3 Methodology

3.1 Research questions and research design

Developing the ability to use L2 grammar structures proficiently, the role of explicit and implicit instruction therein, and the potential interaction between instruction and a number of potentially intervening variables are issues that this study addresses. In design, this study is both explanatory and descriptive, making use of a quasi-experimental repeated measures research design to investigate the development of two grammatical structures by means of explicit and implicit measures of L2 knowledge. In this section, the research questions and the main features of the research design are presented.

One of the purposes of this study was to provide some insight into how second language learners develop the ability to start using new grammar structures. In the previous chapter, the emergence of grammar was described as a process of abstracting grammar from exemplars, and seemingly algorithmic use of grammar structures is expected develop gradually and slowly. Fine-grained, longitudinal analyses of the emergence of grammatical phenomena during second language acquisition are scarce (DeKeyser, 2001; N. Ellis, 2003), and little is yet known about how L2 learners develop the ability to use grammar structures in spontaneous situations. As this study involved three repeated measures of spontaneous written response data over a period of four months, a descriptive exploration of these data was undertaken. These analyses were intended to gain qualitative knowledge of how becoming proficient in the two grammatical phenomena in focus arises. In addition, the outcomes were used to inform subsequent analyses. Thus, the first question to be addressed is:

RQ 1: How do second language learners develop the ability to use the target structures in spontaneous situations of second language use?

The primary goal of this study was to investigate the importance of explicit knowledge to the development of second language proficiency development. As already pointed out in 2.4, FFI research can provide evidence for the existence of
an interface between explicit and implicit knowledge, if explicit types of instruction can be found to be more beneficial to the development of L2 proficiency than implicit types of instruction. It was also pointed out that FFI research has not properly addressed the interface debate yet. It has tended to use explicit measures of progress, and studies that did measure implicit progress often did not compare explicit and implicit types of instruction. Comparing explicit and implicit grammar instruction, this study intended to establish whether explicit and implicit grammatical development were differently affected by these two types of instruction, and whether explicit instruction is indeed superior in promoting implicit knowledge development, as many researchers have argued (Bienfait, 2002; Doughty, 2003):

**RQ 2:** How are explicit and implicit FFI related to the development of explicit and implicit grammatical knowledge?

Another goal of this study was to investigate the interaction of explicit and implicit types of instruction with a number of individual characteristics. The notion that differences between individuals may explain the some extent why a particular kind of instruction is effective is actually common currency in SLA thinking. However, their precise relation to specific types of instruction has not been studied very frequently. This study has explored the relation between a number of defining individual characteristics and effective FFI. Developmental readiness, L1 similarity and Individual Differences such as aptitude, attitude, and age were selected for study:

**RQ 3:** How do developmental readiness, L1 similarity, and Individual Differences affect the success of explicit and implicit FFI?

Finally, another pervasive notion is that the value of explicit knowledge and FFI varies between different grammar structures. Although FFI research has addressed this question on a number of occasions, no coherent picture has emerged yet. Results seem to be contradictory, and most studies addressing this issue have used explicit measures of grammatical knowledge. Thus, the relationship between effective FFI and the instructed grammar structure is still rather obscure, and an investigation structures complexity while making a distinction between explicit and implicit grammatical development may provide valuable new insights. This issue was investigated by comparing the effect of FFI on two grammar structures varying in complexity, and the following research question was addressed.
RQ 4: Does the effectiveness of explicit and implicit FFI depend on the complexity of the instructed grammar structure?

In this study, three different treatment groups have been compared: two experimental groups and one control group. Participants of the experimental conditions took part in a computer-assisted language learning experiment in which they received information – either explicitly or implicitly – about two grammar structures. In 2.4.4, it was asserted that – in order to make for a fair comparison – exposure to the target structures should be kept equal for students in different treatment conditions. For this reason, the amount of input and the linguistic context in which the input was offered was precisely matched, the only difference being the degree of explicitness in presentation of the target structures. However, a true control group was included as well, in order to investigate if and how L2 learners develop their grammatical knowledge without additional focused input. The subject sample of this study consisted of approximately fourteen-year old learners of Dutch as a second language who were enrolled in intensive, often full-time, language learning programmes.

The development of L2 proficiency has been monitored by means of two measures of grammatical knowledge: grammaticality judgements and free written production, tapping into explicit and implicit grammatical knowledge respectively. The review of FFI research has sufficiently demonstrated that too little is known about the effect of FFI on L2 implicit knowledge. However, the use of both an explicit and an implicit measure of grammatical development allows for a deeper investigation of the interface debate. The two tests have been administered at three points in time during the experiment: once before the treatment, once immediately after, and once with a delay of 2 months after the treatment. This way, the participants of this study were followed over a period of three months’ time.

Two contrasting grammar structures have been instructed, and the explicit and implicit tests of grammatical knowledge assessed the students’ knowledge of these two structures. One structure was functionally simple and morphologic, while the other was functionally complex and syntactic. All participants in the experimental conditions received explicit instruction in only one of the two target structures, and they would receive implicit instruction in the other structure. Given this crossed design, this study is best thought of as two separate experiments in one, each focusing on a different grammar structure.

Besides monitoring students’ grammatical knowledge of two grammar structures, information was gathered to enable a rather precise characterization of the participants of this study. They took additional aptitude and proficiency
tests; their performance during the instruction was monitored by the computer; teachers were asked to make judgements about their motivation, (verbal) intelligence, and cognitive style; and the participants were asked to fill out background questionnaires about their age, first and second languages, and educational experience. All this information was used both to control for differences between the treatment conditions, and to investigate their relevance to developing grammatical knowledge of the target structures in focus.

The setting in which this research was conducted was the Dutch L2 classroom. Because of the focus on spontaneous second language use, laboratory-like research settings are precluded. In fact, the notion that explicit instruction may be more effective than implicit types of instruction is to a considerable extent based on studies that have taught and tested grammar rules in isolation; and in a number of cases impoverished artificial languages have been used. Such experiments lose out on ecological validity: if too much of the context in which language learning normally takes place is ruled out, variables may be lost that are of crucial importance to language learning, reducing the learners' task to what De Graaff (1997b) calls a "cognitive puzzle". His advice: "Only a careful combination of the advantages of both realistic and optimally controlled L2 learning environments can provide real opportunities for studying the effect of instruction on L2 acquisition" (p. 273).

An important design feature of this study was the use of the computer as instructor. This offered a number of advantages. First, the instruction was precisely the same for all subjects in one particular treatment condition. This way, control was exercised over the intervening effects that are introduced if different teachers are involved in explaining the rules of grammar. Second, the amount of exposure to the target structures could be carefully controlled and kept equal between the conditions. A third advantage was that the computer tracked the participants' performance while they were working with the programme. And the fourth and very important advantage of using the computer was that it offered a convenient way to randomly assign the participants to one of the two treatment conditions. As a result, no set classes were used to populate the individual treatment conditions which would likely have introduced bias. Rather, the two treatment conditions were represented by students from all the classes participating in this study. Only the 'true' control group was created on the basis of two set classes, because they did not work with the computer programme.

In the remainder of this chapter, the research design shortly outlined here will be described in detail, starting with a description of the target structures and how
they were instructed (3.2), and followed by an explanation of how grammatical development was assessed (3.3). Next, the participants of this study and their background are introduced (3.4), and they are characterized in terms of potentially decisive individual characteristics (3.5). The chapter is concluded with expounding the organisational and statistical procedures (3.6).

3.2 The instruction

3.2.1 The target structures

The structures selected for instruction and comparison are the degrees of comparison and subordinate clauses. Both theoretical and practical issues have been considered when choosing the target structures, although the practical issues have prevailed.

In 2.3.6, it was argued that there are no fundamental differences in learning different grammar structures: they start as exemplars and gradually become open class structures. However, particular features of the target structure may affect a structure’s learnability and perhaps teachability. The literature mentions a number of such features, such as functional and formal complexity, reliability, scope, and the linguistic nature of the target structure (morphologic, syntactic, and formulaic). In addition, hierarchical depth was put forward as one of these factors. Choosing the proper target structures is not simple though. A practical point is that it is very difficult if not impossible to control for all the features of grammar structures simultaneously. For example, if one decides to compare two grammar structures contrasting with respect to scope, they are likely to differ as well according to one of the other features. This makes it difficult to attribute found differences to one particular feature. A complicating issue is that characterizing a grammar structure according to these features does not constitute categorical choices. Most of these factors are best seen as gliding scales rather than dichotomous choices. Another important practical consideration was the requirement that the participants had to be able to produce a fair amount of Dutch, given the focus on proficiency in this study. Beginning L2 learners simply cannot be expected to be very productive in their L2 yet. Obviously, the target structures had to be more or less new to these somewhat advanced L2 learners. Finally, a design requirement specifically related to this study was that the two grammar structures occur more or less simultaneously in the students’ interlanguage.
Several steps have been undertaken to make a well-informed choice of target structures. A first selection of structures was made on the basis of Zebra (1999), a widely used method for teaching Dutch as a second language (see also 3.4.2). The teacher manual belonging to this method provides detailed information about the underlying structural syllabus. A small scale but intensive pilot study was then conducted in order to assess whether the selected structures were indeed emerging in the interlanguage of students that had had approximately one year of Dutch training. This pilot involved inviting students to write short pieces of text, and was also intended to evaluate how well the target structures could be elicited implicitly (see also 3.3.3 on the measurement of implicit grammatical knowledge).

The degrees of comparison and subordinate clauses were found to meet the set requirements best. L2 learners start using these (correctly) rather late; the structures were found to occur more or less simultaneously; and they can be elicited with considerable success. The structures also differ fundamentally from each other: the degrees of comparison are functionally simple and morphological, while subordinate clauses are functionally complex and syntactic. In the remainder of this subsection, both structures will be described in more detail.

**Degrees of comparison**

The degrees of comparison (DoC) in Dutch are very similar to the English degrees of comparison. The comparative and superlative are formed by attaching –er and –st to the adjective or adverb. Both suffixes carry independent meaning, and thus change the meaning of the adjective or adverb they are attached to. As each suffix carries a straightforward meaning, the structure must be considered functionally simple. However, Dutch also allows for periphrastic DoC by means of the use of the quantifiers meer (more) and meest (most). Periphrastic use is restricted to complex adjectives, though (Aarts & Welker, 1993); and in the vast majority of cases, periphrastic realization of the DoC will be considered ungrammatical in Dutch. Periphrastic comparison was not subject of the instruction, simply because the kinds of adjectives that are realized periphrastically in Dutch are generally not part of the vocabulary of the L2 learners participating in this study.

As explained in 2.2.4, functional complexity refers to the relationship between meaning and form. If this relationship is straightforward and transparent, then the structure can be considered functionally simple. At first sight, the suffixes that mark the degrees of comparison seem functionally simple in that they have a straightforward meaning. However, Dutch uses the –er suffix also to make nouns...
of verbs (*werken, werker*), which may be a complicating factor for L2 learners. Formal complexity refers to the number of operations that have to be performed in order to use the rule correctly. Realization of a form the degrees of comparison requires suffixation only, and can therefore be considered formally simple. Such suffixation does cause changes in spelling of the adjective or adverb sometimes (for example, -f becomes –v when suffixed: *lief, liever, liefst*). These cannot be heard, and they are in accordance with normal Dutch spelling conventions.

The notions of *scope* and *reliability* refer to the extent to which a particular rule is probabilistic. Hulstijn and De Graaff (1994) define *scope* as the absolute number of cases covered by a rule and reliability as the percentage of instances for which the rule holds (number of exceptions). They also draw borders, which they admit are arbitrary: a rule is large in *scope* when it covers more than 50 cases, and its reliability is high when it applies to over 90% of all cases. According to these definitions, the degrees of comparison can be considered large in scope, as there are over 50 adjectives that can be inflected. The rule is quite reliable as well, as periphrastic comparison is not very frequent in Dutch, and occurs with adjectives that probably lie outside the L2 learners’ vocabulary.

**Subordinate clauses**

Subordinate clauses refer to a specific class of embedded sentences in which one clause is subordinate to the other – main – clause. Subordinate clauses are introduced by subordinating conjunctions, which express how the main clause and the subordinate clause relate to each other. For example, subordinating conjunctions can express among others temporal, causal, and conditional relations. Particularly difficult for L2 learners is that the finite verb – by default in second position in main clauses – is positioned at the end of subordinate clauses (see Table 3.2). This repositioning of the finite verb is entirely meaningless, in that it does not change the meaning of the sentence. Neglecting to position the verb at the end of the subclause does not lead to incomprehension or miscomprehension.
Subordination is somewhat difficult to classify in terms of functional complexity. As pointed out, the repositioning itself does not carry meaning. However, the meaning of the clause changes, because of the conditionality or causality it expresses by means of the coordinating conjunction. L2 learners therefore might associate the rule with the relationship the clause intends to express. A complicating factor specifically related to expressing causality is that Dutch also has a coordinating conjunction expressing causality: want. In Table 3.2, example 5 illustrates two sentences that are synonymous in meaning, but not in word order. The only valid cue for repositioning the finite verb, then, is the subordinate conjunction itself. In terms of formal complexity subordination must be considered simple, because it requires only one operation.

In terms of scope and reliability, subordination can be considered a reliable rule that is wide in scope. The occurrence of one of the subordinating conjunctions almost always leads to verb repositioning. The only exception for applying the rule may be when the subordinate clause is too complex to move the verb all the way to the end. However, this occurs seldom. The applicability of the rule is virtually endless, and especially conditional and causal subordinate clauses occur rather frequently in everyday language.

### Table 3.2 Subordinate clauses in Dutch (examples taken from the data collected for this study)

<table>
<thead>
<tr>
<th>Example</th>
<th>Main clause</th>
<th>Subordinate clause</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>Mijn vriend vertelt een grap.</td>
<td>..., als mijn vriend een grap vertelt.</td>
<td>..., when my friend a joke tells.</td>
</tr>
<tr>
<td>2)</td>
<td>De olifant plukt appels</td>
<td>..., omdat de olifant appels plukt.</td>
<td>..., because the elephant apples picks.</td>
</tr>
<tr>
<td>3)</td>
<td>Hij heeft tegen mij gelogen.</td>
<td>..., als hij tegen mij heeft gelogen.</td>
<td>..., when he to me has lied</td>
</tr>
<tr>
<td>4)</td>
<td>Zij wil een ijsje eten.</td>
<td>..., omdat zij een ijsje wil eten.</td>
<td>..., because she icecream wants to eat.</td>
</tr>
<tr>
<td>5)</td>
<td>...</td>
<td>..., omdat de kleine radio goedkoper is.</td>
<td>..., want de kleine radio is goedkoper.</td>
</tr>
</tbody>
</table>
3.2.2 The instruction

The value of three types of instruction has been evaluated: explicit instruction (EI), implicit instruction (II), and no instruction (NI). In both experimental conditions (EI and II), the participants of this study received focused input of the DoC and SubC by means of a computer-delivered language learning programme. Participants in the NI group simply did not receive any focused instruction; and therefore, no extra input. They continued to do their normal classroom activities while the experimental groups were working with the computer programme. The NI condition was included to investigate how L2 learners develop their grammatical knowledge without input specifically focused on the target structures. This way, the effect of receiving focused instruction – either explicit or implicit – could be evaluated with reference to not receiving any focused instruction.

A crossed treatment design was used (see Table 3.3). This meant that participants in treatment group 1 would receive explicit instruction on the DoC and implicit instruction on SubCs; for participants in treatment group 2, this was the other way around. Consequently, all participants in the two experimental conditions received explicit and implicit instruction, but which target structure they dealt with explicitly depended on the treatment group they were in. This way, the treatment groups were each others controls with respect to exposure to the target structures. Also, all participants did a variety of explicit and implicit kinds of exercises, which made the programme more enjoyable. The instruction was relatively short, in that on average three to four hours were spent on the programme by the participants.

The instruction was delivered by means of a computer programme called Taal in Themas, which was developed for the occasion. As already pointed out in 3.1, the use of the computer offered some important advantages. Using the computer as a teacher ensured that the instruction was precisely the same for all subjects. But also the amount of exposure to the target structures could be carefully matched, which is considered to be an important design requirement. Input matching will be discussed in the next subsection. Detailed information about the

| Table 3.3 Organisation of the experimental treatment groups |
|---------------|----------------|
| Condition     | Group 1         | Group 2         |
| Explicit instruction | Degrees of Comparison | Subordinate clauses |
| Implicit instruction  | Subordinate clauses | Degrees of Comparison |
development of the programme and in-class procedures is given in 3.6. In addition, some screenshots of the programme can be found in Appendix A1.

Most would agree that any FFI should be integrated in functional and communicative contexts. Using the computer for the instruction is to some extent at odds with this requirement, simply because computers do not communicate. The kinds of activities that can be offered by means of the computer are limited, and such activities do mostly not serve a functional purpose comparable to those of everyday communication. This is not problematic, though, as long as the explicit instruction leads to explicit knowledge, and as long as the input offered during the explicit instruction is matched with implicit input. The instruction was nevertheless set in a functional context. A series of eight lessons was developed centred around three different themes: Advertising, the Olympic Games, and The Netherlands and Water. Each lesson consisted of short texts about these themes, and students would do a number of unfocused exercises related to the texts. These texts and exercises would be interlaced with focused exercises, either implicitly or explicitly offering FFI in the DoC or SubC. As there were eight lessons in total, four focused on the Degrees of Comparison, and four focused on Subordinate clauses. Whether this focus was explicit or implicit depended on the treatment group the students were in (see Table 3.3). In addition, a detailed outline of the instruction is provided in Appendix A2.

In line with the definition of explicit knowledge given in 2.4.2 and the purpose of this study, the aim of the explicit instruction was to create an explicit understanding about the target structures. This aim was explicitly stated from the outset, this way establishing the intention on behalf of the learner to deliberately learn a particular rule. Focused exercises would be clearly marked by a special logo, caption, and background colour. In addition, the exercises would be introduced by rule explanation, either introducing new aspects of the rule, or repeating earlier discussed aspects. In the subsequent exercises, the participants were continuously reminded to apply the rules just explained.

Following recommendations by Kuiken (1999), and the example of Zebra grammatica (Andringa et al., 2000), rule explanation focused primarily on functional use of the structure. The instruction would be accompanied by examples, and if possible, visual aids. Care was taken not to make the explanation too technical. Given that the participants’ command of specific linguistic terms was limited, such terms were used sparingly; and when used, they would be carefully defined. With respect to subordinate clauses, it should be pointed out that the instruction focused primarily on conditional and causal types of subordinate clauses. These are by far the most frequent in Dutch, and the use of
other subordinate clauses proved difficult to elicit (see 3.3.3). Nevertheless, other
types of subordinate clauses were not ignored: they were simply less frequent.
The Appendices A3 and A4 list all the explicit instruction as it was provided to the
participants.

Although using the computer as instructor does limit the possibilities, several
kinds of exercises were used to strengthen the students’ explicit knowledge. In
the receptive phase of the instruction, students had to recognize the target
structure in example sentences by means of multiple choice or yes/no questions.
Input enhancement was also used in dialogues in which the target structure
would be underlined. As the students progressed, more productive exercises
would be used, such as gap-filling and sentence completion. As indicated,
Appendix A2 provides a precise outline of all the exercises students had to do.

The implicit instruction was created to invite the students to process the
target structure for meaning. However, the students were never made aware of
this: the focus was entirely on meaning in the perception of the participants; and
there were no explicit attempts to get the language learner to notice the target
structure that would be abundantly present in the comprehension exercises. An
implicit focus-on-form was achieved by presenting the target structures in the
context of text comprehension exercises. These exercises were always based on
one of the texts just read, and providing the proper answer simply required
remembering or checking the facts in the text. The participants would have to
read the instruction and possible answers carefully; and thus they were forced to
process the target structure for meaning. In later stages of the programme, the
participants were also asked to type their answers to questions that were
designed to elicit the target structures.

During the focused exercises, immediate feedback was provided. In the
beginning stages of the instruction, the feedback function was often used to
restate the rule (explicit feedback), or repeat the correct answer (implicit
feedback). In later stages, feedback might simply consist of conveying whether
the answer given was right or wrong.

3.2.3 Input matching

As explicit instruction may lead to concomitant implicit learning effects, care
was taken to keep the amount of input and the nature of the input equal. The
input in the experimental conditions was precisely matched, meaning that the
same input sentences were used to create either an explicit or implicit focus on
the target structures. Given that explicit and implicit instruction are very
different in nature, this was not an easy task. In practice, input matching was realized by modifying the text comprehension exercises into grammar exercises. This way, the original implicit focus on the target structures was changed into an explicit focus, while the input itself remained unchanged. Thus, the subjects in both conditions were exposed to the same input, meaning that they had more or less equal chance to learn implicitly from the instruction.

Despite the efforts to keep the input between the conditions similar, there were nevertheless some differences. One obvious difference involves the rule explanation that was part of the explicit instruction. Also, students in the EI condition were given the opportunity to return to the rule explanation screens during the grammar exercises. Providing this opportunity was considered an important aid to effectively develop explicit knowledge, but it may have led to unequal exposure. In order to make up for these inequalities, students in the implicit condition received extra input during the exercises themselves, for example, by means of feedback screens. This way, the number of times that participants in the EI and II conditions encountered the target structures was more or less equal.

Table 3.4 provides examples of how this input matching has been realized throughout the instruction.

### 3.3 Assessing grammatical development

#### 3.3.1 Introduction

The development of grammatical proficiency has been monitored by means of two tests of grammatical knowledge, one tapping into explicit knowledge, and one tapping into implicit knowledge. Several authors have cautioned that the effect of instructional efforts may not be immediate, but delayed (Gass, 1991, 1997; Lightbown, 1994; Gass en Varonis, 1994). The argument is that teaching explicit knowledge may not impact immediately on the development of implicit knowledge, but will facilitate implicit learning once the L2 learners starts acquiring the particular structure. For this reason, many FFI studies have included delayed knowledge tests. In this study, the L2 learners’ grammatical knowledge has been assessed at three different points in time: once before the instruction (T0), once immediately after the instruction (T1), and once with a delay of two months after the instruction (T2).
### Table 3.4: Examples of Input Matching (Translated from Dutch)

<table>
<thead>
<tr>
<th>Explicit Instruction</th>
<th>Implicit Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DoC: Chapter 2, ex. 1; (see App. A2):</strong></td>
<td></td>
</tr>
<tr>
<td>Which word is a form of the DoC?</td>
<td>This statement is about the text you just read. Is it true or false?</td>
</tr>
<tr>
<td>When a commercial is funny, people think it is nicer.</td>
<td>When a commercial is funny, people think it is nicer.</td>
</tr>
<tr>
<td><strong>DoC: Chapter 2, ex. 2; (see App. A2):</strong></td>
<td></td>
</tr>
<tr>
<td>These 3 sentences are taken from advertisements. Which sentence does not contain a form of the DoC</td>
<td>What does this advertisement try to tell you?</td>
</tr>
<tr>
<td>A – X now washes even cleaner.</td>
<td>A – X now washes even cleaner.</td>
</tr>
<tr>
<td>C – X is not expensive.</td>
<td>C – X is not expensive.</td>
</tr>
<tr>
<td><strong>DoC: Chapter 7, ex. 1; (see App. A2):</strong></td>
<td></td>
</tr>
<tr>
<td>Fill in one of the (three) words in the sentence below. Use a form of the DoC</td>
<td>Fill in of the (three) words in the sentence below. Create a logical sentence.</td>
</tr>
<tr>
<td>X is one of Holland’s [high/fast/small] skaters.</td>
<td>X is one of Holland’s [high/fast/small] skaters.</td>
</tr>
<tr>
<td><strong>SubC: Chapter 1, ex. 3; (see App. A2):</strong></td>
<td></td>
</tr>
<tr>
<td>Which sentence is does not have a subordinate clause?</td>
<td>Why does X like the Sony discman?</td>
</tr>
<tr>
<td>A – X like the Sony, because it is very small.</td>
<td>A – X like the Sony, because it is very small.</td>
</tr>
<tr>
<td>B – X likes the Sony, because it looks smart.</td>
<td>B – X likes the Sony, because it looks smart.</td>
</tr>
<tr>
<td>C – X like the Sony: it is very small and it looks smart.</td>
<td>C – X like the Sony, because it is very small and it looks smart.</td>
</tr>
<tr>
<td><strong>SubC: Chapter 8, ex 1; (see App. A2):</strong></td>
<td></td>
</tr>
<tr>
<td>Make from two sentences one complex sentence:</td>
<td>Read the text (excerpt given) and finish the sentence</td>
</tr>
<tr>
<td>1 - You can earn a lot of money. 2 - You are good in sports. You can earn a lot of money if ...</td>
<td></td>
</tr>
</tbody>
</table>
Both the explicit and the implicit test were presented in precisely the same form at each time of measurement. A number of researchers have adopted the technique of changing words in their tests just to give the language learner the impression that their subjects are doing a different test (i.e., Sanz & Morgan-Short, 2004). However, it was felt that even changing words could lead to differences in test difficulty. Therefore, no modifications to the tests were made in order avoid all possibility of introducing such differences. The test takers were explicitly told that the tests were precisely the same, so their teachers could see they had improved their knowledge of Dutch.

In the remainder of this section, both tests are discussed in detail.

### 3.3.2 Explicit grammatical knowledge

In 2.4.2, a number of requirements were put forward that explicit knowledge tests should meet. Most importantly, the task should be such that it expressly calls on one’s knowledge about a particular rule of grammar, and consequently, it should allow the test taker the time to do so. For these reasons, an untimed paper-and-pencil grammaticality judgement task was used. Also, grammaticality judgement tasks call on explicit knowledge without making explicit reference to the target structures in focus, which would offer an advantage to students who had explicit instruction. Grammaticality judgement tasks require the L2 learner to judge whether a particular stimulus is grammatically correct or incorrect. In this particular test, the participants were given target sentences containing incorrect realizations of the target structure, and were instructed to indicate whether they recognized any errors in the sentence. In addition, they were asked to underline the error if they saw one. For the degrees of comparison, subjects had to see that a particular adjective was not marked for comparison while the context required this. For subordination, default word order was used in the subordinate clause. Thus, the target structures were always presented incorrectly in the test. Examples are provided in Table 3.5. The entire test can be found in Appendix B1.

<table>
<thead>
<tr>
<th>Table 3.5: Grammaticality judgement examples.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Anne is best lang, maar Lien is nog veel lang.</td>
</tr>
<tr>
<td>2) Ik wil graag een ringetje in mijn neus, omdat ik vind dat onzettend mooi.</td>
</tr>
</tbody>
</table>
The number of items of which this test consisted was small: it consisted of 20 items in total, six focusing on the degrees of comparison, and eight focusing on subordinate clauses. Five sentences contained errors unrelated to the target structures, and one sentence was correct. The number of test items focusing on the target structures must be considered small, then. In general, test reliability is positively improved as the number of test items increases. However, a small scale pilot (eighteen subjects) with a 40-item GJ test revealed practical problems. As the subjects of this study were selected to have little knowledge of the target structures in focus, taking this test proved simply too difficult. It took too much time, and some students became quite anxious. In light of this experience, it was decided to keep the number of test items small. In addition, the five sentences that contained errors not focusing on one of the target structures were kept quite simple.

The explicit knowledge score expresses the number of times students identified incorrect use of the target structure in the stimulus sentences. Table 3.6 provides the reliability coefficients for both the degrees of comparison and subordinate clauses. Although the test was administered as one task, these coefficients are reported for each target structure independently. The reason is that all analyses in this study will be presented independently for each target structure. The reliabilities for both target structures and at each time of measurement are sufficiently high, especially given the small amount of test items.

By way of validating the explicit knowledge test, a second test was administered (to be found in Appendix B2). This test measured the participants’ explicit knowledge of the target structures by means of controlled production tasks, and because of the more explicit nature of this test, it was administered at T1 only. There were ten items in the test for each structure. For the degrees of comparison, gap sentences were given in which students had to fill in a

<table>
<thead>
<tr>
<th>Table 3.6: Reliability coefficients (KR-20) of the grammaticality judgement test and the controlled production test (T1 only)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td><strong>Grammaticality judgement</strong></td>
</tr>
<tr>
<td>Degrees of comparison</td>
</tr>
<tr>
<td>Subordination</td>
</tr>
<tr>
<td><strong>Controlled production</strong></td>
</tr>
<tr>
<td>Degrees of comparison</td>
</tr>
<tr>
<td>Subordination</td>
</tr>
</tbody>
</table>
comparative form. The adjective to be used was given. For SubC, students had to combine two clauses by means of a given subordinate conjunction. The subjects had to do this task on the basis of an example sentence. Although no explicit reference was made to the target structures, this test more explicitly focuses the students towards the target structures. The reliability coefficients can be found in Table 3.6. After correcting for attenuation, high correlations were found between the two tests (.81 for the DoC, and .73 for SubC). This indicates that the two tests measure the same construct to a considerable extent; and therefore, that the students’ explicit knowledge of the target structures was quite adequately measured by the grammaticality judgement task. The controlled production test was not used for purposes other than validating the grammaticality judgement task.

### 3.3.3 Implicit grammatical knowledge

The primary requirement of implicit grammatical knowledge tests is that they assess the ability to use the target structures in spontaneous situations of language use. Spontaneous use of the target structures means that their use was unplanned or incidental, and that the structures were used to fulfil a clear communicative purpose. In addition, although the notion of proficiency actually covers accuracy, fluency, and pragmatic aspects of language use; in this study the focus is on the ability to use the grammar structures accurately.

Implicit grammatical knowledge has been measured by means of a free written response test. A number of considerations underlie the choice for a written production test. Most importantly, the effects of instruction are expected to affect written proficiency before they do oral proficiency. The reason for this is that conversation – being more time-pressured – requires higher degrees of automatization (Bialystok 1989). Therefore, it is likely that L2 learners have more opportunity to use their explicit knowledge as a monitor in written proficiency tests. A number of FFI studies confirm this. Day and Shapson (2001) and VanPatten and Sanz (1995), for example, found no effect of their instruction on oral proficiency tests, but it did affect written proficiency (see also 2.4.4). Thus, a written proficiency test probably maximizes chances of finding an effect. A more practical consideration to choose for a written proficiency test was that collecting and processing written data is simply faster and less work-intensive, thus allowing for a larger subject pool.

The free written response test was designed to elicit incidental use of the target structures. Students were required to read a short text describing a
particular situation, or to look at a particular comic strip or picture. Questions were asked about these, and the students had to respond to these appropriately in one or two sentences. The situations were designed in such a way that the L2 learner would likely use the target structures, but there was no explicit invitation to do so. The students were expressly invited to use their imagination when replying to the questions, as long as their response was appropriate. Consequently, there was no guarantee that the L2 learners would actually use the target structures in their response. Students were kept unaware of the fact that they were tested for their knowledge of and ability to use a particular grammar structure. The test was presented as a writing proficiency test, and the instruction simply asked them to write short sentences in reply to the questions, and to write carefully and correctly.

Test items that intended to elicit forms of the degrees of comparison placed the test taker in a situation in which they had to make comparisons. For example, they were asked to compare two radios, or to choose between two pairs of jeans. No attempt was made to specifically elicit the use of either comparative or superlative forms, as comparatives and superlatives can be used more or less interchangeably depending on the perspective taken. For example, in response to the question which radio they would buy, students might answer: "I buy the cheapest" or "the one that is cheaper". Both are equally appropriate responses to the question.

Elicitation of subordinate clauses aimed primarily for conditional and causal SubCs. These are by far the most frequent in Dutch. In addition, the pilot study revealed that SubCs expressing relations other then conditional or causal relations proved to be very difficult to elicit. Conditional SubCs can be elicited very effectively by means of questions asking when (under what condition) something may occur. For example, questions starting with the Dutch question word *wanneer* (when, under what circumstances) resulted in a very high rates of conditional SubC use. Situations aiming for the use of causal SubCs proved to be less compelling. Students were asked to explain what caused a particular thing to happen in their opinion, or why they would make a particular choice. For example, when asking why they would choose a particular radio, students might answer: "because it is cheaper". However, they also frequently omitted the subordinate conjunction: "it is cheaper"; or avoided the use of the subordinate conjunction by means of the synonymous coordinate conjunction want: "[want] it is cheaper". For this reason, the test contained only three questions aiming for conditional SubCs, and eighteen questions aiming for causal SubCs.
In total, 29 test items intended to elicit one of the target structures. Eight focused uniquely on the DoC; eleven elicited SubC; and ten items invited the use of both DoC and SubC simultaneously. Consequently, there were 18 DoC items and 21 SubC items. The test items were selected on the basis of small-scale pilots with twelve L2 learners that had already acquired the target structures. If an item elicited the use of one of the target structures in at least 50 percent of the cases from these advanced learners, the item would be considered fit for the test. During the pilot phase, it also became clear that elicitation of incidental use of subordinate clauses other than conditional or causal subordinate clauses proved extremely difficult. For this reason, only conditional and causal SubCs were elicited. The entire test can be found in Appendix B3, and the DoC and SubC items are marked.

The implicit knowledge score calculated on the basis of the free written response task simply consisted of the amount of times the target structures were used correctly in the test: the frequency of occurrence. Each occurrence of the target structure was counted, so if one test item generated multiple uses of a target structure, each use would be counted. For this reason, percentages of correct use could not be calculated: they would in some cases exceed a hundred percent. However, for calculating test reliability coefficients, a test item was scored ‘correct’ if it elicited at least one correct instance of the target structure. Incorrect use, no use or avoided use were scored as ‘incorrect’. Table 3.7 provides the reliabilities for both grammar structures. As can be seen, they are considerably high.

<table>
<thead>
<tr>
<th>Table 3.7</th>
<th>Reliability coefficients (KR-20) of the implicit knowledge tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Free written response task pre (n=101)  post (n=101)  2nd post (n=76)</td>
</tr>
<tr>
<td>Degrees of comparison</td>
<td>.83</td>
</tr>
<tr>
<td>Subordination</td>
<td>.91</td>
</tr>
</tbody>
</table>
3.4 The participants

3.4.1 Introduction

The L2 learners participating in this study were students from ten different schools from the North and West of The Netherlands. The participants in the experimental conditions (EI and II) came from eight different schools, and they were randomly assigned to either treatment condition by means of the computer programme. There was no matching of any kind. As the subjects in the NI condition did not work with the computer programme, this group consisted of two intact classes from two different schools. In this section, the participants are introduced. In 3.4.2, their current educational background is discussed, followed by a discussion of the mortality that this study suffered from in 3.4.3. Subsequently, the students are characterized in terms of general Dutch proficiency (3.4.4), day-to-day exposure to the Dutch language (3.4.5), and performance during the instruction (3.4.6).

3.4.2 The subjects and their educational background

The subject pool for this study consisted of 101 students from so-called 'ISK' classes. ISK is short for 'Internationale SchakelKlas, and refers to a special secondary school educational facility for immigrant students. Mostly, the ISK is an integrated part of a secondary school, but sometimes local authorities install a separate ISK school for the entire district (Berenst, Bienfait, Hofstede & Van der Schaaf 1999). ISK classes provide intensive Dutch courses for students ranging between 12 and 18 years old, with the intention to enable them to follow regular secondary programmes as soon as possible. The students of ISK classes are a diverse group. They came from many different countries, and came to the Netherlands for various reasons. Sometimes they fled their home country because of unstable political regimes or poor economic conditions, but they also came because of parents working in the Netherlands. Often, they are involved in lengthy legal procedures in order to obtain a ‘green card’ that would allow them to stay in The Netherlands. While awaiting the outcome of such procedures, they often live in special housing facilities, and have the legal right to education (Berenst et al., 1999). Generally speaking, these students have a strong motivation to learn Dutch, because for most their future depends on knowing
Dutch. The subjects of this study came from 10 different ISK schools from the North and West of the country. These schools were selected based on proximity, availability of students, and willingness to cooperate.

Generally speaking, grammar is not very prominent in the curricula of ISK classes. As pointed out, the majority of ISK’s use Zebra (Alons, Bienfait, & et al., 1999) as the basis of their Dutch instruction, which prepares immigrant children for the regular secondary educational system. Grammar instruction in Zebra is mainly implicit. Although each lesson does target particular grammatical structures, there is no explicit reference to these structures. Zebra is accompanied by a reference grammar, Zebra Grammatica (Andringa, Bienfait, Kuiken & Van der Schaaf 2000), but it is not introduced until lesson twenty-six. In practice, this means that students do not work with grammar until their second year of Dutch.

Some ISK classes do organise additional grammar instruction based on other methods of Dutch as a second language (mostly for adults), or on self-made materials. Of the ten participating ISK schools, three schools indicated providing such additional grammar hours. In one school, the amount of time spent on grammar exceeded one hour per week. In addition, all participating teachers indicated providing on demand grammar instruction, i.e. when one of their students asked for it. None of the teachers indicated having taught one of the structures focused on in this study.

3.4.3 Mortality

This study suffered from considerable mortality in the subject sample. From pre-test to post-test, 101 students were tested, but a quarter of the subjects was lost between the post-test and the second post-test. Ultimately, the grammatical development of 76 L2 learners was followed over a period of three months. A number of reasons underlie this loss. First, this study was conducted in a time of changing national admittance policies. This led to a sharp decrease in the number of refugees allowed into The Netherlands. In less than two years’ time, the number of ISK students decimated and many ISK schools had to close their doors, or were threatened to do so within the near future. One group was lost because of this: the second post-test could not be organised in the school’s last weeks of class. Another reason for loss was that some students simply left The Netherlands, or were relocated to other schools or to other areas in The Netherlands. Finally, some students were lost because of absence on the days of testing. Table 3.8 provides an overview of the subjects per condition. Especially
the NI group – already the smallest group – suffered from considerable loss, which makes this group rather small. There was another problem with the NI group. One class inadvertently received some explicit instruction in the DoC. For this reason, they had to be excluded from some of the analyses.

A point of concern is whether the loss was random. This concern will be addressed throughout this study by investigating whether the mortality somehow affected the nature of the subject sample and the results presented in this report.

### 3.4.4 General L2 proficiency

The subjects were selected for participation in this study on the basis of their general L2 proficiency. The intention was to put together a subject sample that was homogeneous with respect to their proficiency. In fact, a number of different indicators of proficiency were gathered. They will be described here shortly. Each measures proficiency in a different way, and may tap into different aspects of a students’ proficiency. These measures were merely used to gain insight into the participants’ general level of proficiency.

**De ISK-toetsen**

*De ISK-toetsen* (Schuurs, 1999) were primarily used for subject selection. They are a series of standardized tests especially developed for assessing general proficiency of Dutch of ISK students. The test series is used widely by ISK’s in The Netherlands to monitor progress in Dutch, and it contains listening, speaking, writing, and reading components. As most schools use this test, it provided a good starting point for subject selection. Subjects scoring approximately fifty points on this test were considered fit for participation. For 85 subjects, ISK test scores were obtained. Their mean score was 53.1 ($SD = 5.2$), and observations ranged between 41 and 67. No significant differences were found between the three conditions on ISK test scores, irrespective of mortality.

<table>
<thead>
<tr>
<th>Treatment condition</th>
<th>N (pre post)</th>
<th>N (pre post delayed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition 1: DoC expl – Sub impl</td>
<td>41</td>
<td>32</td>
</tr>
<tr>
<td>Condition 2: Doc impl – Sub expl</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>Control: no instruction (NI)</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>101</strong></td>
<td><strong>76</strong></td>
</tr>
</tbody>
</table>
The C-test

The C-test provides another indication of overall proficiency in the second language (Klein-Braley, 1997). C-tests are a type of cloze tests, in which parts of words are deleted rather than entire words. In C-tests, every \( n \)-th word is partially deleted. Students have to complete the word on the basis of the context and the first half of the word: the first and last sentences of the text are left untouched. The advantage of the C-test is that it generates many test items on a small text. In addition, if multiple texts are used with different topics, knowledge-of-the-world effects can be minimized. Also, it can be scored quickly and objectively (Klein-Braley, 1997; Weir, 1988).

The test used for this study was developed for the occasion. The test can be found in Appendix C1. It consisted of three short and simple texts with different topics. On the basis of small-scale pilots, it was decided to delete every fourth word of the text. In total, the test consisted of 64 items. It was administered to all students at T0 together with a number of other tests. Test reliability was found to be adequate (Cronbach's alpha = .85). The mean score was 64 percent correct (SD = 12.9), and observations ranged between 36 and 97 percent correct. No significant differences were found between the conditions, irrespective of mortality.

Written proficiency

The free written response data gathered to measure implicit knowledge (see 3.3.3) have also been used to determine three different components of written proficiency: fluency, grammatical accuracy and spelling ability. Fluency expresses how comfortable second language learners are in producing written text, and grammatical accuracy refers to freedom from error in relation to native-like use (Wolfe-Quintero, Inagaki, & Kim, 1998). Finally, spelling refers to the ability to form words correctly in the L2. These three measures were included as measures expressing specifically the students’ ability to write in their L2.

The three measures have been calculated on the basis of all free written response task utterances at T0. Following recommendations by Wolfe-Quintero, Inagaki, and Kim (1998), fluency was calculated by counting and averaging the number of words per clause realized per student. The mean fluency score was 5.5 words per clause (SD = 0.5), and scores ranged from 3.8 to 6.9. Grammatical accuracy was obtained by counting the number of grammatical errors per clause. For example, if there was no subject-verb agreement, this would count as one error. Spelling errors were not included in the count; nor were errors in using the
target structures included, as these were intentionally elicited. The mean number of errors per clause was 0.84 (SD = 0.5); scores ranged from 0.015 to 2.04. Finally, spelling was calculated as the number of spelling errors per clauses. On average, 0.31 (SD = 0.21) spelling errors per clause were made, and scores ranged between 0.02 and 0.9. Again, no significant differences were observed between the conditions, irrespective of mortality.

3.4.5 Exposure to the second language

In order to gain insight into the educational background of the participants and the amount of exposure they had had to the Dutch language, a questionnaire was submitted to their teachers. The teachers were asked how much contact the students might have with the Dutch language, and how long their students had been in The Netherlands. There were significant differences between the three conditions.

The amount of exposure outside the ISK was assessed by asking teacher to estimate how much they thought their students would use Dutch outside school. They were given the following options to choose from: often, fairly often, fairly little, and hardly. Data were obtained for 92 participants. According to their teachers, eleven participants used Dutch often; 44 percent used Dutch fairly often; 38 percent used Dutch fairly little; and 6 percent were said to hardly use any Dutch outside school. The Kruskal-Wallis statistic was used to assess how the observations were distributed across the three treatment conditions. A significant difference was observed: $X^2 = 11.99$, df = 3, $p. = .00$, $N = 92$. This significant difference is caused by the NI group: they used Dutch less than the other two groups, and if they were excluded from the analysis, no significant differences were observed.

Another indication of exposure to Dutch is the length of the students' stay in the Netherlands. Teachers were asked to give the date of arrival in the Netherlands. Using this date, the number months until the pre-test date was calculated. Data were obtained from 84 participants. Their mean length of stay was 21.2 months (SD = 7.8). The observations ranged substantially: one student had been in The Netherlands for just 8 months, while another had been in The Netherlands for 46 months already. A one-way ANOVA was run to test for significant between-group differences: there were none for the total sample. However, for the sample that was followed for three months, there were significant between-group differences: $F(2,63) = 5.24$, $p. = .01$. Again, the NI
group was the odd one out, its participants having spent considerably more time in The Netherlands than the participants from the experimental conditions.

3.5 Assessment of individual characteristics

3.5.1 Introduction

One of the goals of this study is to evaluate how differences between learners affect explicit and implicit learning. A number of individual difference (ID) variables have been found to affect L2 learning, and will therefore be evaluated in this study. They are: developmental readiness, L1 similarity, age, aptitude, learning style, and motivation. Each of these factors has been assessed; this section explains how. It starts with developmental readiness (3.5.2), then continues with L1 similarity (3.5.3), Age (3.5.4), Aptitude (3.5.5), and ends with the affective variables learning style and motivation (3.5.6).

3.5.2 Developmental readiness

Developmental readiness refers to the phenomenon that grammar may only be effectively instructed if the L2 learner is in a particular stage of grammatical development. A complicating issue is that there is no evident way to operationalize the notion of ‘developmental readiness’ without extensive prior research. An important reason is that it is simply unclear what underlies the notion of developmental readiness (see 2.2.3): constraints specific to linguistic processing, general processing capacity limitations, or input features. As there is no clear theoretical motivation to operationalize developmental readiness, determining sensitivity to instruction has to be based on the participants’ interlanguage. In FFI research, several approaches have been used, but all are based on emergence of correct use. Some have used previous research to define stages in the acquisition of their target structure (e.g., Pienemann, 1989; White, 1998). Another approach has been to monitor L2 use in formal and informal situations, where differences in use of the target structure between the two situations indicated readiness (e.g., Bienfait, 2002: see also 2.4.4 and 2.4.5). Finally, emergence of correct use in the pre-test has been taken to be an indication of developmental readiness (e.g., Williams & Evans, 1998). As there is no previous research to base stages of acquisition on, and because monitoring
formal and informal language use was considered to be too labour-intensive, the latter approach has been adopted.

Emergence of correct forms of the target structure in the participants’ use of Dutch will be taken to indicate developmental readiness. Naturally, readiness has to be determined for each target structure independently. A two-level nominal variable will be created for each target structure based on whether correct forms of the structures were used at T0. The group of students that already do use correct forms of the target structures are considered developmentally ready, while non-use is taken to indicate unreadiness. When selecting and assigning students to the treatment conditions, developmental readiness was not taken into consideration. In other words, there was no matching, and possible effects of developmental readiness are assessed post hoc.

3.5.3 First (and second) language background

As pointed out, the students of ISK schools are quite diverse with respect to their first language (L1) background. Due to a rapid decline in student population of ISK schools, the intention to select students with similar L1 backgrounds could not be fulfilled. In fact, this study now includes speakers of 33 different languages. The most frequently occurring language was Portuguese as spoken in Angola: it was the L1 of 23 students. There were no other languages with more than ten speakers present in the subject pool. Also, only five out of 101 participants spoke a Germanic language as L1: English. The speakers were almost exclusively originating from Asian or African countries. Eleven participants were from Europe: four of them spoke Russian, four spoke Turkish, there were two speakers of Polish, and participant spoke Bulgarian.

One of the goals of this study is to assess whether effective FFI depends on a student’s L1. To enable answering this question, particular features of the L1 were coded that might affect the students’ L2 proficiency. For the degrees of comparison, the participants’ L1 was coded for how they express comparison: morphologically or periphrastically. If the language exclusively allowed for periphrastic degrees of comparison, it would be coded as marked. Otherwise, it was coded as unmarked. A Chi-square analysis revealed no significant differences in the distribution of marked and unmarked L1s between the three conditions: \( \chi^2 = 4.60, df = 2, p = .10, n = 98 \). The trend towards significance here is caused by the control group (for the delayed group: \( \chi^2 = 3.42, df = 2, p = .18 \)). Because Dutch requires inversion in subordinate clauses, the L1 was coded for their default word order (svo or sov), and whether it has similar inversion in
subordinate clauses. One language – Armenian – seems to have similar inversion in subordinate clauses: there are three speakers of this language in the subject pool. Given the small number of speakers, this similarity was ignored. A Chi-square analysis on the distribution of svo and sov languages did not reveal significant differences between the three conditions: $X^2 = .74$, df = 2, p = .69, n = 98. (for the delayed group: $X^2 = 7.35$, df = 2, p = .03)

The participants were also asked if they spoke any other second languages. Forty-nine students reported speaking an L2 besides Dutch. These languages were mostly regional languages or French or Spanish as spoken in former colonies. The participants were also asked to assess their command of this L2 as compared to their current level of Dutch. Twenty subjects reported speaking their L2 better than they did Dutch. None of the thirteen students that reported a Germanic language as their L2 (all English) spoke it better than Dutch. Three reported it being equal; the other ten indicated being worse in English than in Dutch. Effects of the students’ L2 have not been investigated.

### 3.5.4 Age

Age is another factor that has been found to affect L2 acquisition, and may influence the effectiveness of FFI. Age has been measured in years, and the participants of this study range somewhere between 12 and 18 years old. In fact, 85% of all participants fall within the 13-16 interval: five subjects were twelve years old, seven were 17, and three subjects were 18 years old. Treating age as an ordinal variable, the Kruskal-Wallis statistic for independent samples was used to test how age is distributed across the three conditions. There were no significant differences: $X^2 = .69$, df = 2, p = .71, N = 101.

### 3.5.5 Aptitude

For this study, two components of aptitude were assessed, both based on tests from the Modern Language Aptitude Test (MLAT) battery developed by Carroll and Sapon (1959). In this case, rote memory and grammatical sensitivity (GS) were assessed. Both aptitude components were assessed by means of tests developed for the occasion, as the standardized Dutch form of the MLAT – the Verbale Aanleg Test (Drenth & Van Wieringen, 1969) – was considered to be too difficult for the present subject sample.

In the original, the rote memory test requires the language learner to rote-learn Kurdish words. Since it could not be excluded that speakers of Kurdish or related languages might participate in this study, the test was administered using
Swedish words to learn. None of the participants had to be excluded because of prior knowledge of Swedish. In addition, the test procedure was somewhat simplified as compared to the original: it involved a learning session of 2 minutes, in which the participants were asked to memorize the Dutch meaning of sixteen Swedish words. Immediately afterwards, they were given a sheet containing the Swedish words, and they had to give the Dutch meanings. An item was scored correct if the student reproduced the correct meaning: spelling errors were ignored. Ultimately, memory scores were obtained from 97 of 101 subjects. Reliability was found to be sufficient (Cronbach's alpha = .80, N = 97). The mean score was 8.9 (SD = 3.7), and the scores ranged from 0 to 16. The test can be found in Appendix C2.

The grammatical sensitivity (GS) test assesses the ability to match phrases that have the same grammatical functions in pairs of sentences. In its original form, the GS component is administered in the test taker's L1. However, in the light of the different L1’s of the participants of this study, it was administered in Dutch. Robinson (1997) similarly employed the MLAT, and did not find significant correlations between the MLAT and general proficiency, which is an indication that the MLAT scores were not influenced by L2 proficiency. The test consisted of fifteen items, and an item was scored correct if the grammatical function of the example sentence phrase and the test sentence phrase matched. For this test, there were quite some missing values, as it was the last test to be administered in a row of tests: some participants did not take this test due to lack of time. Scores were obtained from 75 students, and the test was found to sufficiently reliable (Cronbach’s alpha = .81, N = 75). The mean score was 6.7 (SD = 3.8), and the scores ranged from 0 to 15. The test can be found in Appendix C3.

One way ANOVAs were run to assess whether there were differences between the conditions for either memory or GS. No significant differences were found, irrespective of mortality.

3.5.6 Affective variables

A number of personality factors or affective variables may influence the ability to learn. An attempt was made to gain insight into the participants’ personalities by means of a teacher questionnaire. Teachers were asked to judge their students on seventeen aspects relating to attitude, aptitude, motivation, structure-dependence and precision by means of five-point Likert scale items. The questionnaire can be found in Appendix C4.
The results of this questionnaire were subjected to a principle components factor analysis to investigate whether particular sets of items could be combined to represent personality factors such as motivation, aptitude, etc. The data resulting from this questionnaire turned out to be suitable for such an analysis (KMO measure of sampling adequacy = .84, p. < 0.01). A number of items were removed first, though, because they did not correlate sufficiently with other items, which is a requirement for factor analysis. In Appendix C4, the questions marked with an asterisk were removed from the analysis. Once these items had been removed, all the conditions for performing a factor analysis were met. Oblique rotation – to be used when the factors to be extracted may be related to each other – was used to extract the factors. In this case, there are indeed no theoretical grounds to state that the factors should be uncorrelated. The results of the factor analysis can be found in Table 3.9. The outcome was a solution involving three factors with eigenvalues greater than one, and explaining in total 69 percent of the variance. Table 3.9 provides the details.

The resulting factors turned out to be well-interpretable. The first factor has been called 'cognitive style', and it expresses one's operating style; whether one

<table>
<thead>
<tr>
<th>Factor</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>5.13</td>
<td>1.86</td>
<td>1.31</td>
</tr>
<tr>
<td>Explained variance</td>
<td>42.72</td>
<td>15.52</td>
<td>10.94</td>
</tr>
<tr>
<td>Cumulative variance</td>
<td>42.72</td>
<td>58.24</td>
<td>69.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contributing items</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1:</td>
<td></td>
</tr>
<tr>
<td>Cognitive style</td>
<td></td>
</tr>
<tr>
<td>Precise vs. chaotic</td>
<td>.60</td>
</tr>
<tr>
<td>Well-considered (vs. impulsive)</td>
<td>.82</td>
</tr>
<tr>
<td>Intellectual (vs. emotional)</td>
<td>.65</td>
</tr>
<tr>
<td>Independent of others</td>
<td>.74</td>
</tr>
<tr>
<td>Factor 2:</td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
</tr>
<tr>
<td>Enjoys school</td>
<td>.86</td>
</tr>
<tr>
<td>Motivated to learn Dutch</td>
<td>.71</td>
</tr>
<tr>
<td>Likes to speak Dutch</td>
<td>.84</td>
</tr>
<tr>
<td>Factor 3:</td>
<td></td>
</tr>
<tr>
<td>Aptitude</td>
<td></td>
</tr>
<tr>
<td>Has a feeling for language</td>
<td>.91</td>
</tr>
<tr>
<td>Understands grammar instruction easily</td>
<td>.90</td>
</tr>
<tr>
<td>Remembers things easily</td>
<td>.80</td>
</tr>
</tbody>
</table>
adopts a precise and well-considered approach to things to be undertaken, or a more chaotic and impulsive approach. Independence and intellect (which was contrasted with emotion) also loaded on this factor, which seems plausible. The second factor clearly represents the students’ motivation to be in school and learn Dutch. Finally, the three questions that relate to aptitude were found to load on the third factor.

One way ANOVAs were used to investigate whether the three conditions differed significantly with respect to either one of these factors. For none of the three factors, however, significant differences were observed, again irrespective of mortality.

3.6 Procedures and data processing

3.6.1 Introduction

In this section, matters of research organisation are discussed. The procedures and steps that were taken with regard to the collection of the data are described in 3.6.2. Next, in 3.6.3, a description is given of how the instruction was practically organised at schools. Issues of coding and data processing are dealt with in 3.6.4; and the section is concluded with a discussion of the statistical procedures followed, in 3.6.5.

3.6.2 Data collection

All data have been collected over a period of three to four months. The first test session took place at T0, immediately preceding the instruction. The second test session (T1) would immediately follow the instruction, and the third session (T2) took place approximately two months after T1. It should be pointed out that the pre-test was not always immediately followed by the instruction. In some cases, technical problems with the computer programme led to a delay of two or three weeks before the students started with the instruction. Table 3.10 provides an overview of the tests and times of administration. As can be seen, the free written response task and the grammaticality judgement test were administered at all three times of testing. At T0, three additional tests were administered: the paired associates test, the grammatical sensitivity test, and the c-test. At T1, the controlled production test was administered as an extra (see 3.3.2).
All tests were administered in class by the researcher, mostly in absence of the students’ regular teacher. The students were told that they participated in a study into the acquisition of Dutch as a second language, and that their Dutch would be monitored for this purpose. In addition, they were reassured that the information would be used for research purposes only, and that they did not serve any additional purposes related to their school career. No mention was made of grammar or the specific grammar structures in focus. Nor was the computer programme mentioned that was part of the research. In a number of cases, however, their teacher had explained that the testing and their computer work were related.

At T0, a total of five tests were administered. These tests were administered in a fixed order. The first test would always be the paired associates test, because it involved strict timing. The students were challenged to try and remember as many words as they could in two minutes time. After this time, the students had to turn around their paper, and they were asked to give the correct Dutch meanings of sixteen Swedish words. They were allowed to think for as long as they needed, but it generally did not take much more time than five minutes to fill in the meanings. The next test was the free written response task, which was always given before the grammaticality judgement test. The test was introduced as an assessment of their writing proficiency, and the students were instructed to respond appropriately to the situations given to them. They were told explicitly that they were free to be creative in their response: all answers would be considered correct, as long as they took care of their Dutch.

The free written response test would be followed by the grammaticality judgement test, the c-test, and the grammatical sensitivity test. Because there were substantial differences between students in the time it took them to finish the free written response task, students did not have to wait for the rest to finish. As soon as they finished the test, they would continue with the grammaticality judgement test. This way, no students ever had to wait and perhaps take the opportunity to distract their classmates. As a result, the grammaticality judgement test, the c-test and the grammaticality judgement test were not explained in class, but individually. Explaining the grammaticality judgement test and the c-test proved rather straightforward, and did not lead to problems. Care was taken, though, to make sure students would underline the error in the grammaticality judgement sentences. The last test to be taken was the grammatical sensitivity test. This test was most difficult to explain. It was presented as a puzzle that students needed to solve on the basis of the examples
(see Appendix C3). The examples were all read out and discussed, but no clues were given as to why phrases matched. All in all, administering these five tests would take approximately two hours.

At T1 and T2, testing took much less time, as the participants only took three (T1), or two (T2) tests. The order of the tests was the same: so the free written response task was administered before the GJ task and the controlled production task. At this time, students were told that they would do the same tests as before. In addition, they were told that the results would be compared in order to investigate how much they learned, and they were encouraged to show they had learned. They understood why the tests were the same, and did not object to this.

Finally, it should be pointed out that a strong dissociation was created between testing and instruction. All the tests taken by the students were paper and pencil tests, and they were administered by the researcher at all three times of testing. The instruction was delivered by the computer, and the students’ regular teacher would accompany them during their computer work. In addition, although the computer instruction did contain exercises resembling grammaticality judgements, the instruction exercises and the test exercises were different. This way, effects of task familiarity were hoped to be reduced to a minimum.

### 3.6.3 Organisation of the instruction

As indicated, all instruction was delivered by means of a computer programme called *Taal in Themas*. The programme was created for the occasion by means of Authorware (1999) software especially designed to develop e-learning applications. The programme was designed to work on stand-alone computers, but has also been used successfully on school networks. It was distributed by
means of a CD-rom, and would be installed by the schools’ system managers on
demand of the participating teachers.

A special feature of the programme was its database connectivity. This enabled
random assignment to the treatment groups. Microsoft Access databases were
used for this. Before the start of the instruction, the teacher would provide the
names of the participating students, which were then entered into the database.
At this point, the students were randomly assigned to either treatment groups
one or two (see 3.2.2). Also, they would be assigned their own login number.
During the instruction, the programme would interface with this database and
record student performance. When all students had finished the instruction, the
school’s system manager would return the databases by email.

Once the programme was installed and the students had been pre-tested, the
instruction started. Students would work for one hour per week with the
computer programme. At some schools, they worked two hours per week.
However, it also occurred that students skipped a week because of previous
engagements in the school curriculum. The students worked individually on
separate pc’s in computer classrooms, and their regular teacher would accompany
them. Students were not told that there were two versions of the programme, and
their neighbour might actually be doing grammar exercises about a different
grammar structure. Most never found out, but when they did, the teacher would
explain. To start the programme, students had to enter their login number. On
first time login, students received an introduction about the programme and its
workings. As all progress was recorded, students could stop at any point during
the programme: on subsequent login, the programme automatically returned to
the point the student stopped. On average, three one-hour sessions were needed
to pass through the entire programme. As soon as all students had finished, the
immediate post test was administered.

3.6.4 Coding

Each item of the free written response test was coded for a number of
different aspects. First of all, actual use was coded, and whether this use was
correct, incorrect or clearly avoided. In addition, the way the structure was
realized in the utterance was also coded in order to assess whether correct use
depended on the structure’s realization.

For the DoC, three different aspects of structure realization have been coded.
First, an obvious feature of use that was coded was the actual degree used for
### Table 3.11 Examples of DoC use

<table>
<thead>
<tr>
<th>Use</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct use</td>
<td>Omdat nog kleiner is.</td>
</tr>
<tr>
<td></td>
<td>De meneer wordt steeds dikker</td>
</tr>
<tr>
<td></td>
<td>Ik kies de goedkoopste en mooiste</td>
</tr>
<tr>
<td></td>
<td>De ene is duner dan de tweede</td>
</tr>
<tr>
<td></td>
<td>Ik vind deze mooier dan die andere</td>
</tr>
<tr>
<td></td>
<td>Ik kies de mooiste broek, met de mooiste kleur</td>
</tr>
<tr>
<td></td>
<td>Ik ga eerst kijk welk is de mooist</td>
</tr>
<tr>
<td></td>
<td>Hij heeft het snelste tijd</td>
</tr>
<tr>
<td>Incorrect use</td>
<td>Dan hij ga steeds dik worden</td>
</tr>
<tr>
<td></td>
<td>hij is een beetje oude</td>
</tr>
<tr>
<td></td>
<td>omdat die veel groot en beter is</td>
</tr>
<tr>
<td></td>
<td>Bril 1 is mooi dan bril 2</td>
</tr>
<tr>
<td></td>
<td>Want op de bank is veilig dan op de vloer</td>
</tr>
<tr>
<td></td>
<td>Ik kies de nieuw en de mooi broek</td>
</tr>
<tr>
<td></td>
<td>Want is de goedkoopster auto</td>
</tr>
<tr>
<td></td>
<td>Omdat is de mooie</td>
</tr>
<tr>
<td></td>
<td>Want is de sportief auto</td>
</tr>
<tr>
<td>Avoided use</td>
<td>Hij wordt meer dik</td>
</tr>
<tr>
<td></td>
<td>Hij is meer groot geworden en sterk</td>
</tr>
<tr>
<td>No use</td>
<td>Omdat hij gaat snel</td>
</tr>
<tr>
<td></td>
<td>Omdat het is groot voor ik en mijn 3 kinderen</td>
</tr>
<tr>
<td></td>
<td>Die heeft meer garantie [garantie] dan radio 2.</td>
</tr>
<tr>
<td></td>
<td>The man can become sick</td>
</tr>
</tbody>
</table>

Comparison: either comparative or superlative degree. Second, the syntactic function was coded. Adjectives and their inflected forms can appear either attributively or predicatively (Aarts & Wekker, 1993). When used attributively, adjectives modify a noun phrase (de snelste auto: ‘the fastest car’), while predicative adjectives function as subject or object attributes (die auto is het snelst: ‘that car is the fastest’). If comparatives were used predicatively, the presence or absence of a comparative clause would also be coded (Deze radio is groter dan de andere: ‘This radio is larger than the other’). The third and last aspect of realization to be coded was the adjective itself.
Coding decisions for the degrees of comparison were rather straightforward. If the participants used the degrees of comparison, they would mostly use them correctly. In Table 3.11, examples of correct use can be found. As soon as -er or -st was properly attached to the adjective to express comparison, an item would coded as correct. The schwa-inflection that Dutch adjectives receive in particular circumstances (een groter huis; het grotere huis) was ignored, as this inflection is not related to the use of the DoC. Errors in spelling were coded, but they did not affect the decision of whether a form would be coded as correct or incorrect.

Incorrect use was infrequent. Rather than producing incorrect forms, students would simply neglect to use the DoC. Sometimes, though, they would provide a context in which the use of DoC is obligatory. For example, the Dutch adverb steeds requires a comparative form. An error that was actually quite frequent was the omission of the comparative marker in combination with comparative clauses. Omission of a form of the DoC while the context clearly demands this would be marked as incorrect. Examples can be found in Table 3.11. A special case of omission of DoC markers is avoidance of use by means of the adverbs meer and meest (more and most). Because these adverbs betray a clear intention to express comparison, such cases would be coded as avoidance (see Table 3.11 for examples). They occurred infrequently. Items would be coded as no use in case no form of the DoC was used. In these cases, there is simply no telling as to whether the student intended to express comparison or not. If the question was left blank, it was also coded as no use.

A final remark is required with respect to coding the use of the DoC. It is important to point out that incorrect use of the DoC may go unnoticed. Quite frequently, students were found to omit the comparative suffix in phrases containing a comparative clause: “one is small than the other”. In the process of acquiring the DoC, L2 learners apparently fail to mark the adjective for comparison. When comparative clauses are used, such omissions can be noticed and marked as incorrect. However, if one does not use a comparative clause, there is no telling what someone writing “one is small” actually intended to say. Such phrases were coded as ‘no use’. Given that omission was quite frequent with comparative clauses, it is likely that ‘incorrect use’ has gone unnoticed, and that incorrect use scores are therefore an unfair reflection of real incorrect use.

For SubC, three different aspects of the structure’s realization have been coded. First, the type of SubC has been coded; or, in practice, the subordinate conjunction used. The second aspect pertained to verb phrase complexity. A distinction was made between single-verb phrases consisting of be or have
<table>
<thead>
<tr>
<th>Use</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Correct use| Als ik iets grappigs zie of hoor  
              Als er een grapje gezegd wordt  
              Wanneer je blij bent  
              Omdat die mooi is  
              Omdat het nat wordt door de regen  
              Omdat die in minder tijd heeft gewonnen  
              Omdat de trein zonder te stoppen gaat  
              Zodat de olifant niet meer kan storen  
              Doordat ze op een ijs staan |
| Incorrect use | Als je doet een sport  
                        Als ik wil iets kopen  
                        Als ze roeelt over een andere vriend van mij  
                        Omdat is goedkoper is  
                        Omdat op de grond heb je meer plaats dan op de bank  
                        Omdat de olifant kan niet meer de fruiten pakken  
                        Omdat Jan heeft 7 minut gedaan  
                        Omdat hij kan goed luisteren |
| Avoided use | Want is goedkoopste  
                      Want hij wil fruit uit de boom halen  
                      Want zij is bang |
| Not analysable | Als je rent  
                        Omdat ze willen slapen  
                        Omdat ze wil dat iederen haar hoort |
| No use      | De ene vogel valt door de liefde  
                        De olifant eet de appels  
                        Deze auto is iets groter dan de andere twee autos  
                        Boes wil mol slaan |

auxiliaries; single-verb phrases other than consisting of *be* or *have* auxiliaries; and multiple-verb phrases. Finally, clause complexity was also considered by way of the verb phrase complements realized. Three types of clauses have been distinguished: adjectives; direct objects, either with or without attributive adjective; and all other, more complex types of complements.
For SubCs, coding decisions were also not too difficult. When the verb was properly placed at the end of the subordinate clause, it would be considered correct. Table 3.12 provides examples. Incorrect SubCs were very frequent, and errors always consisted of using the default main clause word order. Causality was regularly expressed by means of the coordinate conjunction want instead of the subordinate conjunction omdat, which would be coded as avoidance. It should be noted that although the Dutch may have preferences to use either want of omdat in particular contexts, the use of want is perfectly grammatical. Utterances can only be judged for their correctness if the predicate contains a complement. In cases there was no complement realized or when the complement was a relative clause, the utterance would be coded as not analysable. Finally, if no subordinate conjunction was used, or if the question was left blank, the utterance was coded as no use.

3.6.5 Statistical procedures

A number of different statistical methods have been used to answer the research questions posited in 3.1. The first research question pertaining to the nature and development of implicit grammatical knowledge was addressed by examining the free written response task data. In the previous subsection, the contexts in which the target structures may appear have been identified. By means of descriptive statistics, the use of the target structures in each of these contexts has been characterized. In addition, a comparison of percentages of correct use provided insights into the relationship between correct use and context of appearance. No inferential statistics were used.

The second research question addresses the effect of different kinds of instruction on the development of explicit and implicit grammatical knowledge. The independent variable ‘instruction’ was operationalized as a three-level between-subjects variable (EI, II, and NI). Two dependent variables were used: the grammaticality judgement scores and the free written response task scores, both obtained at three different points in time. Univariate analysis of variance with repeated measures was used to test the hypotheses related to the third research question. The analyses were performed separately for both target structures and both dependent variables.

To answer the third research question, two different statistical techniques were used, depending on which variable was addressed. In addition, rather than using three-level within-subjects dependent variable for grammatical
development, gain scores were computed for explicit and implicit grammatical progress in both target structures. Developmental readiness and L1 similarity were both defined as two-level between-subjects variables. Therefore, univariate analysis of variance was used to explore for each of these two variables separately how they related to explicit and implicit gain. The remaining ID variables (the aptitude measures, cognitive style, motivation, and age) were all operationalized as interval variables. Therefore, correlation analyses were used to explore the relationship between these ID variables and explicit and implicit gain.

The fourth and final research question has not been addressed statistically, because the nature of the grammar structure has not been operationalized as a statistical factor. The research design actually involved two parallel experiments and does not logically allow for statistical comparison of the two structures. In addition, the unequal scales of measurement for both the grammaticality judgement task and the free written response task were another reason not to introduce the kind of grammar structure as an additional statistical factor. Thus, the question was simply answered by accumulating and comparing the results on each of the previous research questions.

Statistical assumptions

Normally, analyses of variance require that the assumption of homogeneity of variance be met. However, in repeated measures designs, which involve comparing data from the same subjects over time, not only variances across conditions should be equal, covariances between pairs of conditions need to be as well. This is referred to as compound symmetry or sphericity (Field, 2000). If the assumption of sphericity is violated, there are a number of corrections that can be applied to the data. In this study, the most conservative correction will be reported in case of a violation of sphericity: the Greenhouse-Geisser estimate.

Analysing skewness and kurtosis values for each measure demonstrated that the data were not always normally distributed. In fact, at T0, all distributions were positively skewed. Skewness diminished at T1; and at T2, there were no significantly skewed distributions. As the participants had been selected to have little knowledge of the target structures, a positively skewed distribution at T0 is hardly surprising. Violations of kurtosis occurred as well, but not very often. Although these violations were unfortunate, analysis of variance is robust against such violations (Van den Berken & Voeten, 2002), especially if these violations are not too grave, as was the case here.
Post hoc analyses

If necessary, post hoc tests were carried out for a deeper investigation of found effects. For these analyses, the post hoc tests available in SPSS’s GLM repeated measures procedure were not used, as they do not compare differences between group means at each point of measurement. Group means are conflated over time, which makes for an uninteresting comparison between groups. In addition, there are technical issues. The error term used in post hoc repeated measures analyses to compute significance is based on the mean error of all cells in the design. This is problematic in case of violations of equality of variance. For these reasons, post hoc analyses were performed for each point of measurement separately, using one-way ANOVA’s comparing the groups one-by-one. In case of significant differences, effect sizes (Cohen’s $d$) were reported.