The acquisition of verbal morphology in coclear-implanted and specific language impaired children

Hammer, A.

CHAPTER 4

Language assessment and research method

1. Introduction

The aim of this dissertation is to assess the development of verbal morphology in hearing- and language-impaired children. In this chapter we will focus on the methodological issues in language assessment in general and the methodology used in this dissertation in particular.

This chapter will start with a short outline of the language assessment procedures. As already pointed out by Masterson and Kamhi (1991): *'one extrinsic factor affecting the language knowledge attributed to a child is the way that information about language abilities is obtained'* (p:549). Language assessment procedures can be divided into two main categories: psychometric testing and spontaneous language sampling. The choice in favour of one assessment tool rather than the other is determined purely by the objectives of the researcher. We will elaborate on both assessment procedures and discuss the methodological issues of reliability and validity.

In this dissertation, we want to compare the scores of the hearing- and language-impaired children with a group of hearing peers with typical language development. We therefore use a norm-referenced standardized test. For Dutch, the only norm-referenced test that assesses morphosyntactic complexity and correctness is the STAP test. This test provides norms for children aged between 4 and 7 years (Verbeek, Van den Dungen & Baker, 1999). The STAP test will be discussed in more detail in section 3.

Prior to the implementation of the STAP test in our research, a small study was set up to explore whether this test satisfied psychometric criteria. This study is presented in section 4.

2. Language assessment

2.1 Objectives in language assessment

A psychometric test is defined as a behavioral measure in which a sample of behavior is obtained in a highly structured setting and under conditions in which the child is assumed to perform at his or her best (McCauley, 2001). As a language evaluation tool, psychometric testing is an efficient method to assess maximum language behavior, especially when the clinician or researcher is interested in a specific linguistic area. Comparing an individual score to a group score easily assesses language proficiency (Bacchini, Kuiken & Schoonen, 1995; Braam-Voeten, 1997). However, the language measures are obtained in a structured setting that deviate from daily speech. Therefore generalizability of the results is limited.

The goal of spontaneous language sampling is to obtain a language sample that maximally corresponds with the daily speech of the child and to offer insight into the full repertoire of syntactic structures that a child has at his or her disposal. This language evaluation tool does not follow a clear-cut format; samples vary in format from an unstructured setting in which a child plays with toys and the clinician/researcher only marginally stimulates the child to talk, to a more structured setting where the child is asked to tell a story using picture books. The effects of sampling method on language use have been addressed in several studies.

For instance, Southwood and Russell (2004) investigated the effects of three different sampling methods in 5-year-old typically developing African boys. The first method was a conversation between researcher and child structured by a questionnaire. The second method involved a play session with toys, called freeplay. The last sampling method was a story generation task in which a child was asked to tell the researcher something that happened to him/her. All samples were time-framed in 15 minutes. The last method elicited significantly less utterances as compared to the conversation and freeplay methods. However, the utterances produced were significantly longer. Both conversation and story generation tasks stimulated children to use complex syntactic structures, whereas freeplay did not. With respect to the variety of syntactic structures and errors, no differences between sampling methods were found.

Longer utterances in story generation tasks as compared to conversations were also found by Wagner, Nettelbladt, Sahlén and Nilholm (2000), who studied 28 SLI children aged between 4;11 and 5;9. They attributed the difference in utterance length to the occurrence of more elliptical answers in conversations as compared to story generation tasks. When questionnaires are used, children easily respond with a one or two-word answer. The inclusion of elliptical answers could also account for the results found by Southwood & Russell (2004), as they used questionnaires in their conversations.

In addition to the longer utterances in story generation tasks, Wagner et al. (2000) report a higher number of grammatical morphemes per utterance in this task as compared to conversations. Grammatical morphemes included the free-standing grammatical morphemes (articles, prepositions, auxiliaries) and the bound morphemes (verb inflections and plural).

With respect to the analysis of grammatical morphemes, a more sophisticated approach is to calculate the proportional use of a particular morpheme in an obligatory context. The method of spontaneous language sampling, irrespective of the format used, has a serious drawback in this respect. An adequate number of obligatory contexts need to be present in a child's sample in order to evidence the production (or non-production) of a particular morpheme (Lahey, Lievergott, Chesnick, Menyuk & Adams, 1992; Sealey & Gilmore, 2008).

The production of obligatory contexts for finite verbs and the accuracy of these finite verbs across four different sampling formats was taken up in the study of Sealey & Gilmore (2008). The first format was a freeplay session with minimal interference by the researcher, the second format was called 'storyboard' and involved a story-telling task using props. The child was given a model story, after which the he or she could tell his or her own story using the props. During the third sampling session the child was asked to retell a story that was first told by the researcher. In the fourth sampling format the child had to tell a story using a wordless picture book. No model story was given beforehand. All sampling sessions were time-framed in 15 minutes. Five SLI children and 5 TD children, aged between 3;11 and 5;6 participated in this study. Results showed that the number of obligatory contexts for finite verbs was the highest in freeplay sessions as compared to the other sampling formats.

However, interestingly, when samples are controlled for language production, rather than time, effects between sampling formats disappeared. No significant difference between sampling formats was found in the overall proportion of finite verbs to the total number of morphemes (lexical and grammatical). Moreover, when adult target-like use of finite verb forms was expressed as a proportion of the number of obligatory contexts, no significant sampling effects occurred. This suggests that not only should the method of sampling be carefully considered, but also how the samples are approached in language assessment.

2.2 Methodological concepts: reliability and validity

2.2.1 Defining reliability

The language evaluation tools of spontaneous language sampling and psychometric testing can be placed on a continuum from an unstructured setting to a relatively highly structured setting. The language samples lacking

any form of structure, such as freeplay, are located at the right-hand end of the continuum. The use of questionnaires and picture books in eliciting speech gives more structure to the language samples, causing a move to the left. The left-hand end of the continuum is taken up by the more experimental approach to language assessment, called psychometric testing.

The benefit of obtaining language measures in a structured setting is the likelihood of replicating findings when the same individual is tested on another occasion. Replication in test theory is an important methodological concept and is referred to as reliability. The term reliability is defined as the consistency between measures, i.e. the agreement in scores when a test is applied multiple times. In order to measure a linguistic aspect consistently, one needs to measure that aspect systematically. To clarify this it is helpful to express reliability mathematically as has been done in the Classical Test Theory.

The underlying idea of the Classical Test Theory is that the *observed score (X)* of an individual is composed of the individual's *true score (T)* and an *error score (E)*. The component of interest is the true score, which is a constant. A random variation, or error, is added to this constant. This variation occurs through factors related to the individual (e.g. fatigue, loss of attention, low motivation) and to the test situation (e.g. noisy environment, room is to warm). An individual's observed score can be expressed in the following way:

$$X=T+E$$

The Classical Test Theory makes two assumptions when the same individual repeats a test multiple times. The first assumption is that the mean of the errors is 0, i.e. positive and negative values level each other out. This means that the average of the observed score estimates the true score. The second assumption is that the random error is normally distributed around the true score. This is called the standard error of measurement. The fact that the error scores do not correlate with each other and the true score, means that the error term does not allow any systematic variance. In the case of one individual repeating the test multiple times, the standard deviation of the errors equals the standard deviation of the observed scores. From here it follows that the smaller the standard deviation of the errors, the more compactly the random errors are grouped around the true score.

Instead of giving one individual the same test over a hundred times, a sample of 100 different people can be given the test. The same assumptions apply for this sample as for the individual case. It is interesting to note here that the variance in the observed scores is the sum of the variance in true scores (i.e. not everybody has the same true score) and the variance in the random error.

$$VAR\ (X) = VAR\ (T) + VAR\ (E)$$

To calculate reliability, the variance in true scores has to be divided by the variance in the observed scores.

$$R=VAR(T)/VAR(X)$$

A test has perfect reliability when the variance in the true scores equals the variance in the observed scores (R=1). Poor reliability is obtained when the variance of the observed scores equals the variance in the error scores (R=0). This becomes clearer when the above formula is rewritten.

$$R= 1- (VAR(E)/ VAR(X))$$

The Standard Error of Measurement and the reliability are both important to estimate the accuracy of a measure. When the variance of the observed scores is kept constant, the formula below shows that with increasing reliability the variance in random error decreases.

$$VAR(E)= VAR(X) \sqrt{(1-R)}$$

Using the standard error of measurement a 95% confidence interval can be calculated. This is the range that reflects a 95% probability that it includes the true score of an individual and 5% that it does not. A small range indicates a higher accuracy of estimating a child's true score as compared to a large range.

The most straightforward way to calculate the reliability coefficient is to collect two spontaneous language samples from the same individual or to give an individual a psychometric test at two consecutive moments. The correlation coefficient between the two test moments can be taken as the reliability coefficient. This procedure is termed the test-retest method. This term is interchangeably used with stability, because it measures the stability of a test over a period of time (Van den Brink & Mellenberg, 1998; McCauley, 2001; Drenth & Sijtsma, 2006). The measurement of test stability also hints at a practical problem with the test-retest method, namely the determination of the time-interval between the two test moments. This interval should be large enough to optimize the independency of the two test scores and small enough to ensure stability of the matter to be measured within the subject.

In terms of the reliability of language assessment tools, the primary interest is not the stability of a test in assessing a child's language proficiency at consecutive intervals. On the contrary, it is believed that a language assessment tool should identify the language development experienced over consecutive intervals. This does not mean that reliability is of no importance in language assessment.

With regard to spontaneous language sampling, we want to prevent the language measures derived from one set of utterances differing from the measures obtained from another set of utterances, when both sets are taken from the same sample. This type of reliability is also called internal consistency reliability. An effective procedure for calculating this reliability coefficient is to split a test in half, the split-half method. The correlation coefficient between both halves can be considered as the reliability coefficient for half the test. We can correct the reliability coefficient for the complete test using the Spearman-Brown formula. In the formula below R_k stands for the reliability coefficient calculated for half the test, K refers to the number of parts in which a test is divided and R is the reliability coefficient for the complete test.

$$R= (KR_k)/ 1 + (K-1) R_k$$

The conclusions drawn from test reliability are related to the purpose of a study. For individual comparisons, a reliability of >.90 can be considered as acceptable. As the main purpose of this dissertation is to compare groups of atypical developing children with TD children, a reliability of >.70 can be regarded as a rule of thumb. It has to be kept in mind that the accuracy of measurements in group comparisons is also determined by group size (Drenth & Sijtsma, 2006).

2.2.2 Defining validity

The methodological concept of reliability is closely related to the concept of validity. Validity refers to the degree to which a test measures the intended objective for which the test was designed. The metaphor of archery can easily demonstrate the relation between both methodological concepts. If an archer hits the mark consistently, his aims are reliable and valid. A second possible situation is that the archer neither hits the mark nor shoots consistently, resulting in invalid and unreliable aiming. When an archer hits a target consistently but near the mark, his aim is reliable, though invalid (McCauley, 2001). From here it follows that a valid measure needs to be reliable, whereas the opposite, a reliable measure is valid, is not true.

Despite the clear definition of validity, many subtle changes have been made to this definition depending on the purpose of the test for which validity measures were taken. All subtypes can be placed under the umbrella term of construct validity. This umbrella distinguishes two broad categories. The first category is content validity, which *'involves the demonstration that a measure's content is consistent with the construct or constructs it is being used to measure'* (McCauley, 2001 p:56). With respect to language assessment, this entails that the language measures need to be independent of non-verbal skills, such as memory and auditory processing.

The second category is criterion-referenced validation, that *'refers to the accumulation of evidence that the measure being validated is related to another measure – a criterion – where the criterion is assumed to have been shown to be a valid indicator of the targeted construct'* (McCauley, 2001 p:61). One type of criterion-referenced validation is concurrent validity that measures how well test scores correspond with a criterion, which are both taken at the same moment. The criterion should be a 'gold standard', reflecting the 'true' measurement of the behavior under validation (McCauley, 2001 p: 218). For language assessment tools such a gold standard is not available and the alternative method of contrasting groups is used (Aram et al., 1993; McCauley, 2001). The contrasting groups method determines how well a test discriminates between subjects with and without the disorder, in which the groups are chosen with prior knowledge that they differ on the construct to which the test applies.

The discriminating abilities of a test are expressed in terms of sensitivity and specificity. Sensitivity indicates the percentage of SLI children who are correctly identified as such by a particular test (in Table 1: A/A+B). Specificity indicates the percentage of children with normal language who are also identified as such by the specific test (in Table 1: D/C+D) (Dunn, Flax, Sliwinski & Aram, 1996). The accepted level for sensitivity is 90% or above. For specificity, 70% is considered 'fair', and 80% is 'good' (Plante & Vance, 1995 as cited by McCauley, 2001).

Table 1. Outline of the sensitivity and specificity of a test to identify SLI children using the contrasting groups method.

Test outcomes → *Group ↓*	*SLI*	*Non-SLI*	
SLI	A	B	A+B
non-SLI	C	D	C+D

2.3 Methodological concepts in language assessment tools

The validation of language assessment tools is driven by the clinical need for consistent criteria in identifying children with SLI. In clinical practice we want to assess a child's language proficiency to draw conclusions about the age appropriateness of the child's language level. The language measures as well as the normative data provided by the test should therefore distinguish children with SLI from the children with age-appropriate language levels. However, when it comes to language assessment tools there is a false believe that below a

predefined cut-off point in a norm-referenced test a second population exists, which can be labeled language delayed (Gavin, Klee & Membrino, 1993). This second population is generally not statistically underpinned in test manuals. The best indicator of validity is therefore the demonstration that a language test accurately discriminates between language delayed and non-delayed children.

In diagnostic evaluation, psychometric language tests seem to under-identify children with language impairments. Plante and Vance (1994) validated four psychometric language tests that met a high number of psychometric criteria (e.g. description of the normative samples in test manuals, sample sizes, means and standard deviations). Forty 4 to 5-year-old children, equally divided in a group of SLI and TD children, were given the tests. Results revealed that a sensitivity percentage of 90% was reached for only one test measuring morphosyntactic production. The other tests all had sensitivity percentages lower than 80%, indicating that in less than 80% of the cases a congruence was found between the clinical diagnosis of SLI and the psychometric diagnoses of SLI. These results are compatible with the results of Aram et al. (1993) from their earlier large-scale study. From the 252 SLI children, only 20% to 70% were correctly identified as SLI using psychometric language tests. This implies that if the diagnosis of SLI is solely based on psychometric test outcomes, between 30% and 80% of the clinically diagnosed language delayed children will be misidentified.

Dunn et al. (1996) analyzed spontaneous language samples of SLI children, with the objective to extract language measures that adequately classify children with SLI. When base rate information was taken into account (i.e. correcting for unequal sample sizes of SLI and TD children included in the study) SLI was correctly predicted in 90.2% of the cases. This optimal classification was reached with a combination of the MLU, percent structural errors and age measures. Qualitative analysis revealed that the majority of structural errors involved morphological errors.

As pointed out by Bedore & Leonard (1998), one way of increasing accuracy in identifying SLI is to include a clinical marker, which is variable in the SLI population and stable in the children with typical language development. A large number of studies have indicated weaknesses in the area of morphology in SLI children, especially in their use of verbal morphology (e.g. Leonard et al., 1992; Conti-Ramsden & Jones, 1997; De Jong, 1999; Blake, Myszczyszyn & Jokel, 2004; Rice et al., 1995). In the study by Bedore & Leonard (1998) good sensitivity (>90%) was reached when SLI diagnosis was based on the production of regular past tense, 3rd person singular present inflections, copulas and the auxiliary *be* and MLU.

It thus seems that spontaneous language samples robustly discriminate between children with and without SLI, on condition that valid measures are included in the formula. However, in-depth qualitative analysis of the sample is essential in pinpointing the language difficulties of a particular child.

3. The STAP test

3.1 The STAP method

The STAP test requires conversational languages samples, in which topics of interest to the child are discussed. The test procedure consists of recording conversations between a child and an adult. The child is followed in his or her spontaneous speech as to limit the interference of the adult. The manual indicates that approximately 10 to 20 minutes of recording is needed to elicit sufficient speech to conduct analysis. Sufficient speech means that the transcribed recordings include minimally 50 utterances. The definition of an utterance is adopted from Hunt's T-unit: *one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it* (Hunt, 1970 p:4 as cited by Verbeek, Van den Dungen, Baker, 1999). Conjunctions are analyzed separately.

The morphosyntactic analysis is based on the 50 utterances selected from the transcript. The following types of utterances were discarded: repeated and unintelligible utterances and idioms (e.g. 'weet ik niet' *I don't know*) as well as elliptical answers i.e. answers to preceding questions without a finite verb and/or other utterance parts that can be inferred from the preceding question (e.g. adult: 'does it hurt' child: 'a little bit').

The analysis of the STAP test can be divided into two parts: the first part includes the quantifying morphosyntactic measures and the second part includes the qualifying morphosyntactic measures. The first part consists of language measures, which indicate the number of grammatical elements produced in a 50-utterance sample (e.g. number of finite verbs). All utterances are judged on morphological and syntactic correctness. Morphological errors include the incorrect use of inflectional suffixes for nouns, verbs and adverbs. On a syntactic level, errors include deletion, insertion, substitution and inversions of utterance parts. These utterance parts refer to nouns, verbs and determiners.

The main purpose of this dissertation is to compare the production of verbal morphology of hearing- and language-impaired children with the production of TD children. The measures belonging to the verbal domain are therefore of interest here. These measures are summarized and specified in Table 2.

Table 2. Overview and specification of the quantifying and qualifying morphosyntactic measures pertaining to the verbal domain.

quantifying measures	*specification*
finite verb	Total number of finite verbs. Fifty finite verbs are to be expected, -1 is scored when a finite verb is lacking, +1, 2 …k is scored with each extra finite verb that is produced (in the case of subordinations).
composed verbal predicates	Total number of composed verbal predicates, which include COP/AUX + past participle, COP/AUX + infinitive, COP/AUX + *aan het* infinitive, COP/AUX + *om te* infinitive. Both verbs need to be produced.
past participle	Total number of past participles, correct and incorrect (e.g. prefix is omitted).
past tense	Total number of past tenses, correct and incorrect.
qualifying measures	
main verb absent	Total number of omissions of the main verb (lexical or modal).
agreement errors	Total number of agreement errors, including incorrect subject-verb agreement (NB. Subject needs to be realized) and the deletion of the copula or AUX when main verb is present.
past participle error	Total number of past participle errors, including the deletion of the prefix.
past tense error	Total number of past tense errors, excluding the cases in which the context requires a past tense and the present tense is produced.

3.2 Psychometric review

To enhance our knowledge of the psychometric characteristics of the STAP test, an evaluation was carried out using the 10 psychometric criteria listed by McCauley & Swisher (1984). This is not an exhaustive list of criteria, but highlights a number of important psychometric criteria. The list can serve as a guideline to explore the potential use of the STAP test for the purpose of this dissertation. The STAP manual and its supplement (Van den Dungen & Verbeek, 1994; Verbeek et al., 1999) were consulted to complete the criteria list. The criteria and their content are presented in Table 3. A positive judgment is

given whenever the information is included in the manual and/or its supplement. The absence of information resulted in a negative judgment.

Table 3. Psychometric review of the STAP-test (psychometric criteria taken from McCauley & Swisher, 1984). A positive judgment is indicated by ✓ and a negative judgment by ✗.

no.	*criteria*	*definition*	*judgment*
1.	Description of the normative sample	Clarification of normative sample, including geographic information, 'normalcy' of subjects and socioeconomic status.	✓
2.	Sample size	Adequacy of (sub) sample size, subgroups with a minimum of 100 participants.	✗
3.	Item analysis	Systematic item analysis evidenced by quantitative methods.	✓
4.	Means and SD	Measures of central tendency and variability should be available.	✓
5.	Concurrent validity	Empirical evidence should be provided that the test categorizes children as normal or impaired.	✗
6.	Predictive validity	Empirical evidence should be provided that the test predicts performance on another, valid measuring the same aspect of language behavior.	✗
7.	Test-retest reliability	Empirical evidence should be provided that the test has a stability coefficient of .90.	✗
8.	Inter-examiner reliability	Empirical evidence should be provided for congruence between examiners, with a correlation coefficient of .90.	✓
9.	Description of test procedures	The test procedure should be described in such a detail that the test user can duplicate test administration.	✓
10.	Description of tester	The test manual should provide information about the qualifications a tester needs to adequately perform the test.	✓

The STAP manual and its supplement provide information to meet 6 out of the 10 psychometric criteria. The sample sizes on which the norms are based are too small. A total of 240 children, divided over 60 children per age group is to small to receive a positive judgment. Moreover, all children were drawn from the same geographical area of the Netherlands, namely the Amsterdam region. Careful consideration should be taken of the STAP norms.

According the STAP supplement, language measures have concurrent validity when a score below ≤ -2 SD can be obtained. This is closely related to the frequency of occurrence of a particular morphosyntactic element. For example, within the TD population, children are found who produce no past tenses within a spontaneous language sample. The consequence of this is that SLI children cannot be discriminated on the production of past tenses. However, this reasoning does not provide sufficient (empirical) evidence for concurrent validity.

No information has been found on predictive validity in the STAP manual and its supplement. Predictive validity requires a follow-up of the children included in the norms. This is time-consuming and is usually lacking in test manuals (McCauley & Swisher 1984; Plante & Vance 1994; Drenth & Sijtsma 2006).

The manual briefly reports on the test-retest reliability. Two language samples were obtained from 8 children. The time interval between these samples is not mentioned. The reliability of a measure was compromised if: 1) one score of the child was within 1 SD from the mean (i.e. reflecting a normal score) whereas the other score is 2 SD from the mean (i.e. reflecting a deviant score) or 2) one score was positively deviant and the other negatively deviant. Results of their analysis revealed instability on 14 language measures on at least one sample pair (i.e. one child). However, the authors do not elaborate on these language measures.

3.3 Implications and considerations

As mentioned in subsection 3.2, the reference population is rather small. Therefore, the accuracy of the norms is substantiated if the same scores are obtained with another group of TD children. Another motivation for including this second group is that the hearing-impaired children participating in this study were all monolingual speakers of Flemish. As the norms are based on Dutch-speaking children, the norms need to be tested for regional robustness.

The lack of empirical support for the concurrent validity of the STAP test motivated a validity study. The discriminating abilities of the language measures can be observed when a group of SLI children is contrasted with the group of TD children. Good validity implies that the language measures can be used to pinpoint children with a language performance beyond age expectancy and can readily be used for diagnostic evaluation.

Due to time limitations, a full study of the reliability is not possible. Some insight into the reliability can be obtained when applying the split-half test for internal consistency reliability, as mentioned in subsection 2.2.1. The underlying idea of this test is that utterances selected from a speech sample do not (or to a limited extent) differentiate from another set of utterances taken from the same

sample in its grammatical and syntactic structure. Stability within one sample increases the likelihood that the true score of the child can be estimated.

4. Reliability and validity testing

4.1 Participants

Fourteen TD children aged between 72 and 82 months (M=77) were selected to participate in this study. They were all monolingual speakers of Dutch, attended mainstream education and had no cognitive, perceptual or attentional disorders. The children were drawn from the East and North West of Flanders (Limburg). Another group of 15 6-year-old children clinically diagnosed with SLI participated in the present study. The group of SLI children had a mean age of 76 months (ranging from 72 to 83 months). All children in this group attended special education and did not have additional problems, besides their language impairment. Seven children lived in the Netherlands (Amsterdam) and 8 children lived in the East and North West of Flanders. For an overview of the participants see Appendix.

4.2 Data collection

Spontaneous language samples were recorded for the children in an individual setting, following the STAP protocol. For the TD children, the setting involved the researcher and the child. In the case of the SLI children conversations were held by a speech/language therapist, who knew the child well. These conversations were recorded by the researcher. According to the STAP guidelines, a familiar interviewer could ameliorate speech production in SLI children and therefore providing a language sample, which is comparable with the daily speech production of the SLI child. The topics of conversations varied from one sample to another, as the adults encouraged the child to discuss his/her own interests to reduce as much as possible any silent periods during the registration session. If the conversation was strained, picture books were used to elicit speech. Transcriptions were made according to the CHAT conventions, available through the Child Data Exchange System (MacWhinney, 2000)

4.3 Results

Norms

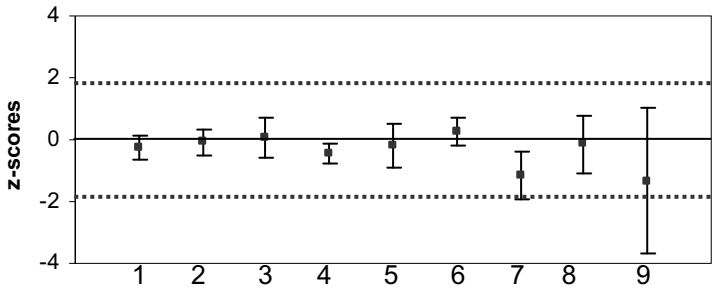
The means and standard deviations of the 14 TD children on MLU, quantifying and qualifying language measures are presented in Table 4. These results are compared to the scores of the reference group by transforming the raw scores of the 14 TD children into z-scores. This places each individual score within the normal distribution of the reference group. The mean z-score and standard

error of the mean per language measure is visualized in Figure 1. Z-scores are given for the reference group on the vertical axis; a z-score of 0 corresponds with the mean of the reference group, the dotted lines indicate the 95% confidence interval (i.e. between z ± 1.96).

Table 4. Mean scores and standard deviations per language measure for the 14 typically developing children.

language measures	*mean*	*SD*
MLU	6.58	0.61
quantifying verb measures		
finite verb	52.29	2.92
composed verbal predicates	14.86	6.69
past participle	3.86	2.38
past tense	11.21	10.28
qualifying verb measures		
main verb absent	0.79	1.05
agreement errors	1.43	1.22
past participle error	0.14	0.54
past tense error	0.50	1.34

Figure 1. Mean z-scores and standard deviations for the 14 TD children compared to the reference group. A z-score of 0 corresponds with the norm mean and the dotted lines indicate the 95% confidence interval. 1. MLU, 2. finite verb production, 3. composed verbal predicate, 4. past participle, 5. past tense, 6. main verb absent, 7. agreement error, 8. past participle error, 9. past tense error.



The mean z-score for the 14 TD children is for the measure of agreement errors and past tense errors over a standard deviation discrepant from the group on which the test norms are based. For all other language measures, mean scores are comparable between the two groups of TD children.

Reliability testing

The raw scores of the TD and SLI children are used to calculate the internal consistency reliability coefficient. The set of 50 utterances per child is divided in two. For both halves raw scores are calculated for MLU, the quantifying and qualifying verb measures of the STAP.

The split results in two raw scores per language measure and per child. The raw scores are entered in a correlation analysis. The Pearson correlation coefficient is taken as the reliability coefficient for half the test, i.e. R_k. This coefficient is used to calculate the reliability coefficient for the complete test (R) with the Spearman-Brown formula (see subsection 2.2.1).

The reliability coefficient can be used to calculate the Standard Error of Measurement (Var (E)). With this error term it is possible to calculate a 95% confidence interval. This is the range that includes the true score of the child with a 95% probability.

The results are summarized in Table 5.

Table 5. Reliability results for the STAP measures. R_k indicates the reliability coefficient for half the test (with p indicating if R_k is statically significant), R indicates the reliability coefficient for the complete test, var (E) refers to the standard error of measurement.

language measures	R_k	*p*	R	*var (E)*	*95% interval*
MLU	.88	<.01	.93	.21	.82
quantifying measures					
finite verb	.91	<.01	.95	1.14	5.53
composed verbal predicates	.54	<.01	.70	2.90	11.83
past tense	.38	ns	.55	5.23	20.05
past participle	.17	ns	.29	.84	6.24
qualifying measures					
main verb absent	.78	<.01	.88	.41	1.62
agreement error	.77	<.01	.87	.29	1.13
past tense error	-.09	ns	-.19	.32	1.28
past participle error	.74	<.01	.85	.12	.47

The MLU and finite verb production are reliably measured within one speech sample. Coefficients for both measures are above .90. The omission of the main verb (Main verb absent) and agreement errors are also fairly consistent across the speech sample, as indicated by consistency coefficient >.80. The occurrence of composed verbal predicates has a coefficient of .70, which is fair. However, the high variance in observed scores results in a high loss of accuracy. This means that a child's true production score lies almost 6 predicates above or below his/her observed score.

The occurrence of past tenses and participles is variable within a speech sample, resulting in a low reliability coefficient (<.70). The low reliability for the production of past tenses within one sample may not be surprising, as the use of the past tense is strongly dependent on context. When a child talks about something that happened to him/her in the past, then past tenses will occur. Including these utterances in the analysis will give a result for this variable. Choosing an utterance set excluding the 'past-tense' utterances then no score can be given for the past tense variable. Yet, the latter zero score does not mean that the child has no mastery over past tense morphology, it was just not present in the sample. The same accounts for the low reliability of errors in past tense production.

Validity testing

The contrasting-groups method is used to test the discriminating abilities of the language measures included in the STAP analysis. This method of validity testing is chosen because a criterion, or gold standard, for morphosyntactic development to which the STAP can be compared is not available for Dutch. Moreover, the evaluation of the diagnostic value of a test by means of the contrasting-groups method is frequently reported in the literature (see subsection 2.3).

STAP analyses were conducted for 15 SLI children, all 6 years of age. The means and standard deviations of the measures of interest to this dissertation are presented in Table 6. The raw scores are transformed into z-scores according the data given in the STAP manual. The mean z-scores as well as the standard deviations of this SLI group are plotted with the scores of the 14 typically developing children in Figure 2.

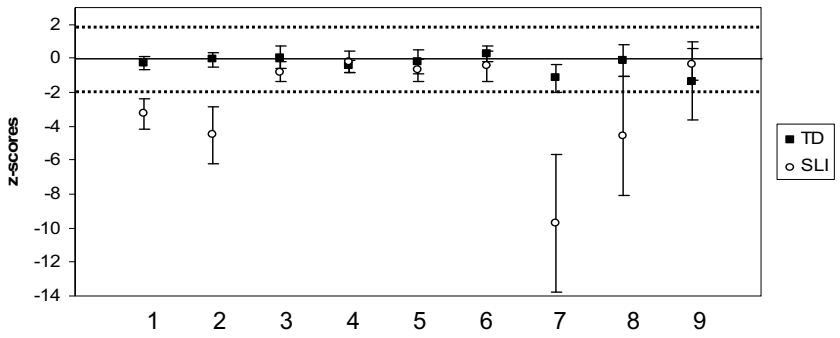
The graph in Figure 2 shows that the SLI children's score deviates from the reference group and the Flemish TD children on the MLU measures, finite verb production, agreement errors and past participle errors. The SLI children produce significantly shorter utterances when compared to their chronological peers with normal language ($U=22.0$, $p<.01$). Also, the SLI children produce significantly less inflected verbs as compared to their TD peers ($U=26.0$, $p<.01$). This could be explained by a high production of elliptical utterances

and/or the omission of copula and auxiliaries to a greater extent (no. 7 in Figure 2).

Table 6. Means and standard deviations of the language measures for the SLI children.

language measures	*mean*	*SD*
MLU	4.19	1.41
quantifying measures		
finite verb	35.47	12.71
composed verbal predicates	10.20	6.14
past participle	4.87	4.49
past tense	7.33	10.70
qualifying measures		
main verb absent	1.47	1.96
agreement errors	8.47	6.78
past participle error	1.47	2.07
past tense error	0.33	0.72

Figure 2. Mean z-scores and standard deviations for the SLI group compared to the 14 TD children and the reference group. A z-score of 0 corresponds with the norm mean and the dotted lines indicate the 95% confidence interval. 1. MLU, 2. finite verb production, 3. composed verbal predicate, 4. past participle, 5. past tense, 6. main verb absent, 7. agreement error, 8. past participle error, 9. past tense error.



The frequently reported weaknesses in the use of inflection morphemes for the SLI children is supported by the results in Figure 2. At the age of 6, the speech of TD children hardly contains any subject-verb agreement errors or errors in the production of past participles. By contrast, the SLI children produce significantly more agreement errors and past participles errors when compared to their typically developing peers (respectively $U=13.0$, $p<.01$ and $U=11.0$, $p<.01$).

4.4 Conclusion

With respect to the norms of the STAP test, this small scale study indicates that these are robust for MLU, all qualifying verbal measures and for the variable measuring the omission of the main verb and past participles. High to very high internal consistency reliability was attained for MLU, finite verb production, composed verbal predicates, omission of the main verb and agreement errors. Only four measures discriminated between SLI and non-SLI children. These were MLU, finite verb production, agreement errors and past participle errors. Therefore, the outcomes on MLU, finite verb production and agreement errors are reported in chapter 5, because these language measures are valid and reliable.