Journal of Clinical and Experimental Neuropsychology

ISSN: 0168-8634 (Print) (Online) Journal homepage: http://www.tandfonline.com/loi/ncen19

An approach to analyzing a single subject's scores obtained in a standardized test with application to the aachen aphasia test (AAT)

K. Willmes

To cite this article: K. Willmes (1985) An approach to analyzing a single subject's scores obtained in a standardized test with application to the aachen aphasia test (AAT), Journal of Clinical and Experimental Neuropsychology, 7:4, 331-352, DOI: 10.1080/01688638508401268

To link to this article: http://dx.doi.org/10.1080/01688638508401268

Journal of Clinical and Experimental Neuropsychology 0168-8634/85/0704-0331 $3.00
1985, Vol. 7, No. 4, pp. 331-352

An Approach to Analyzing a Single Subject's Scores Obtained in a Standardized Test with Application to the Aachen Aphasia Test (AAT)*

K. Willmes
RWTH, Aachen, Federal Republic of Germany

ABSTRACT

Methods for the analysis of a single subject's test profile(s) proposed by Huber (1973) are applied to the Aachen Aphasia Test (AAT). The procedures are based on the classical test theory model (Lord & Novick, 1968) and are suited for any (achievement) test with standard norms from a large standardization sample and satisfactory reliability estimates. Two test profiles of a Wernicke's aphasic, obtained before and after a 3-month period of speech therapy, are analyzed using inferential comparisons between (groups of) subtest scores on one test application and between two test administrations for single (groups of) subtests. For each of these comparisons, the two aspects of (i) significant (reliable) differences in performance beyond measurement error and (ii) the diagnostic validity of that difference in the reference population of aphasic patients are assessed. Significant differences between standardized subtest scores and a remarkably better preserved reading and writing ability could be found for both test administrations using the multiple test procedure of Holm (1979). Comparison of both profiles revealed an overall increase in performance for each subtest as well as changes in level of performance relations between pairs of subtests.

Although a number of aphasia tests for the English language are available, the objectivity, reliability, and validity of which have been studied more or less extensively, little attention has been given to a psychometrically sound examination of a single subject's total score or profile of subtest scores. Also, the problem of identifying significant changes in performance between two test administrations is neglected in the manuals of these aphasia tests, even if retest reliability estimates are known. For the most part, test authors limit their attention to problems of selection and classification, using either taxonomic approaches or descriptive comparisons with some "typical" profile or mean profile of a syndrome group of aphasic patients. Additionally, it is often the case that only insufficient norm tables from rather small standardization samples are reported.

* The critical reading and helpful suggestions of one of the editors and anonymous reviewers is gratefully acknowledged.
Send reprint requests to: Klaus Willmes, Abt. Neurologie, RWTH Aachen, Pauwelsstr., 5100 Aachen, West Germany (FRG).

Accepted for publication: August 17, 1984.

There are at least three major reasons why a detailed and psychometrically sound analysis of single subjects' aphasia test data is of great importance. First, a comprehensive diagnosis cannot be given without a differentiated account of the patient's language disorders in the most relevant language modalities – repetition, naming, reading and writing, and comprehension. That is why the reporting of only a total aphasia score must be replaced by a comparison of performances for different modalities and, if possible, for different components (e.g., phonology, semantics, morphology, syntax). This problem of making decisions based on a given patient's profile of scores obtained in a standardized test is a general one for those engaged in the assessment of certain disorders. Obviously, the clinician/researcher is not satisfied with the mere description of the patient's test scores, be it in a differentiated or in a global manner. He wants to draw inferences about the likelihood of relations between observed and expected test performances within the reference population to which the patient of interest belongs. Consequently, procedures are needed for individual cases which relate diagnostic hypotheses to statistical hypotheses and then statistical decisions based on suitable test statistics back to diagnostic decisions. These diagnostic decisions are thus in turn affected by the same controllable types of errors as statistical decisions.

Secondly, this diagnostic information can have a direct impact on planning a systematic language therapy treatment. As speech therapy has become more elaborate and has been shown to be efficient in recent years (Basso, Capitani, & Zanobio, 1982; Benson, 1979; Reinvang & Engvik, 1980; Springer & Weniger, 1980; Weniger, 1982), evaluation for (longer) periods of routine treatment of single patients also turns out to be helpful. A speech therapist thus can obtain more detailed information regarding the course of recovery of a patient as well as a rational aid to decide on modifications, continuation, or termination of treatment.

Finally, for research purposes, groups of patients can be made more homogeneous and patients used in single-case experimental studies can be made more comparable if they are also tested routinely with some general purpose aphasia test.

One reason for the insufficient use of evaluation methods for single patients in aphasia test manuals may be that no comprehensive collection of procedures is available for standardized test scores in the English literature. The concept of psychometric single case assessment of Huber (1973, in German) provides one solution to the diagnostic problem(s) described above. The procedures are based on the classical test theory model (Lord & Novick, 1968). They can be applied to any test for which norm data from a large sample ($N \geqq 400$) and satisfactorily high reliability estimates are available. Contrary to the classical test theory model which was developed for raw scores, these procedures can be applied to standardized scores.

Although not all aspects of Huber's approach can be covered fully, the two most relevant types of inferential comparisons for single (or subsets of) subtests from a test profile will be demonstrated by using two test profiles of one aphasic

patient obtained with the Aachen Aphasia Test (AAT; Huber, Poeck, Weniger, & Willmes, 1983, in German; Huber, Poeck, & Willmes, 1984; Willmes, Poeck, Weniger, & Huber, 1983) before and after a period of speech therapy. These inferential procedures are as follows: (i) comparisons between (subsets of) subtest scores on one test application; and, (ii) comparisons between (subsets of) subtest scores from two test applications. For each of these comparisons, two different aspects are of interest: (i) Is there a reliable (significant) difference between scores beyond measurement error? (ii) How likely is the difference obtained or a still larger difference in the reference population of the subject under study? That is, what is the diagnostic validity of the difference between scores?

For the actual analysis of the AAT profile data two FORTRAN computer programs CASE1 and CASE12, written by the author, were used. These are general purpose programs which, after minimal changes, can be used for the analysis of any test for which the single-case assessment procedures are applicable (see the appendix for a description of these requirements). Compared to Huber's treatment of the topic, application of more recent multiple test procedures (Holm, 1979) made possible a unified outline for several of the procedures. The computer output of these programs given in the subsequent figures is used as a basis for explaining the single-case assessment procedures*. For a better understanding of the test procedures, a short description of the AAT and its psychometric properties is necessary (cf. Huber et al., 1984; Willmes et al., 1983).

SHORT DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF THE AAT

The AAT consists of six 6-point spontaneous speech rating scales and five subtests: Token Test (TT), German version of Orgass (1976), Repetition (REP), Written Language (WRIT), Confrontation Naming (NAME), and Comprehension (COMP), in which different units (phonemes, mono- and polysyllabic nouns, sentences) and linguistic rules are incorporated. Each subtest is composed of three to five parts having 10 items each. Except for the Token Test, which has dichotomous scoring, the responses to all items are scored on a 0-3 scale. Although defined more precisely in linguistic terms, for each subtest the scoring is as follows: 3 indicates a correct response; 2, mild or little departure from the stimulus; 1, severe departure from the stimulus; 0, complete error or no response. Thus, scores range from 0-50 for the Token Test, 0-150 for repetition, 0-90 for written language, and 0-150 for confrontation naming and comprehension. In the following, only the five subtests will be of interest.

*Copies of the two source programs as well as test data and a technical report describing the mathematical properties of Huber's procedures can be obtained upon request if a computer tape is also provided.

A detailed description of the linguistic structure of the test items is given in Willmes et al. (1983) along with a report on the construct and differential validity of the test. Validity and reliability studies are based on a sample of 120 aphasic patients (30 of each of the four standard aphasic syndromes). Test characteristics of the AAT are reported in Table 1. The standardization sample is comprised of 376 aphasic patients (90 globals, 74 Wernicke's, 79 Broca's, 71 amnesic, 62 nonclassifiable aphasics), almost exclusively with vascular etiology, and 100 control patients (41 left brain-damaged patients without language disorders, 29 right brain-damaged patients, and 30 normal controls).

A nonparametric discriminant analysis program (ALLOC, Habbema, Hermans, & van den Broeck, 1974) is used routinely for the classification of patients with language disorders. It investigates (i) what the (posterior) probability for a new patient is to be aphasic or not and, (ii) if the probability for aphasia is at least 80%, what the probability is of belonging to one of the four standard syndromes. Only if the highest of the four "syndrome" probabilities is above 70% is the patient considered to belong to one of the standard aphasic syndrome groups; otherwise, the patient is taken to be nonclassifiable.

Test-retest reliability was studied in 40 patients (10 of each standard syndrome) over an interval of 2 days. High reliability coefficients and no significant changes in level of performance (no practice or learning effects) were found both for

Table 1

Psychometric Properties of the AAT Needed for Psychometric Single-Case Analysis

	(1) Raw scores			(2) Consistency	(3) *T* score correlations				
AAT subtest	scale	*M*	*SD*	coefficient	TT	REP	WRIT	NAME	COMP
Token Test (TT)[1]	0- 50	27.13	15.08	.978	—	.702	.831	.835	.772
Repetition (REP)	0-150	96.49	44.67	.989		—	.781	.808	.689
Written Language (WRIT)	0- 90	43.73	29.74	.985			—	.838	.781
Confront. Naming (NAME)	0-120	63.96	38.85	.983				—	.808
Comprehension (COMP)	0-120	73.42	27.00	.928					—

[1] Error scores

(1) Arithmetic means and standard deviations for the AAT standardization sample (N = 376 aphasic patients: 90 global, 74 Wernicke's, 79 Broca's, 71 amnesic, 62 nonclassifiable aphasic patients)

(2) Reliability estimates (consistency coefficients X, n = 120: 30 global, 30 Wernicke's, 30 Broca's, 30 amnesic aphasic patients)

(3) Intercorrelations (Pearson) between *T* scores based on the standardization sample

patients with duration of aphasia below and above 3 months. Thus, one can feel safe in attributing substantial (significant) changes in performance to other factors such as spontaneous recovery or language therapy treatment.

Several norm tables have been set up for the AAT. Based on the whole standardization sample of 376 aphasic patients, raw scores for each subtest are transformed into centile ranks first. These centile ranks are then transformed further into *T* scores. By this 2-step procedure, the nonnormal distributions of raw scores, a problem for all aphasia tests, are turned into (quasi-) normally distributed standard scores (Guilford, 1965, Table 19.3).

For a short description of the patients' level of performance, a stanine scale derived from the centile rank distribution is used. For a still broader categorization, the language impairment is called severe for stanine scores 1-3, medium for 4-5, mild for 6-7, and minimal/not impaired for 8-9. These severity categories are also given as different shades of grey in the *T* score profile sheet of the AAT (Figure 1).

The lowest raw score equivalent to a stanine score of 8 is in very close agreement with the cut-off score between the aphasic and the nonaphasic control group obtained in nonparametric discriminant analyses for each subtest separately. For each patient assigned to one syndrome group by the ALLOC-procedure, a second impairment grade (severe, medium, or mild) is available. It depends on whether the subtest raw score is below the 33th centile, below the 67th centile or above the 67th centile of the syndrome group's raw score distribution.

BASIC FEATURES OF HUBER'S APPROACH

The mathematical basis for Huber's approach to psychometric single-case analysis is given by the classical test theory model as explicated in Lord and Novick (1968, pp. 30). In order to draw statistical inferences about a single subject's test performance, one needs an estimate of the individual error variance (i.e., variability of test performances in hypothetical replications of assessments with the same test). For a homogeneous population of subjects, one can assume that these individual error variances are approximately equal. In that case, the square of the standard error of measurement, which is computed from the test score variance and the test reliability, can be taken as a good approximation of any subject's error variance within the population of interest.

If one wants to compare a subject's subtest scores from a test profile, this only makes sense if raw scores are converted into standard scores. Using the same standardization for all subtests provides comparability of observed scores. This common approach used for almost all available tests has one drawback: Diagnostic hypotheses about an individual's abilities aim at the identity of true scores and not of observed scores. It can be shown that identity of raw true scores does not imply identity of standardized true scores (cf. Huber, chapter 4.4) unless the

reliability parameters of the subtests are identical. Huber suggests τ-standardization (τ indicating true scores) which yields comparability of standardized true scores in all cases. Conversion of a standardized observed score y reported in the norm tables of a test to a τ-standardized score y^{τ} is accomplished by the following formula using the reliability parameter ρ and mean L of the particular standardization used:

$$y^{\tau} = y/\sqrt{\rho} + L(1 - 1/\sqrt{\rho}).$$

If the test reliability is high (i.e., ρ is close to 1), y and y^{τ} do not differ much.

For the subsequent test statistics to be distributed as normal or chi square (random) variables one has to assume independent and normally distributed errors and, sometimes (for tests of diagnostic validity), multidimensional normality of true scores (Huber, 1973; Willmes, 1984).

ANALYSIS OF AN INDIVIDUAL AAT *T* SCORE PROFILE

In the following, the AAT test results of an aphasic patient H.C. are used to demonstrate the application of the single-case analysis estimation and test procedures. Instead of providing a formal derivation of the formulas (Huber, 1973; Willmes, 1984), the respective parts of the computer outputs of CASE1 and CASE12 for H.C. given in Figures 2 to 6 will be explained and discussed.

Case History

H.C., a 56-year old clerk, was admitted to the Neurology department in Aachen on March 10, 1982. On admission there was an aphasia plus a somatosensory impairment of the right arm and leg. No hemiplegia or hemianopia were present. A CT-scan examination on the same day showed a not clearly demarcated hypodense lesion in the territory of the posterior temporal artery.

Because of the patient's good general condition, the AAT could be administered 5 days post onset (first examination). After a 14 days' stay on the ward, during which H.C. received 1 hour of speech therapy every day, the patient was discharged and outpatient speech therapy was continued three times per week, in sessions of 1 hour each.

From the beginning, speech therapy was comprised of structural language training for different language modalities using the better preserved oral reading capacity of the patient. The number of severe semantic and phonemic errors (semantic jargon) in spontaneous speech, repetition, and confrontation naming were reduced to a substantial extent. From May to the end of June, the speech therapy continued in a rehabilitation center, including at this time training of comprehension.

After leaving the rehabilitation center, H.C.'s language was still substantially impaired and control examinations with CT-scan and AAT were done at the Neurology department in Aachen. This time, CT-evaluation revealed a sharply demarcated lesion in the whole superior temporal gyrus as well as small multiple infarcts in the basal ganglia and the posterior white matter.

In Table 2, spontaneous speech ratings, raw scores, centile ranks, stanine scores, and the level of severity assignment compared to the whole aphasic group are listed for both test administrations. The severity grading takes into account 90% confidence intervals for raw scores. Furthermore, the ALLOC classifications, including posterior probabilities, are given.

Although the posterior probability for Wernicke's aphasia falls just short of 70% for the second examination, the neurolinguistic diagnosis is that of a well-recovered Wernicke's aphasia. Qualitative analysis of errors also shows that sentence-semantic and paragrammatical disorders in spontaneous speech, in repetition of sentences (part 5 of the Repetition subtest), and in part 4 of the Confrontation Naming subtest, in which one sentence is

Table 2

AAT Test Results of Patient H.C. for Both Test Administrations, Including Standardized Scores, Overall Severity Grading, Psychometric Syndrome Classification with Posterior Classification Probabilities (ALLOC-Procedure), Syndrome-Specific Severity Grading

	1. EXAMINATION (15.03.1982)				2. EXAMINATION (24.06.1982)			
	Raw Score	Centile Rank	Stanine Score	Severity Grading	Raw Score	Centile Rank	Stanine Score	Severity Grading
TT*	40	28	4	severe - medium	10	81	7	mild - minimal
REP	106	47	5	medium	133	76	6	mild
WRIT	63	65	5	medium - mild	85	95	8	mild - minimal
NAME	38	31	3	severe - medium	95	72	6	medium - mild
COMP	46	35	4	severe	92	70	6	medium - mild
SPONTANEOUS SPEECH RATINGS	1 5 3 1 2 3				3 5 4 3 4 3			
	COM ART AUT SEM PHO SYN				COM ART AUT SEM PHO SYN			
ALLOC Classification			Syndrome Severity Grading		ALLOC Classification		Syndrome Severity Grading	
Aphasia : 100%	TT : severe				Aphasia : 100.0%	TT : mild		
	REP : medium					REP : mild		
Wernicke's : 100%	WRIT : mild				*Wernicke's* : 66.1%	WRIT : mild		
Amnesic : 0%	NAME : severe-medium				Amnesic : 33.4%	NAME : mild		
	COMP : severe					COMP : mild		

* Error Score

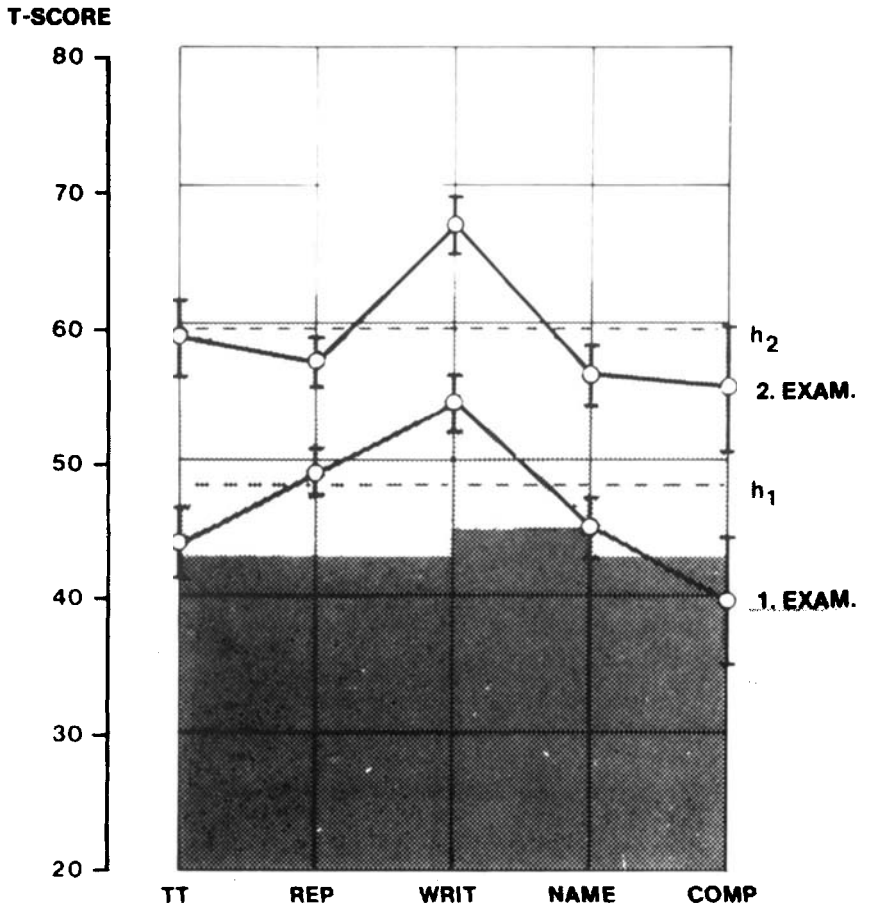

Figure 1. τ-standardized T scores with 90%-confidence intervals for both AAT examinations of patient H.C., and the estimated profile levels, h_1 and h_2 (dashed lines).

required as response, prevailed. Additionally, the severity of naming and comprehension difficulties is very similar. The syndrome-specific level of severity also given in Table 2 is a mild one for all subtests.

For the psychometric single-case analyses, only the T scores for the five subtests of both test administrations are needed (Figure 1). Because the computer programs used are general-purpose programs, instead of AAT subtest labels, only numbers 1 (Token Test), 2 (Repetition), 3 (Written Language), 4 (Confrontation Naming), 5 (Comprehension) are given in the printouts.

Standard Scores and Confidence Intervals

Using the formula for τ standardization given above, first the (point) estimates of the T-standardized true scores are given (Figure 2). Because of the high reliability estimates (Table 1), the estimates are very close to the T scores; only for comprehension is there a somewhat larger difference. For a diagnostic judgement concerning the subject's ability level, the 90%-confidence intervals covering the τ-standardized true score with 90% probability have to be considered. This leads, for example, to the diagnosis of a severe to medium naming disorder, the observed score itself indicating a severe disturbance. In general, the gradation aids reported in test manuals for scores standardized in the usual way (e.g., "below average", "average", etc.) are also valid for τ-standardized scores.

Global Profile Characteristics

The profile level is used as a global index of the overall level of performance. It is computed as a weighted sum of τ-standardized subtest scores, with the weights summing to unity and being functions of the reliability coefficients of the subtests. The more reliable a subtest is, the more weight the actual score obtained in that subtest gets. If the reliabilities of a test battery are identical, the profile level is, simply, the arithmetic mean of the subtest scores.

For H.C., the profile level estimate is $h = 48.08$ (Figure 2) as compared to an arithmetic mean of 46.31. Again, a 90%-confidence interval must be computed for the τ-standardized true profile level. For an aphasia test like the AAT, which is composed of subtests for various language modalities, reporting a profile level as an overall measure of the language disorder is not very useful. Although aphasia is usually viewed as being multimodal, there is no reason to assume that all modalities must be impaired to the same degree. The scatter of the individual profile is thus an indicator of "real" differences between subtest performances, i.e., differences which cannot be accounted for by measurement errors alone. In order to test whether numerical score differences are beyond chance, one computes the sum of squared deviations of subtest scores from the profile level. These are again weighted with a function of the respective subtest reliability (i.e., the inverse of the square of the subtest's standard error of measurement for τ-standardized scores). To prevent too large a Type II error, the profile is judged to be real if the value of the test statistic is above the 90%-quantile of the χ^2 distribution, with the number of degrees of freedom being one less than the number of subtests. For H.C., the first profile is clearly real.

A real profile calls for a more detailed analysis of the shape of the individual profile to reveal differences between pairs or groups of subtests. If no diagnostic hypotheses are specified in advance, systematic (post hoc) comparisons are useful. For the AAT, all pairwise comparisons of subtests are of interest. Planned comparisons are performed if one or several diagnostically relevant contrasts of subtest groups, specified before the global, overall analysis of the individual profile, are of interest. For example, for an intelligence test like the WAIS, one linear contrast of verbal versus non-verbal subtests might be useful.

```
**********************************************************
*                                                        *
*              STANDARD SCORES & CONFIDENCE INTERVALS    *
*                                                        *
**********************************************************

 SUB-  STANDARD  TAU-STANDARD.   90.0 %-CONF   INTERVAL
 TEST   SCORE        SCORE         LOWER   -    UPPER

   1      44         43.73         41.47   -    46.40

   2      49         48.99         47.26   -    50.73

   3      54         54.03         52.00   -    56.06

   4      45         44.96         42.79   -    47.12

   5      40         39.62         35.04   -    44.20

**********************************************************
*                                                        *
*                    GLOBAL PROFILE CHARACTERISTICS      *
*                                                        *
**********************************************************

     (1) PROFILE LEVEL:

      LEVEL ESTIMATE              =     48.08

     (2) 90.0%-CONF.-INTERVAL:

      LOWER LIMIT                 =     47.08
      UPPER LIMIT                 =     49.08

     (3) TEST FOR REAL PROFILE:

      TEST STATISTIC              =     46.51
      90.0%-CHISQUARE QUANTILE    =      7.78   DF=  4

      DECISION  THE PROFILE IS REAL.
```

Figure 2. Results of single-case analysis procedures for the first AAT examination of patient H.C. (output of program CASE1):
Upper part: confidence intervals for τ-standardized T scores.
Lower part: analysis of global profile characteristics.
(The AAT subtests are numbered in the following way: 1 Token Test, 2 Repetition, 3 Written Language, 4 Confrontation Naming, 5 Comprehension)

All Pairwise Comparisons of Subtest Scores

The results are given in Figure 3. Each pairwise difference has to be divided by the standard deviation of that difference. For the resulting standard normal z value, the two-sided p value is computed. In order to decide which of the 10 differences are significant for an overall Type I error of 10%, one of the following two multiple-test procedures can be adopted. With the Bonferroni approach, each individual p value from a pairwise contrast is compared to a Type I error of 10%, divided by the total number of pairwise comparisons, which equals 1% for the AAT. In the computer printout, a "1"

```
**********************************************************************
*                                                                    *
*              ALL PAIRWISE COMPARISONS OF SUBTEST SCORES             *
*                                                                    *
**********************************************************************
```

SUBTESTS	DIFF	T-VALUE	RELIABILITY ASPECT BONFERRONI P-VALUE (IN %)	BONFERRONI DECISION (TWO-SIDED)	HOLM RANK	HOLM DECISION	VALIDITY ASPECT P-VALUE (IN %, ONE-SIDED)
1 - 2	-5.06	-2.76	0.5769	1	5	1	25.7788
1 - 3	-10.10	-5.20	0.0000	1	10	1	4.2659
1 - 4	-1.02	-0.51	60.7682	0	1	0	42.9944
1 - 5	4.31	1.36	17.2719	0	2	0	26.6637
2 - 3	-5.04	-3.10	0.1920	1	6	1	22.4838
2 - 4	4.04	2.40	1.6615	0	4	1	25.8828
2 - 5	9.38	3.15	0.1645	1	7	1	12.2504
3 - 4	9.07	5.03	0.0001	1	9	1	5.6913
3 - 5	14.41	4.73	0.0002	1	8	1	1.6711
4 - 5	5.34	1.73	8.3123	0	3	0	20.0205

```
*** DECISION FOR OVERALL TYPE-1 ERROR =10.0 % :
*** (1: SIGNIFICANT; 0: NOT SIGNIFICANT)
```

```
**********************************************************************
*                                                                    *
*                      PLANNED LINEAR CONTRASTS                      *
*                                                                    *
**********************************************************************
```

THE SUBTESTS COMPARED IN THE LINEAR CONTRASTS ARE

CONTRAST	SUBTEST - GROUPS
1	3 - 1 2 4 5
2	1 5 - 2 3 4

TEST RESULTS

CONTRAST	CONTRAST-VALUE	RELIABILITY ASPECT Z-VALUE	P-VALUE (IN %)	DECISION (ONE-SIDED)	DIAGNOSTIC VALIDITY ASPECT Z-VALUE	P-VALUE (IN %)	DECISION (ONE-SIDED)
1	7.86	5.54	0.0000	1	1.63	5.1711	1
2	-6.49	-4.36	0.0007	1	-1.28	9.9777	1

```
*** DECISION FOR OVERALL TYPE-1 ERROR
*** RELIABILITY ASPECT =10.0 %  DIAGNOSTIC VALIDITY =20.0 %
*** (1: SIGNIFICANT; 0: NOT SIGNIFICANT)
```

Figure 3. Results of single-case analysis procedures for the first AAT examination of patient H.C. (output of program CASE1):
Upper part: all pairwise comparisons,
Lower part: planned linear contrasts, both for the reliability aspect and the diagnostic validity aspect.

indicates that the scores of two subtests are judged to be significantly different: that is, the difference in performance cannot be attributed to measurement errors alone but indicates a reliable difference in level of aphasic language disturbance.

Holm (1979) provides a more powerful procedure. The p values for the 10 pairwise comparisons are ordered with the lowest p value obtaining the highest rank. The test procedure is a sequentially rejective one. If the product of the lowest p value and rank 10 is below the overall Type I error, the two subtests are reliably different and the product of the second lowest p value and rank 9 is again compared to the 10% level. If the value of the product is below 10%, the two respective subtests are also judged to be reliably different, etc. The procedure stops as soon as the product of a p value and its rank is above the overall Type I error level.

From the printout in Figure 3, one can see that one more pairwise comparison (repetition vs. confrontation naming) is judged to be significant if Holm's procedure is applied. Arranging subtests according to increased T' values, and underlining subtests which are not significantly different, is a good means of visualizing the test results:

COMP TT NAME REP WRIT.

Written Language abilities are obviously better preserved in this patient, especially reading aloud which is at the 78th centile. Putting together words and sentences (centile rank = 59) and writing to dictation (centile rank = 61) are above average. Auditory and reading comprehension, however, are both equally low.

The better preserved performance in the subtest Written Language can be investigated further for its diagnostic validity. To demonstrate that written language is better preserved, the diagnostic validity of all four pairwise comparisons with the other four subtests must be computed. The probabilities (p values, one-sided) for obtaining still larger pairwise T-score differences are given in the right-hand side column of Figure 3. Applying Holm's procedure for the four comparisons of interest at an overall Type I error of 20% (as suggested by Huber, 1973, p. 116), only the contrast of written language and repetition performance is not diagnostically valid. Such a high Type I error is recommended because diagnostic decisions usually are different from decisions in experimental psychology. In experimental psychology it is widely accepted that one should be hesitant to reject a null hypothesis. That is, the Type I error is usually kept low. With diagnostic decisions, one mostly tries not to overlook "real" symptoms. Symptoms in aphasia testing are often related to differences in true level of performance among subtests. This implies using a rather liberal Type I error level of 10% or even 20% in order to prevent large Type II errors. If for some particular diagnostic problem a large Type I error is critical, a smaller Type I error level could be used. One only has to change the relevant parameter values in the computer programs CASE1 and CASE12.

Planned Linear Contrasts

For an aphasia test having different subtests that assess different language modalities, it is not the most adequate procedure to compare two subgroups of subtests in order to delineate overall profile differences (e.g., to detect particularly well preserved or strongly impaired language modalities). Thus, only for demonstration purposes, written language is compared to the subprofile composed of the other four subtests.

The subprofile level is 46.17, leading to a linear contrast value of 7.86, which indicates a reliable difference of performance (cf. Figure 3, p value = .0000). More interestingly, the diagnostic validity p value = 5.17% is also well below the required level of 20%.

Another reasonable planned linear contrast might be to compare receptive and expressive subtest performances (contrast 2 in the printout). Expressive abilities are significantly better preserved than receptive ones: the diagnostic validity p value of 9.98% is just below 10%, which is the Type I error level per comparison if the Bonferroni procedure is applied to a total of two planned comparisons.

Analysis of the Second Test Profile

The results are given without details. The profile level has risen to 59.55 and the profile is again real. Pairwise comparisons with the p values assessed by Holm's procedure resulted in the following configuration of subtests:

COMP	NAME	REP	TT	WRIT.

All pairwise comparisons of the Written Language subtest and the remaining four subtests are diagnostically valid as well. Similarly, the same planned linear contrasts as used for the first examination reveal better preserved written language ability that is diagnostically valid. However, there are no reliable differences between receptive and expressive modalities.

COMPARISON OF TWO AAT PROFILES OF THE SAME SUBJECT

The procedure used in the following can be applied if either the same subtests or parallel forms of the subtests are compared. The two test administrations may, for example, be before and after some treatment or may follow each of two different treatments. For H.C., AAT performances before and after about 3 months of speech therapy are compared.

Global Profile Comparisons

For a test of overall *profile identity* the sum of (weighted) squared differences of τ-standardized subtest scores is considered (Figure 4). The resulting value of the test statistic (= 187.77) is well above the 90%-quantile of the chi square distribution with 5 d.f. Thus, the two AAT profiles are judged to be significantly different. The two profiles may be nonidentical because of differences in profile levels and/or shapes, each of which can be assessed separately. The difference in *profile level* of 11.45 is highly significant. A one-sided test was performed because there is no reason to assume increase of impairment for an aphasic patient with vascular etiology who also received intensive speech therapy and has a short duration of aphasia. Before testing for *identity of profile shapes*, the two profiles have to be adjusted for differences in level of performance. After subtracting the respective profile level estimate from each subtest score, the weighted sum of squared

(adjusted) subtest scores is again computed. This test statistic value of 10.31 is still above the 90% quantile of the chi square distribution with 4 d.f.

This overall difference in profile shape requires a more detailed analysis of differential changes in relations of subtest performances between the first and second AAT examination. For an aphasia test consisting of one subtest per language modality, again, all 10 pairs of subtests should be analyzed. If diagnostic hypotheses concerning more general changes in relations of subgroups of subtests are of interest, they can be tested as well.

All Comparisons of Pairwise Profile Differences Between Subtest Scores

The difference of each two subtest differences (Figure 5) is divided by the square root of its variance. The resulting standard normal z values are subjected to either

```
**********************************************************
*                                                        *
*                          GLOBAL PROFILE COMPARISONS    *
*                                                        *
**********************************************************

     (1) TEST OF PROFILE IDENTITY:

      TEST STATISTIC            =    187.77
      90.0%-CHISQUARE QUANTILE  =      9.24  DF =  5

      DECISION: THE PROFILES ARE DIFFERENT.

     (2) TEST OF IDENTICAL PROFILE LEVELS:

      PROFILE LEVEL DIFFERENCE  =     11.45
      (2ND EXAM.-1ST EXAM)
      TEST STATISTIC (Z-VALUE)  =     13.32

      ONE-SIDED TEST:
      P-VALUE =   0.0000  (TYPE-1 ERROR = 10.0%)

      DECISION: SIGNIFICANT CHANGE IN PROFILE LEVEL.

     (3) TEST OF IDENTICAL PROFILE SHAPES:

      TEST STATISTIC            =     10.31
      90.0%-CHISQUARE QUANTILE  =      7.78  DF =  4

      DECISION: THE PROFILE SHAPES ARE DIFFERENT.
```

Figure 4. Results of single-case analysis procedures for the comparison of both AAT examinations of patient H.C. (output of program CASE12): global profile comparisons.

ALL COMPARISONS OF PAIRWISE DIFFERENCES BETWEEN SUBTEST SCORES

SUB-TESTS	DIFF 1ST EXAM	DIFF. 2ND EXAM	DIFF 1ST-2ND	Z-VALUE	RELIABILITY ASPECT P-VALUE (IN %)	BONFERRONI DECISION TWO-SIDED)	HOLM RANK	HOLM DECISION
1 - 2	-5.06	2.06	-7.12	-2.75	0.6010	1	10	1
1 - 3	-10.10	-8.03	-2.07	-0.75	45.1263	0	3	0
1 - 4	-1.02	3.05	-4.07	-1.44	14.8779	0	7	0
1 - 5	4.31	3.91	0.40	0.09	92.8180	0	1	0
2 - 3	-5.04	-10.09	5.05	2.20	2.7692	0	9	0
2 - 4	4.04	0.99	3.05	1.28	20.0711	0	6	0
2 - 5	9.38	1.85	7.53	1.79	7.3951	0	8	0
3 - 4	9.07	11.08	-2.00	-0.79	43.2031	0	4	0
3 - 5	14.41	11.94	2.47	0.57	56.6069	0	2	0
4 - 5	5.34	0.86	4.48	1.03	30.4145	0	5	0

*** DECISION FOR OVERALL TYPE-1 ERROR =10.0 %
*** (1: SIGNIFICANT; 0: NOT SIGNIFICANT)

PLANNED PROFILE COMPARISONS

THE SUBTESTS COMPARED IN THE PROFILE COMPARISONS ARE

COMPARISON	SUBTEST - GROUPS
1	3 - 1 2 4 5
2	1 5 - 2 3 4

TEST RESULTS :

PROFILE COMPARISON	CONTRAST-VALUE 1ST EXAM	2ND EXAM	DIFFERENCE (1ST - 2ND)	RELIABILITY ASPECT Z-VALUE	P-VALUE (IN %	DECISION TWO-SIDED)
1	7.86	10.04	-2.18	-1.09	27.7634	0
2	-6.49	-1.65	-4.83	-2.30	2.1609	1

*** DECISION FOR OVERALL TYPE-1 ERROR =10.0 %
*** (1: SIGNIFICANT; 0: NOT SIGNIFICANT)

Figure 5. Results of single-case analysis procedures for the comparison of both AAT examinations of patient H.C. (output of program CASE12):
Upper part: all comparisons of pairwise subtest comparisons.
Lower part: planned linear profile contrasts.

Bonferroni's or Holm's multiple test procedure by examining the two-sided p values related to the z values. For H.C., only the relation between the Token Test and the repetition subtest has changed significantly (reliably) for both test strategies. Whereas the T' score for the Token Test is numerically lower than the one for repetition at the first examination, the opposite relation is found at the second test administration.

Planned Profile Comparisons

The same linear contrast of subgroups is computed for both examinations and the difference of both values is tested to determine if it is significantly different from zero. The rather few changes in the relation of AAT performances for H.C. detected by using pairwise profile comparisons is also reflected in the following result. The profile comparison for the relation of written language performance to the linear combination of the other four subtests is $7.86–10.04 = -2.17$. Thus, it is not significantly different from zero, although the superiority of the written language ability has become more pronounced numerically. The z value of the related test statistic results in a two-sided p value of 27.76%. If one is also interested in a possible change of the relationship between receptive and expressive tasks, one obtains a significant result (cf. the second profile comparison in Figure 5). The distance between receptive and expressive impairment has diminished from a contrast value of -6.49 to a value of -1.65. The two-sided p value of 2.16% indicates that a still larger amount of change is very rare for the reference population of all aphasics.

Comparison of Subtest Performances

If one is not interested in detailed profile comparisons after two profiles were found to be nonidentical, one can also compare performances for the five subtests separately and apply either Bonferroni's or Holm's multiple test procedure. As Figure 6 shows, the z values of the test statistics give rise to very small one-sided p values. Each subtest performance has improved significantly.

If one compares two AAT profiles before and after a period of speech therapy, significant improvement is not necessarily due to the therapeutic treatment alone. Especially if both test administrations are within the first 6 months post onset, spontaneous recovery may have occurred as well. The amount of spontaneous recovery is difficult to estimate, especially for a single patient.

Data on the spontaneous recovery of groups of aphasic patients having received no formal speech therapy are available for the AAT (Willmes & Poeck, 1984). CVA-patients (21 global, 19 Wernicke's, 12 Broca's, 32 amnesic, 12 not classifiable aphasic patients) had been tested 1, 4, and 7 months post onset in that study of spontaneous recovery. The two test administrations for patient H.C. correspond quite well to the first two test occasions of the recovery study. One could use the mean change in T scores for each subtest and the profile level as the best (point) estimate available for the influence of spontaneous recovery. After subtracting

COMPARISON OF PERFORMANCES PER SUBTEST

SUB-TEST	STANDARD SCORES 1ST EXAM	2ND EXAM	TEST OF DIFFERENCES(2ND-1ST) TAU-DIFF	Z-VALUE	P-VALUE(%) *** (ONE-SIDED, INCREASE) ***	DECISION BONF	HOLM(RANK)
1	44	59	15 17	7 15	0 0000	1	1 (4)
2	49	57	8 04	5 39	0 0000	1	1 (2)
3	54	67	13 10	7 51	0 0000	1	1 (5)
4	45	56	11 09	5 97	0 0000	1	1 (3)
5	40	55	15 57	3 95	0 0039	1	1 (1)

*** DECISION FOR OVERALL TYPE-1 ERROR =10 0 %
*** (1 SIGNIFICANT, 0 NOT SIGNIFICANT)

Figure 6. Results of single-case analysis procedures for the comparison of both AAT examinations of patient H.C. (output of program CASE12): All comparisons of subtest performances.

these recovery estimates from the observed differences of τ-standardized T scores the residual differences can be tested for reliable improvement with the same test statistics as for the uncorrected differences. This part of the analysis is not included in the general computer programs and the results are reported without computational details. If one applies Holm's procedure, all residual improvements are still significant at an overall Type I error level of 10%. The residual increase in profile level of 4.40 is also highly significant.

Summing up the psychometric analyses for patient H.C., a detailed account of the considerable improvement could be given. Relative superiority of written language abilities did not decrease. This might also be due to the therapeutic technique of using a better preserved modality to correct mistakes made in other language modalities. It is worth noting that improvement is especially large for the Token Test. There was also substantial improvement in several spontaneous speech ratings, especially for communicative ability and semantic and phonemic structure (Table 2). Continuation of speech therapy was therefore highly recommended; in addition, the decision to allow the patient go on retirement was postponed.

DISCUSSION

Huber's concept of psychometric single-case analysis offers a unified approach to the analysis of profile scores of a single subject obtained in any standardized test, not only in an aphasia test. In principle, the procedures can be applied to all tests

constructed and standardized along the lines of the classical test theory model; this is the case for the great majority of assessment procedures available. Once it has been determined that, for a particular test, Huber's methods are applicable, the (profile-)scores of each subject tested can be analyzed. However, the requirements of practical invariance of reliability and correlation estimates (see Appendix) and for standardization samples with at least about 400 subjects may pose difficulties for many tests.

The introduction of τ standardization helps to clarify the diagnostic hypotheses that a test user wishes to test. Only inferences related to true (profile-) scores meet these substantial diagnostic problems. A τ standardization is always possible if standard norms from large samples are reported for a test. However, if reliability estimates are high, the reported standard scores and τ-standardized scores are usually close together.

Huber's approach attempts to follow Zubin's postulates (1950) for the statistical analysis of intraindividual series of observations. These postulates state that (i) each individual must be assumed to be an independent universe characterized by (ii) a given level and (iii) a given degree of variability of performance, both of which characterize the hypothetical distribution of potential scores of the subject and (iv) both of which can be changed by some sort of exogeneous factors (treatments) or endogenous changes. The postulate of a characteristic degree of variability of performance is the most crucial one for Huber's approach. Because there is no realistic way of obtaining a good estimate of the error variance of a subject in a specific test by repeated administrations of the same test or several parallel forms of it, the standard error of measurement is used.

A further crucial assumption is that the applications of two subtests of a profile or of parallel tests or two replications of one test form have uncorrelated errors. However, there is evidence (Rozeboom, 1966) that errors can be positively correlated because of illness, fatigue, noise, cheating, etc. Zimmerman and Williams (1977) and Williams and Zimmerman (1977), in deriving the classical test theory model without the assumption of uncorrelated errors, demonstrated that, if the reliability parameters of two tests are high ($> .90$) and the correlation not much below .50, even rather substantial positive correlations between scores do not affect the true correlation between the two tests to a large extent. But large effects are possible under different circumstances. They also demonstrated that the reliability of a difference score which, according to the classical test theory model can be more fallible than both tests themselves, must not be lower. Again, if errors as well as observed scores are positively correlated ($>.50$), and if reliability parameters are high, the reliability of the difference score may even be considerably higher than the reliability of both tests. Put another way, the standard error of measurement for a difference score (or any linear contrast) may be smaller than for the individual scores so that Huber's test procedures may even be conservative for well-constructed test batteries.

In any case, the choice of an adequate reliability coefficient is an important

substantial issue. Although all reliability coefficients can, in principle, be used to estimate the standard error of measurement, the diagnostic problem at hand should guide the choice. If inferences about differences between two parallel tests are of interest, reliability estimates for parallel tests should be used. When testing for score differences within a profile, split-half or consistency coefficients should be preferred. Furthermore, reliability estimates for subpopulations and/or specific age groups should be used although, particularly for clinical subpopulations, they are often not available or are based on very few subjects.

For clinical populations like the aphasic population, there are additional problems. Because raw score distributions often are highly skewed, quasi-normalization by computing T scores from the centile rank distribution may distort the extremes of the scale to some extent. Thus, analyses for patients with very poor or very good performances should be interpreted with some reservation, especially because skewness may be different though high for different subtests contained in the test profile. It may, for example, be impossible to detect a reliable change in performance if the initial score is already quite high. This would also bring into question the model assumption of a constant standard error of measurement over the whole scale of a subtest. The distribution of potential scores of a subject must have a smaller spread at the extremes of a scale. If a clinical population is composed of several subpopulations (e.g., syndrome groups), the normality assumption for true scores is also questionable. The distribution might be bi- or polimodal or may run much flatter than the normal curve. So it would definitely be useful to have normative data for each syndrome group separately. However, for an aphasia test, this will take several more years and the availability of a powerful classification procedure.

The constructional principles of an aphasia test also have consequences for the strategy with which a profile of scores is evaluated. Only if several subtests are assessing the same or a very similar ability is computation of the profile level more than a technical step within the profile analysis procedure. For an aphasia test in which each subtest represents a different modality, pairwise comparisons between subtests should be preferred. With linear contrasts containing several subtests, compensatory effects of some subtest(s) can otherwise disguise substantial differences between some pair of subtests.

Although aphasia is usually taken to be a multimodal deficit, this does not mean that the level of impairment is (almost) identical for all modalities. Consequently, "real" profiles are the rule rather than the exception. Diagnostically valid patterns, on the contrary, are not that frequent. Probably the most interesting kind of profile pattern for an aphasic patient is that of a unimodal deficit or peak in performance. This theoretically based expectation calls for testing the diagnostic validity of the pairwise difference between a certain subtest (modality) and all (several) of the other subtests (modalities). If several subtests for one modality are available in a test, planned comparisons should contain pairwise contrasts of two subgroups of tests, each of the two subgroups assessing one language modality.

The procedures for the detection of a unimodal problem could also help to provide an operational definition for nonstandard aphasias such as transcortical or conduction aphasia. In addition to giving a definition confined to centile rank differences, as do De Bleser, Huber, Willmes, and Blunk (1981) and Reinvang and Engvik (1980), one could also define such an aphasia as follows: "A patient has a transcortical aphasia if his repetition performance, compared to the reference population of aphasic patients, is significantly above his performance on each of the following subtests: Written Language, Confrontation Naming, and Token Test (aspect of diagnostic validity). The repetition score should be at least at the 60th centile rank". For a conduction aphasia, repetition performance should be no better than a T score of 50 and be significantly below all other subtests considering the aspect of diagnostic validity. Also, the patient should have no severe dysarthria and/or speech apraxia.

With the aid of the two computer programs, CASE1 and CASE12, a complete analysis of two AAT test profiles requires only a negligible amount of work. Thus, obtaining detailed diagnostic information on the relative impairment of different language modalities is not restricted to research purposes encompassing only some patients. Rather, it can be used routinely. As has been shown, application of the single-case analysis procedures is also not restricted to diagnostic problems; it can also guide the choice and combination of language therapeutic measures.

APPENDIX

Requirements for the applicability of Huber's procedures

In the derivations of the different test statistics needed for the single-case assessment procedures it is always assumed that the population parameters are known constants. But, for each psychological test, only estimates of them exist. Consequently, one has to assure that sampling errors are negligible in practical applications. The mean and standard deviation of the raw scores are needed for establishing norm scores. The sampling error of both is sufficiently small (cf. Huber's chapters 4.8-4.10) only if the standardization sample size is larger than about $n = 400$ and if the reliability *parameter* is above 0.60. But the reliability parameter itself has to be estimated. Again, it can be demonstrated that a sample size of about 400 is just large enough to yield sufficiently precise reliability estimates. But, for almost all known tests, the reliability studies are based on far less than 400 subjects. Huber suggests that one accept a reliability estimate in case of n less than 400 only if it is *practically invariant*: that is, if the lengh of the 95%-confidence interval of the (true) reliability parameter, ρ, is below 0.1. To examine this property, one first subjects the reliability parameter estimate $\hat{\rho}$, reported for some test, to Fisher's z' transformation, as follows:

$$z' = (\log(1 + \hat{\rho}) - \log(1 - \hat{\rho}))/2;$$

where *log* indicates the natural logarithm.

In terms of the z' scale, the lower z_l and upper z_u 95%-confidence interval boundaries are $z_{l,u} = z' \mp 1.96/\sqrt{(n-3)}$. These have to be transformed back to the original ρ scale again by using the following transformation:

$$\rho_{l,u} = (e^{2z_{l,u}} - 1)/(e^{2z_{l,u}} + 1)$$

In some formulas related to the diagnostic validity aspect, intercorrelations between subtests occur. If they are estimated from samples of less than 400 subjects, their 95%-confidence intervals have to be checked in the same way as the reliability coefficients.

For the AAT, the size of the standardization sample $N = 376$ is just sufficient to guarantee precise enough estimates of the expectation and the standard deviations of the subtests' raw scores. The intercorrelations between subtests are also computed from this sample and are thus practically invariant. Only the reliability estimates are computed from $n = 120$ aphasic patients of the validation sample of the AAT. The Comprehension subtest has the lowest coefficient ($\hat{\rho} = 0.928$). Applying the two formulas given above, one obtains $z_{l,u} = 1.644 \mp 0.1812$ and $\rho_l = .898$, $\rho_u = .949$. The difference between ρ_l and ρ_u clearly is less than .10. Due to the functional form of the above formulas, the other reliability coefficients need not be checked, because higher coefficients result in smaller confidence intervals.

One can also determine, for a given sample size n, what the minimum $\hat{\rho}$ must be in order to guarantee a confidence interval below 0.1. Table A below gives minimum values of $\hat{\rho}$ for several values of n.

Table A

Minimal Reliability Coefficient Values ρ_{min} Which are Practically Invariant for a Given Sample Size n for a Type I Error of 5%

n	ρ_{min}	n	ρ_{min}	n	ρ_{min}	n	ρ_{min}	n	ρ_{min}
10	.975	110	.861	210	.797	310	.745	410	.698
20	.953	120	.853	220	.792	320	.740	420	.693
30	.937	130	.846	230	.786	330	.735	430	.689
40	.924	140	.840	240	.781	340	.730	440	.684
50	.913	150	.833	250	.776	350	.726	450	.680
60	.902	160	.827	260	.770	360	.721	460	.675
70	.893	170	.821	270	.765	370	.716	470	.671
80	.884	180	.815	280	.760	380	.711	480	.666
90	.876	190	.809	290	.755	390	.707	490	.662
100	.868	200	.803	300	.750	400	.702	500	.658

REFERENCES

Basso, A., Capitani, E., & Zanobio, M. E. (1982). Pattern of recovery of oral and written expression and comprehension in aphasic patients. *Behavioral Brain Research, 6,* 115-128.

Benson, F. (1979). Editorial: Aphasia rehabilitation. *Archives of Neurology, 36,* 187-188.

De Bleser, R., Huber, W., Willmes, K., & Blunk, R. (1981, October). Transcortical aphasia. Paper presented at the 19th Annual Meeting of the Academy of Aphasia. London, Ont.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Habbema, J. D. F., Hermans, J., & van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In: G. Bruckmann (Ed.), *COMPSTAT 1974, proceedings in computational statistics,* Wien: Physica.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.

Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *Der Aachener Aphasie Test (AAT).* Göttingen: Hogrefe.

Huber, W., Poeck, K., & Willmes, K. (1984). The Aachen Aphasia Test (AAT). In F. C. Rose (Ed.), *Progress in aphasiology* (pp. 291-303). New York: Raven.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Orgass, B. (1976). Eine Revision des Token Tests. Teil I und II. *Diagnostica, 22,* 70-87 and 141-156.

Reinvang, I., & Engvik, H. (1980). Language recovery in aphasia from 3 to 6 months after stroke. In: M. T. Sarno & O. Höök (Eds.), *Aphasia: Assessment and treatment* (pp. 79-88). Uppsala: Almqvist & Wiksell.

Rozeboom, W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey Press.

Springer, L., & Weniger, D. (1980). Aphasietherapie aus logopädisch-linguistischer Sicht. In G. Böhme (Ed.), *Therapie der Sprach-, Sprech- und Stimmstörungen* (pp. 190-207). Stuttgart: G. Fischer.

Weniger, D. (1982). Therapie der Aphasien. In: K. Poeck (Ed.), *Klinische Neuropsychologie,* Stuttgart-New York: Thieme.

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement, 37,* 679-689.

Willmes, K. (1984). An approach to analyzing a single subject's scores obtained in a standardized test: Description of the method. Unpublished manuscript.

Willmes, K., & Poeck, K. (1984). Ergebnisse einer multizentrischen Untersuchung über die Spontanprognose von Aphasien vaskulärer Ätiologie. *Nervenarzt, 55,* 62-71.

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.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology, 16,* 135-152.

Zubin, J. (1950). Symposium on statistics for the clinician. *Journal of Clinical Psychology, 6,* 1-6.