 5.8%, respectively. The percentage outside ± 3 is even less than 1%. Only two typical examples of such score patterns will be described here. For part 2 there is one patient with Wernicke's aphasia with large standardized residuals for the easy items 1 and 2, for which he scored 0 contrary to an average overall performance (ability estimate of 0.38) leading to an expected score of 3 for item 1. It is not unlikely that aphasic patients need more than the regular practice item to fully grasp what they are requested to do in a meta-linguistic task such as choosing the appropriate letters from a 16-choice set in the correct order. Again in part 2, another marked similar type of deviation from the model occurred in a patient with Broca's aphasia with an ability estimate of -0.30: for items 6-10 he obtained scores of 0,0,3,3,2 compared to expected scores of 1,1,1,0,0. The patient seems to exhibit a strong set effect taking him two items until he was acquainted with selecting from an altered multiple-choice layout now containing words instead of graphemes.

5. Discussion

The AAT has been quite successful for practical diagnostic purposes. The rather close fit of the items in the subtest Written Language to the one-dimensional Partial Credit model could not be expected right away, since the ordinal scoring scheme has to capture a large variety of different types of qualitatively different deviations from the targets. However, the intention expressed in the AAT manual that the graded scores should denote similar degrees of deviations from the target word(s) or sentences across all items of the test part was too optimistic. Only the least restrictive Partial Credit model was adequate, mostly because of substantial skewness parameters which could be both positive or negative. So far, the assessment of skewness has been predominantly confined to the study of response sets in attitude questionnaires. This study has shown that it is also very helpful in revealing rather subtle differences in processing for individual language items. Since the intermediate scores 1 and 2 are tied to a fixed proportion of correctly processed item constituents of possibly varying difficulty, substantial skewness is even to be expected. On the other hand, such powerful probabilistic test models, particularly when applied to large samples clearly reveal even rather minor constructional bugs in individual items, e.g., minor threshold reversals. These inconsistencies might be corrected in a revised edition of the test but then one would have to wait a long time before enough patients have been examined to check for better fit. Therefore, the analyses presented should be viewed more as a post hoc support for rather successful item construction in the applied field of aphasia assessment.

Acknowledgements

This study could not have been carried out without the colleagues I had the pleasure to work with on the AAT: Walter Huber, Klaus Poeck, Dorothea Weniger and the many other colleagues who have contributed by examining aphasic patients. Many thanks are also due to David Andrich who provided me with a copy of the SKEWLOC program in 1982.

References

- Adams, R. J., & Khoo, S.-T. (1993). *Quest: The interactive test analysis system* [Computer program manual]. Hawthorn: The Australian Council for Educational Research.
- Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N.B. Tuma (Ed.), *Sociological methodology 1985* (pp. 33-80). San Francisco, CA: Jossey-Bass.
- Andrich, D., De'Ath, G., Lyne, A., Hill, P., & Jennings, J. (1983). *DISLOC - A Fortran program for a Rasch model which has both a location and a scale parameter for subtests* [Computer program manual]. Nedlands: University of Western Australia, Department of Education.
- Butter, R., De Boeck, P., & Verhelst, N. (1994). *Model with internal restrictions on item difficulty allowing heterogeneous item discrimination* (Research Report 94-20). Leuven: Catholic University Leuven, Department of Psychology.
- De Renzi, E., & Vignolo, L. (1962). The Token Test: a sensitive test to detect receptive disturbances in aphasia. *Brain*, *85*, 665-678.
- Fischer, G. H., & Ponocny, I. (1994). An extension of the Partial Credit model with an application to the measurement of change. *Psychometrika*, *59*, 177-192.
- Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *Der Aachener Aphasie Test (AAT)*. Göttingen: Hogrefe.
- LaPointe, L. L. (Ed.) (1990). *Aphasia and related neurogenic language disorders*. New York, NY: Thieme.
- Masters, G. N., & Wright, D. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*, 529-544.
- Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie*. Bern: Huber.
- Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.
- Shye, S., & Elizur, D. (1994). *Introduction to facet theory*. Newbury Park, CA: Sage.
- Willmes, K. (1981). A new look at the Token Test using probabilistic test models. *Neuropsychologia*, *19*, 631-645.
- Willmes, K. (1993). Diagnostic methods in aphasiology. In G. Blanken, J. Dittmann, H. Grimm, J.C: Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies. An international handbook* (pp. 137-153). Berlin: De Gruyter.
- Willmes, K., Poeck, K., Weniger, D., & Huber, W. (1983). Facet theory applied to the construction and validation of the Aachen Aphasia Test. *Brain & Language*, *18*, 259-276.
- Willmes, K. & Ratajczak, H. (1987). The design and application of a data- and methodbase system for the Aachen Aphasia Test. *Neuropsychologia*, *25*, 725-733.
- Wright, D. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: Mesa Press.