Measuring Receptive Vocabulary Size

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone : +31 30 253 6006
fax : +31 30 253 6000
e-mail : lot@let.uu.nl
http://www.lotschool.nl

Cover illustration: Drawing by Sir John Tenniel and quote by Lewis Carroll. From: Alice in Wonderland and Through the Looking Glass, © Wordsworth Editions Ltd 1993.

ISBN 90-76864-56-X
NUR 632

Measuring Receptive Vocabulary Size

Reliability and Validity of the Yes/No Vocabulary Test for French-speaking Learners of Dutch

Een wetenschappelijke proeve op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
op gezag van de Rector Magnificus Prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op maandag 5 juli 2004,
des namiddags om 3.30 uur precies

door

June Eyckmans

geboren op 28 februari 1972
te Mechelen

Promotor :	Prof. dr. R. van Hout
Co-promotores :	Dr. H. Van de Velde (Universiteit Utrecht/Université Libre de Bruxelles) Dr. F. Boers (Erasmushogeschool Brussel/Universiteit Antwerpen)
Manuscriptcommissie :	Prof. dr. C.L.J. de Bot (Rijksuniversiteit Groningen, voorzitter) Prof. dr. A. Janssen-van Dieten Prof. dr. P. Meara (University of Wales Swansea)

Voor Renaud

Voorwoord

Een tijdje geleden maakte iemand me de opmerking dat één van mijn kwaliteiten is dat ik me weet te omringen met zeer competente mensen. Daar ben ik het volmondig mee eens. Ik heb dan ook een aanzienlijk lijstje van personen die ik graag eens in de bloemetjes zou willen zetten nu mijn proefschrift beëindigd is. Laat ik beginnen bij de "onderzoeksbeesten". Voor diegenen die het begrip niet kennen: het verwijst naar een bijzondere variëteit van de menselijke soort die je weleens pleegt tegen te komen in onze contreien. Je herkent ze aan hun feilloze neus voor ambitieuze onderzoeksprojecten, hun scherpzinnige analyses en overmatige werkdrift. Ze zijn volhardend en gedreven voor hun vak, ze bezitten buitengewone "problem solving skills" en ze worden wild van data-analyses. Ik reken mezelf niet tot de soort maar ik vertoef graag in hun gezelschap.

Hans is zo'n "onderzoeksbeest" in hart en nieren en de rol die hij heeft gespeeld in dit onderzoeksproject is groot. Tijdens mijn jaren aan de Université Libre de Bruxelles loodste hij me van het ene contract naar het andere en deed al het mogelijke om in de eerste plaats mijn werkgelegenheid te verzekeren, een rol die Frank nadien met veel glans overnam. Hij is ook de man die, waar ook ter wereld hij zich bevond, nooit onbereikbaar was. We zagen elkaar niet veel, maar naar mijn aanvoelen was hij er altijd. Na het overlijden van Renaud heeft hij menig uur in de half-afgewerkte data-analyses van dit onderzoek gewroet. Zonder zijn inzet en standvastigheid zou dit boek er niet zijn. Roeland behoort ook tot de bovenvermelde variëteit. Het voorbije jaar leerde ik hem kennen als een efficiënt wetenschapper en een uitmuntend statisticus. Hij maakte de nodige financiële middelen vrij waardoor ik meer tijd kon wijden aan het schrijven en ik wil hem graag bedanken voor zijn bereidwilligheid om een project te begeleiden waarvan de krijtlijnen reeds lang waren uitgezet. Ik ben er mij van bewust dat het een ondankbare taak was. Frank is de derde in het rijtje "onderzoeksbeesten". Een conceptueel wetenschapper van formaat met een geweldig gevoel voor zelfrelativering en humor. De man die productief is voor twee, de workaholic die ik nooit zal kunnen zijn. Aan hem "a heart-felt thank you" voor het verbeteren van de Engelse tekst, het nalezen en de rustige stem in paniekerige momenten. Hij heeft me meermaals getoond dat het, temidden van het relatieve en het absolute van de wetenschap, toch de moeite waard bleef om steeds de volgende stap te zetten.

Ik wil ook graag mijn vroegere collega's van de ULB bedanken voor hun bemoedigende woorden, in het bijzonder Inez, Vera en Alain. Inez stond altijd klaar om me alweer uit een nieuwe "excel-impasse" te halen, de formule "index/match" in de aanslag! Vera wekte mijn interesse voor

woordenschatverwerving en is in vele opzichten mijn petemoei in de wetenschap. Na het aanhoren van mijn ergernis over het traag opschietende schrijfproces, gecombineerd met de eindeloze verbouwingsmalaise en het woekerende verdriet over verloren "compagnons", bracht ze aan mijn verstand dat er een verdienste school in het "leren roeien met riemen van stro". Alain was nabij in moeilijke tijden, luisterde naar eindeloos geweeklaag, ploeterde door teksten, matrices en bibliografieën en gaf me geen ruimte om aan mezelf te twijfelen. Ook de waardering voor mijn werk die ik kreeg van collega's in het buitenland tijdens de vele congressen heeft me erg aangemoedigd. Ik denk hierbij met een warm hart aan Jan Hulstijn, Anne Vermeer en Paul Meara.

Natuurlijk hebben ook mijn familie en vrienden op een belangrijke manier bijgedragen tot het voltooien van dit project. Mijn vader wil ik bedanken voor "het geloof in eigen kunnen" dat hij zijn kinderen van jongsaf heeft meegegeven. Net als mijn zus, ben ik grootgebracht met de boodschap dat onze mogelijkheden onbegrensd waren. "The sky is the limit", was een veel gehoord adagium en ons talent voor volharding danken we zonder twijfel aan hem. Ook mijn grootmoeder hoort hier thuis omwille van de warme zorgzaamheid waarin ze me steeds heeft gekoesterd. Hoe spijtig dat ik haar dit boek niet meer kan tonen. Al was het in het Chinees geschreven, ze had er zeker haar leesbril voor bovengehaald. Een speciale vermelding voor mijn moeder, die de kunst verstond om me aan te porren maar tegelijkertijd nooit mijn welzijn uit het oog verloor. De manier waarop ze zorg voor me droeg en draagt, verdient niks minder dan een cum laude. San en Ralph haalden me uit de diepste dalen. Ze brachten stilte in mijn tumult en hoop in mijn wanhoop. Tegelijkertijd zorgden ze vaak voor de vrolijke noot. De Brusselse vriendinnen zorgden eveneens voor vertier en polsten regelmatig naar de vorderingen. Van de vriendenploeg uit café Lipstick, hebben o.a. Herwig, Bart en An me meermaals een hart onder de riem gestoken. Annicks aanwezigheid in Weert bestreed de eenzaamheid en zorgde voor troost. Ze schotelde me regelmatig een "echte" maaltijd voor en keek meewarig naar mijn chaotische manier van werken. Haar hulp bij de grafische vormgeving van dit boek was bovendien onontbeerlijk. Ook Bart vermeld ik nog omdat het hem werkelijk nooit heeft kunnen schelen of ik dat proefschrift nu schreef of niet en dat bedoel ik op een uitermate positieve manier. Hij is misschien de enige die nooit zou be-, ge- of veroordeeld hebben als ik dit werk niet had voltooid. Toen hij er nog voor me was, was hij er met hart en ziel.

Tot slot, Renaud, de laatste der "onderzoeksbeesten" en boven alles een dierbare vriend. Aan hem draag ik dit boek op. Woorden schieten tekort om zijn bijdrage tot dit onderzoek uit te drukken. Hij leerde me alles wat ik weet over taaltoetsing en de *Yes/No Vocabulary Test* was zo'n beetje ons gezamenlijk zorgenkind geworden. Ik denk met pijn in het hart terug aan onze chaotische werkpret en de uitdijende gesprekken over wetenschap, leven, liefde en maatschappij. Dit alles op zijn tijd doorspekt met "une blache et une

cigarette", zelfs toen het al lang niet meer mocht. Hij was een man van uitersten. De controlefreak in slobbertrui. De creatieve psychometrist. De nonchalante "m'enfoutist" die bleef broeden op probleemstellingen wanneer iedereen er al lang schoon genoeg van had. De man die nooit voor een publiek wou spreken hoewel hij de meest complexe materie kon omzetten in zulke klare taal dat je de deuren in je hoofd voelde opengaan. Een ongeëvenaard methodoloog ook. In het interpreteren van resultaten of het uitdenken van nieuwe werkhypothesen liet hij iedereen ver achter zich. Begeesterd voor en door de wetenschap. "La démarche scientifique est de pouvoir vérifier ce que prétendent les théories", ik hoor het hem nog zeggen. Hoe ironisch dat diezelfde wetenschap hem niet heeft kunnen redden. Na zijn overlijden bleef ook dit onderzoeksproject verweesd achter en het plezier dat ik er voordien aan beleefde, sloeg om in een tocht door de woestijn. Het was erg moeilijk om het zonder hem te (willen) vervolmaken en mijn bezigheden leken alsmaar in het niets te verzinken in het licht van het sterfelijke. Maar kijk Renaud, het is volbracht.

Contents

Chapter 1

Introduction

Today, no teacher or researcher would contest the importance of the lexical dimension in second language learning. Everyone who has ever learned a foreign language will agree that vocabulary knowledge is a prerequisite for the development of any form of language proficiency. The relatively recent revaluation of the lexical dimension in language learning coincides with the shift in perspective from defining foreign language learning as primarily involving "top-down processing" (language learning is essentially learning grammar rules and applying them to concrete examples) to perceiving it as being driven by "bottom-up processing" skills (recognizing and acquiring frequent word combinations – also called "chunks" - from which more general patterns can be extracted) (Nattinger and DeCarrico 1992, Ellis 2002). Changes in the characterization of language proficiency have led to a shift of focus away from grammar. In the 1990s several language teaching approaches popped up emphasizing the importance of a lexical approach to language learning, as can be seen in Lewis' teacher training manuals that revolve around the adagio that language consists of "grammaticalised lexis", not "lexicalized grammar" (Lewis 1993, 1997) and in scholars' assertions that vocabulary needs to be systematically integrated into any course (Schmitt 2000, Nation 2001, Meara 2002). It is clear that the lexical nature of language and the implications for language pedagogy have been widely assessed.

The upsurge of the role of vocabulary in foreign language acquisition went hand in hand with a growing interest in vocabulary testing in SLA research. Meanwhile, researchers have been able to ascertain that the size of one's vocabulary seems to be a determining factor for second language learning (e.g. Meara 1996). Obtaining a sufficiently large vocabulary appears to correlate strongly with other linguistic competences in the target language. Therefore, much recent work on vocabulary testing has focused on estimating how many words learners know in their L2 (e.g. Laufer 1998). To accomplish this goal, vocabulary size tests have been developed. These are premised on the belief that learners need a certain amount of vocabulary in order to be able to operate independently in the target language (Alderson and Banerjee 2001). Two vocabulary size tests have seen wide recognition and application and they have in common that they present the testee with fairly straightforward tasks: the Vocabulary Levels Test (Nation 1990) requires test takers to match a word with its definition; the Yes/No Vocabulary Test (Meara and Buxton 1987) requires test takers simply to say which of the words in a list they know.

This book centers around the use of the latter test format, the Yes/No Vocabulary Test. The central question of this study is whether the Yes/No Vocabulary Test is a reliable and valid test for measuring foreign language learners' receptive vocabulary size. The data were collected from French-speaking learners of Dutch who had to be placed into the appropriate course programme. The study finds its origin in the fact that the use of the Yes/No Vocabulary Test at the language centre of the Université Libre de Bruxelles raised a lot of questions concerning the test's reliability and validity. It appeared that, despite its widespread and manifold use, the format still needed thorough analysis from a measurement perspective. The crux of this research project focusses on handling and suppressing the response bias that the Yes/No Vocabulary Test seems to provoke in the test takers, which seriously endangers the validity of the format. In the subsequent chapters of this book, the validity of the test format is re-assessed on the basis of theoretical considerations as well as experimental data. When we fail in confirming the format's validity, an alternative vocabulary test is presented that retains the universal properties of the Yes/No Vocabulary Test but outperforms it in circumventing the response bias problem. This new vocabulary size test is called the Recognition Based Vocabulary Test.

The results of this study are relevant for language testers and test developers as well as for anyone who is concerned with evaluating language skills. The construction of a valid instrument for measuring the receptive vocabulary size of learners in a foreign language is also of interest to vocabulary researchers because it will provide insights into how learners acquire a target vocabulary, at which speed their lexical progress takes place and how their vocabularies evolve. The opportunity of measuring learners' vocabulary size at different stages of their learning process also serves a pedagogical end. For example, when the objective of a course is that the learners should have mastered the core vocabulary of Dutch by the end of their course programme (as is the case in the objectives formulated for high school education in Brussels), the use of a thoroughly validated standardized vocabulary test allows teachers and administrators to verify if this goal has been achieved. Apart from serving a diagnostic purpose, a standardized vocabulary measure is at its most useful as a placement test, where it can serve as a rapid and therefore powerful tool to assign learners to classes of the appropriate level.

In this introductory chapter, we will first elucidate on the call for standardization in testing (Section 1.1). This is followed by a description of the role of the European Union in developing standardized language measures (Section 1.2). In Section 1.3, the Yes/No Vocabulary Test will be described and we will illustrate why it has gained popularity as a standardized vocabulary size test. Then, we will sketch the markedly applied nature of this research project and the consequences thereof (Section 1.4). Finally, in Section 1.5, we will describe the outline of this book.

1.1 A call for standardization in language testing

In their State-of-the-Art review "Language Testing and assessment (Part I)" Alderson and Banerjee (2001:218) explain that the word "standards" can be found to refer to various meanings in the literature. It can be used to denote procedures for ensuring quality (standards to be upheld) but it can also refer to a level of proficiency or, when we talk about "standardized test", it points to tests whose difficulty level is known and which have been adequately piloted and analysed.

In many educational settings, vocabulary measures – much like other language measures - are mostly used as one-off tests that are designed for use with particular groups of learners and for particular purposes that are often course-related. Since these measures cannot be compared with each other, it is difficult to integrate the data they produce. Such divergent practice contributes to the fragmentation of the SLA field which explains the call for standardisation in testing. Language testers argue for standardization in assessment in the belief that such methods of examining performance will contribute more to reliable measurement than assessment by individual teachers who may have access to a wide range of evidence about the performance of their learners, but whose standards or criteria may vary and whose focus of observation may be unsystematic (Skehan 1998).

The possibilities offered by computerized testing also serve as an incentive for developing methods of evaluation that yield reliable and comparable results regardless of the test takers' L1 or their cultural and educational background. Groot (1990) argues that the standardisation of procedures for test construction and validation is crucial to the exchangeability of test results across different education settings. Through systematic data collection, generalizations can be made to a wide range of contexts going well beyond the test itself. This trend towards international standards for language proficiency and assessment procedures is boosted by the process of European unification, which has created a situation in which internationally interpretable language tests are requested (De Jong 1992).

1.2 The influence of a growing European context

In order to overcome the linguistic challenges concerning educational exchange and employment mobility presented by the European multilingual context, there is an increasing demand for international recognition of certificates. Universities allow their students to follow courses abroad, companies send their employees to subsidiaries in another country and through the use of the internet, communication between people with different L1's has expanded enormously.

Presently, language qualifications, whether they are provided by schools or by private organisations, vary in their standards and the language levels descriptors they use. The Council of Europe is concerned with the international comparability of certificates because it has become an economic as well as an educational imperative (Bologna Declaration 1999). In order to establish a common scale of reference and comparison between language tests and certificates the Council founded the Common European Framework. This Framework is gaining influence in language education as schools are advised to relate their courses and certificates to the common scale of levels that the Council of Europe has developed. The council also set up two influential initiatives, the European Language Portfolio and the diagnostic testing system DIALANG, to arrive at valid records of competence regardless of country, region, sector, or institution of origin (Alderson and Banerjee 2001).

The European Language Portfolio was launched throughout Europe in 2001. It is a personal document that is intended to facilitate mobility within Europe by documenting language skills in a clear and internationally comparable way. The Portfolio consists of a Language Passport (which contains 6 levels of competence), a Language Biography (a personal record of the individual's language learning), and a Dossier of certificates and documentation. Within this portfolio, learners have to profile their language skills and this is where the development and use of language competence descriptors comes in. Reliable and standardized instruments are needed to map the linguistic knowledge of foreign language learners.

In order to enable language learners to identify their level of proficiency in a target language, a diagnostic testing system for languages was set up, called DIALANG. DIALANG is an on-line diagnostic language testing system (http://www.dialang.org). It is set up as a European project for the development of diagnostic language tests in 14 European languages. The tests for each language are anchored in the same scales of proficiency levels and these levels are based on the Council of Europe's scales, which are part of the Council's Common European Framework of reference. The system covers all levels, from beginner to advanced and it has been fully operational since spring 2003. It is predicted that the system will play a major role in language teaching institutions, as an instrument for placement purposes and for diagnosis of learning needs. The measure that is used to obtain information about the learners' vocabulary size in the target language in order to present the test taker with further language tests of the appropriate level, is the Yes/No Vocabulary Test (Meara and Buxton 1987).

1.3 The Yes/No Vocabulary Test as a standardized vocabulary measure

Read (2000) deplores the fact that, despite the growth of second language vocabulary studies, the design of tests that could function as standard

instruments for research or assessment purposes lags behind. The Yes/No Vocabulary Test is presented as the most authoritative vocabulary size test in this light, together with Nation's Vocabulary Levels Test.

The Yes/No Vocabulary Test is a checklist test that presents learners with a list of words in the target language and requires them to indicate if they know these words or not. It includes a number of pseudowords in order to be able to adjust the learners' test scores in case they have overrated their vocabulary knowledge. Selecting this particular test format as vocabulary size test in the DIALANG test battery is undoubtedly based on the same pragmatic arguments as those used by researchers who need estimates of vocabulary size of their non-native-speaking participants or teachers who want to evaluate the girth of their learners' vocabulary knowledge at the beginning or end of a language course: the test is simple to construct, it sets minimal demands on the testee and as a result of the format's simplicity a large number of words can be covered in a short time span, which allows obtaining a reliable estimate of vocabulary size. Add to this list of advantages the fact that the format can very easily be computerized and the test's popularity becomes self-evident.

Even those who object to discrete vocabulary measures - tests in which vocabulary knowledge is seen as a distinct construct and evaluated separately from other components of language - have to acknowledge that the Yes/No Vocabulary Test seems a very practical user-friendly tool. On top of that the test is reported to correlate well with global proficiency tests, which makes it a powerful indicator of language skill without having to subject participants to hours of test taking and administrators to hours of marking.

It were exactly those arguments that convinced us to use the format as vocabulary size test in the placement procedure for Dutch at the language centre of the Université Libre de Bruxelles. However, a first test use and analysis of the results revealed the presence of a response bias in the data. Further inquiry into the Yes/No literature and research into the response bias problem led us to suspect that most of the evidence cited in support of the reliability of the Yes/No test might be overestimated. This convinced us that the format should be re-assessed in terms of contemporary standards of test validation, which in turn led to a series of Yes/No experiments which will be reported in this study.

It has to be noted that an investigation of the Yes/No literature has shown that the format has rapidly acquired a reputation and wide application without necessarily being subjected to the required reliability and validity checks, which certainly was not the intention of its developers. In the article that reported the development of this new vocabulary measure, Meara and Buxton (1987) naturally stressed the positive features of the format and the possibilities it could offer when further researched. Since then, the format has been picked up by many other researchers who have sometimes neglected to acknowledge the limitations of the early published results. One could say that

fairly quickly after the Yes/No Vocabulary Test was introduced, it took off and began to lead a life of its own.

In this study we will attempt to shed more light on the different variables that come into play when using the Yes/No Vocabulary Test for measuring the receptive Dutch vocabulary size of French-speaking learners. The central aim is to find a suitable way of dealing with the response bias we encountered, which we would consider a valuable contribution to the improvement of the Yes/No format.

1.4 The "applied" nature of the research project

The research presented within the scope of this dissertation has a markedly "applied" nature. It arose from a need to select a receptive vocabulary test to include in the placement procedure of the language centre of the Université Libre de Bruxelles. The pragmatism that lies behind this research has its consequences for the way the experimental set-up is organized and for how the data were collected. One of the disadvantages of gathering data within a realistic setting is that the test administration is primordial. Also, the scoring of the test (which this study will demonstrate to be a complicated matter) gets absolute precedence. Test results that suffer a bias have to be transformed into reliable estimates of the students' vocabulary size, one way or the other. In instances where the test responses provide no basis for making a meaningful estimate of the testee's vocabulary size (because of ticking too many pseudowords), the learner's effort in taking the test becomes worthless and this is unacceptable.

However, this kind of pragmatism could be considered an asset rather than a setback, not only because one is continuously reminded of the complex ecological reality of a testing situation and all the elements it involves but also because the research aim and the utility of the test continue to collide. The incessant feedback of confirmed or refuted hypotheses pointing towards new, improved test formats keeps one firmly in touch with the central aim of any test development enterprise: constructing reliable and workable measures of linguistic knowledge and linguistic skills.

1.5 Outline of this study

This book consists of ten chapters. Chapter 2 delineates the gradual recognition among applied linguists of the central role of the lexicon in second language acquisition. It makes a strong case for measuring vocabulary knowledge, and more in particular, vocabulary size. Chapter 3 presents an extensive description of the Yes/No Vocabulary Test and the variables that come into play when constructing or taking the test. In view of the "pragmatic" origin of this research project, Chapter 4 sketches a brief outline of the daily functioning of the language centre where the data were collected. The design of the different

experiments is described in order to provide the reader with a perspective of how the study came to be. Chapter 5 reports on the experimental data we obtained when we used the Yes/No Test for the first time in the placement procedure for Dutch. The high false alarm rate displayed by the participants called for an in-depth analysis and discussion of how the correction formulae proposed in the literature deal with this phenomenon and how the test reliability is influenced as a consequence. These experimental findings and the theoretical discussion they inspired concerning the Yes/No correction formulae have formerly been published (Beeckmans, Eyckmans, Janssens, Dufranne and Van de Velde 2001). This chapter takes a central position in this study because its discussion of the response bias problem determines how the subsequent experiments are conceived and analyzed. In Chapter 6, empirical evidence is collected concerning the format's validity. An experiment is set up to examine the influence of using different correction formulae on the correlation between Yes/No Vocabulary Test results and the results on a translation task of the same words. When the concurrent validity is found to be disappointing, a series of experiments is set up, aiming to isolate the variables responsible for causing the high false alarm rate and trying to evaluate the influence of this problem on the test's validity. These experiments are described in the subsequent chapters. Chapter 7 consists of two parts which report two different experiments that are both aimed at reducing or eliminating the response bias. In the first part, the relation between different test instructions and their influence on the participants' response behaviour is investigated. In the second part, the impact of different computer software applications on the participants responses is examined. Chapter 8 reports a last ditch attempt to reduce the response bias by abandoning our self-made test content and turning to the content of the DIALANG diagnostic language testing system. The resulting Yes/No vocabulary test (infused with the DIALANG test content) was administered to French-speaking learners of Dutch and, in a subsequent experiment, to Dutch native speakers. In Chapter 9, a new vocabulary test is introduced, the Recognition Based Vocabulary Test, that retains the attractive features of the Yes/No Vocabulary Test but is designed to sidestep the response bias problem. Two variants of this new test format are compared to the Yes/No Vocabulary Test in an experimental design that includes validating the test data by means of a translation task. Finally, Chapter 10 summarizes the main conclusions of the aforementioned 7 experiments and reiterates that validation research should include how the nature of the test task interacts with various features, including the characteristics of the test takers and the testing context. It also puts forward some research options for the future development of vocabulary size tests.

Chapter 2

Vocabulary size

The importance of vocabulary in language acquisition goes uncontested. It is evident that vocabulary is indispensable for successful communication in any language. However, the key role vocabulary plays in language learning has not always been reflected in the amount of attention that has been given to it by language teachers and researchers in applied linguistics.

The evolution towards a recognition of the importance of lexical competence within second language learning[1] will be briefly sketched in the first Section of this chapter. In Section 2.2, the question is addressed which specific part of the target lexicon should be presented to language learners at what stage and it is followed by a short summary of how word knowledge has been defined in the SLA literature. From then on, the focus of attention shifts from vocabulary acquisition to vocabulary assessment. In Sections 2.4 and 2.5 the reasons for assessing vocabulary knowledge and, more particularly, vocabulary size, are articulated. This is followed by an attempt to define the construct receptive vocabulary (Section 2.6). Finally, in Section 2.7, the characteristics of standardized vocabulary measures are discussed.

2.1 Changing attitudes towards vocabulary acquisition

In the - not so distant – past, mastery of grammatical structures was seen as central to learning a foreign language. The main focus in classroom activities and FLA research was on the acquisition of grammatical competence and the development of functional communication skills. Vocabulary development was seen as some kind of secondary or auxiliary activity and it usually involved memorizing word lists. By no means could we speak of a principled approach towards the acquisition of the target lexicon (Nation 2001).

Gradually second language acquisition researchers have come to recognize the central, or even preconditional, role of the lexical dimension for fluent language use, whatever skill concerned. Many applied linguists have demonstrated, for instance, that the nature of the language threshold for reading is largely lexical. Anderson and Freebody (1981) reported the high correlation between tests of vocabulary and reading comprehension as a

[1] In this study, we will not make the distinction between second and foreign language acquisition and we will not go into that particular terminological discussion. We will consistently use the term "second language" instead of "foreign language" to refer to the target language.

consistent finding in L1 reading research. Vocabulary difficulty was demonstrated to be a factor of overpowering importance in studies of L1 readability (Wittrock, Marks and Doctorow 1975). Laufer (1989, 1992) showed that the same applies to second language acquisition. She emphasizes the importance of having a vocabulary large enough to provide coverage of 95% of the words in a text. Reading is an important part of most language programmes, no matter whether they are aimed at beginners or intermediate and advanced learners. Learners whose target vocabulary is not large enough to have 95% coverage do not reach an adequate level of comprehension of the texts and are unable to transfer their reading skills from their L1 to their L2. Ellis (1997) has shown that vocabulary knowledge is indispensable to acquire grammar. Knowing the words in a text allows learners to understand the discourse, which in turn allows the grammatical patterning to become more transparent. Nation (1990, 1993, 2001) underlines the critical importance of developing an adequate high-frequency vocabulary since learners' skill in using the language is heavily dependent on the number of words they know, particularly in the early stages of learning a foreign language, with around 3,000 word families being a crucial threshold. He states that a systematic, principled approach to vocabulary development results in better language learning (Nation 1990).

Since the mid-eighties, the study of vocabulary in applied linguistics has been flourishing. Developing lexical competence in the target language is now seen as the crucial factor in language acquisition and there is general agreement that there is a threshold vocabulary below which learners are likely to struggle to decode the input they receive (Alderson and Banerjee 2002). A glance at recent language learning methodologies reveals the priority that is nowadays given to lexical approaches in language learning (Willis 1990, Lewis 1993, 1997, 2000) in which the relationship between vocabulary knowledge and other aspects of linguistic ability is implicit.

2.2 Which words should second language learners know?

Since it has been established that the teaching of vocabulary is crucial and needs to be structured, it has been widely accepted that this structuring needs to be done on the basis of word frequency and text coverage (Meara 1993). It seems evident that the more frequent words are most useful and should be taught first, before spending time on less frequent words or words that only occur in specialised domains. Nation (1990, 2001) reports that frequency-based studies have shown that a small group of very frequent words cover a very large proportion of the running words in any spoken or written text and occur in all kinds of uses of language. In other words: a relatively small amount of well-chosen vocabulary according to frequency and range can enable learners to do a lot. Actually, Nation (1990) divides vocabulary into three groups: (1) a small number of high-frequency words, which are clearly so important that

considerable time should be spent on them by teachers and learners; (2) a very large number of low-frequency words, which require the mastery of coping strategies; and (3) specialized vocabulary which is of interest for learners who are active in specific professional fields. Since the high-frequency words play so prominent a role in vocabulary learning the question arises if this group of words within a language is stable. According to Nation (2001) frequency lists may differ in frequency rank order of particular words but there generally is 80% agreement about what words should be included in the list, provided that the corpus has been well-designed (Nation 2001: 15-16).

With reference to word counts, Nation (2001) holds that knowing a word involves knowing the members of its word family and the number of members of the word family will increase as proficiency develops. A learner may be familiar with the word "rich", "richly" and "richness" in an early stage and expand this word family with "to enrich" and "enrichment" in due time. There is research evidence supporting the idea that word families are psychologically real, and that rather than talking about "knowing a word", we should be talking about "knowing a word family" (Nation 2001:47).

A frequency-based approach to vocabulary learning hinges upon the assumption that frequency is strongly related to the probability that a word will be known. Anderson and Freebody (1981) report that this hypothesis is supported by evidence from a number of L1 areas. Hazenberg and Hulstijn (Hazenberg 1994, Hazenberg and Hulstijn 1996) have researched to which extent word frequency can be used to predict word knowledge. One might expect that the most frequent words are known by all students, whereas more infrequent words are known only by particular individuals, depending on variables such as hobbies, work and experiences. They concluded that the relationship between word frequency and word knowledge appears to depend on vocabulary size. When individuals have a relatively large vocabulary there is no significant relationship. But when individuals have a relatively small vocabulary, word frequency can be used as a criterion to predict word knowledge. It has thus been established that the further you move on from the high-frequency vocabulary, the less significant frequency becomes in an absolute sense. The selection of lower-frequency words depends increasingly on the learners' specific needs and interests. This stresses once more the importance of the 3,000 word family (which corresponds more or less with the 5,000 most frequent words) as a learning objective for any language learner. Beyond the 5,000-word level, Meara (1996) argues that vocabulary size is less important than the way in which the vocabulary is organised in the learner's mind. The hypothesis is that those with a more developed vocabulary knowledge have a more complex and highly structured network of associations among the words they know.

Schmitt (2000) advocates that vocabulary should best be taught to foreign language learners according to a cost-benefit perspective. He mentions

the most frequent 2,000 words as the most commonly cited initial goal for beginners and agrees that these have to be taught explicitly. Meara (1995) claims these are so essential for any real language use that it might be a good idea to teach them right at the beginning of the language course. When learners move on to read authentic texts in the target language, the consensus among applied linguists seems to be that 3,000 to 5,000 word families should suffice. However, Hazenberg and Hulstijn (Hazenberg 1994, Hazenberg and Hulstijn 1996) calculated that foreign students reading university texts need to have 10,000 to 11,000 word families at their disposal. For communication in specific professional domains, it is recommended to have a solid base of high-frequency vocabulary, complemented with the specialized vocabulary required for the domain in question.

Most vocabulary researchers agree that although explicit vocabulary instruction should not cease after the 2,000 most frequent words, it is very important to make the learners responsible for their individual vocabulary learning. Several vocabulary learning strategies should be acquired so that learners can learn words autonomously. Learning word-building processes in the target language, guessing from context and applying mnemonic techniques are strategies that have proven to be very useful (Nation 1990). Through reading, combined with the development of a raised awareness of vocabulary learning strategies, learners can expand their vocabularies far beyond the level of 11,000 word families, even within the realm of a native speaker's vocabulary size that is thought to consist of 15,000 to 20,000 word families (Nation and Waring 1997).

2.3 What constitutes word knowledge?

A great deal has been written on the topic of what it means to "know" a word. Anderson and Freebody parodied this fact by writing that it "(…) is not clear that, if Ludwig Wittgenstein and Bertrand Russell were left alone in a room for three hours, they could decide that they really knew the meaning of dog" (Anderson and Freebody 1981: 90).

Aside from the philosophical speculations that can be raised concerning this issue, the many taxonomies of word knowledge find their origin in the fact that lexical knowledge is not an all-or-nothing phenomenon, it involves degrees of knowledge. Therefore, people's vocabulary knowledge is called incremental: knowledge of a word is to be seen as a continuum from "not knowing" to rich knowledge of a word's meaning, its relationship to other words, and its extension to metaphorical uses (Beck and McKeown 1991:792). Vocabulary knowledge in the mother tongue as well as in a foreign language continues to deepen throughout lifetime: as you grow older, you continue to learn nuances and subtle distinctions conveyed by words. Anderson and Freebody reported that most of the research done on semantics supports the

conclusion that there is progressive differentiation of word meanings with increasing age and experience (1981: 93).

Much of what is written on word knowledge goes back to the well-known vocabulary knowledge framework of Richards (1976). He identifies seven aspects of word knowledge. In his view, "knowing a word" means:

a) knowing the degree of probability of encountering the word in speech or print,
b) knowing the limitations imposed on the use of the word according to function and situation,
c) knowing the syntactic behaviour associated with the word,
d) knowing the underlying form of a word and the derivations that can be made of it,
e) knowing the associations between the word and other words in the language,
f) knowing the semantic value of the word, and
g) knowing many of the different meanings associated with the word.

Applied linguists seem to agree that the same continuous idea of incremental expansion of vocabulary knowledge also applies to the transfer from receptive to productive mastery. The learning of a word is thought to progress from receptive to productive knowledge. This means that a word that can be correctly used, is assumed to be understood by the user, when heard or seen. The opposite however, is not necessarily true. Passive vocabulary size is thus considered to be larger than the active size even though it is not clear how much larger it is. In Nation's (1990) framework for vocabulary knowledge, he therefore distinguishes eight types of word knowledge that are specified both for receptive and productive knowledge.

In the research reported in this study we will only deal with a very basic form of word knowledge. We will settle for the ordinary, everyday sense of knowing a word: recognizing a word in the target language and being able to recount one of its possible meanings in the learners' mother tongue.

2.4 Why measure vocabulary knowledge?

If vocabulary is considered a priority area in language teaching, then it needs to be assessed in some way and test formats are needed to monitor learners' progress in vocabulary learning. There are several arguments to be made in favour of vocabulary testing. First of all there is the affective dimension: tests have consequences far beyond providing estimates of the learners' abilities, they shape the way the learners perceive the content of a course. This is the so-called backwash effect of testing. If students are not tested on vocabulary, they might conclude that vocabulary does not really matter. If teachers want to create a positive attitude towards vocabulary learning, it does not suffice to put

emphasis on vocabulary in the course programme, it needs to be included in tests and exams as well.

A second argument is research-based: vocabulary test results provide useful information on how vocabularies develop. It is important to know how many words foreign language learners know, how fast their target vocabularies grow, and how these factors are related to other aspects of their linguistic competence. Rather than simply measuring vocabulary knowledge, objective vocabulary tests seem to be valid indicators of language ability in a broad sense. If vocabulary levels do reflect language development more generally, then vocabulary testing might offer a relatively quick and easy way for researchers and schools to monitor progress in language development (Cameron 2002). For instance, an assessment of the number of word meanings a reader knows appears to predict this individual's ability to comprehend discourse remarkably accurately. The deeper reasons why word knowledge correlates with comprehension cannot be determined satisfactorily without improved methods of estimating the size of people's vocabularies (Anderson and Freebody 1981). We will return to the potential of vocabulary size as a useful parameter in describing second language ability in Section 2.5.

In the same manner as other language tests, vocabulary tests can serve different purposes: they can be used to assess whether learners have acquired the words they were taught (i.e. achievement testing), they can help detect whether there are gaps in the vocabulary knowledge of learners (i.e. diagnostic testing), they can aim to place students in the appropriate language class level (i.e. placement testing), or they can form part of a more global language proficiency test in order to arrive at an estimate of the learner's skills to perform in the target language (i.e. proficiency testing).

Up until now, the vocabulary measure that is under scrutiny in this study, the Yes/No Vocabulary Test, has been mainly used as a placement test (i.e. the Eurocentres Vocabulary Size Tests, Meara and Jones 1990) or as part of a diagnostic tool (i.e. the DIALANG test battery).

2.5 Why measure vocabulary size?

Vocabulary learning is not only a quantitative issue. Researchers distinguish "breadth" or "size" of knowledge (the number of words of which the learner knows at least some significant aspects of the meaning) from "depth" of knowledge, with which they refer to the quality of vocabulary knowledge, namely how well a particular word is known. Although both measures are considered important - knowledge of words progresses from superficial to deep at various stages of learning - a lot of work on vocabulary testing has focused on vocabulary size.

Even though Meara is convinced that this two dimensional approach is too limited a view (because it does not suffice to explain the diversity that is

found in language learners) and would prefer more research to be done into the accessibility of words in the L2 lexicon (Meara 2002), he endorses that the basic dimension of lexical competence is size (Meara 1996). He states that :

> "All other things being equal, learners with big vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies, and there is some evidence to support the view that vocabulary skills make a significant contribution to almost all aspects of L2 proficiency" (Meara 1996: 37).

Concerning L1 vocabulary knowledge, Anderson and Freebody agree that "Measures of vocabulary knowledge are potent predictors of a variety of indices of linguistic ability" (Anderson and Freebody 1981: 77). In the past, researchers even went as far as saying that the size of a person's vocabulary is a very good predictor of that person's general intelligence (Terman 1918). Another reason for measuring vocabulary size that we have already mentioned in Section 2.1, is that vocabulary size was found to be a good predictor of reading comprehension (Anderson and Freebody 1981). It has also been shown to be an important factor for obtaining fluency in speech (Coady, Magott, Hubbard, Graney and Mokhtari 1993).

From a pedagogical perspective it is useful to know how much vocabulary instruction is needed before learners have reached the vocabulary threshold level which is necessary for the comprehension of written texts. As we have already mentioned, it is assumed that in order to reach text comprehension, readers need to be familiar with 95% of the words in a text (Hirsch and Nation 1992) and it has been claimed for various languages that the 5000 most frequent words yield a coverage of 90% to 95% of the word tokens in an average text (Sciarone 1979, Laufer 1992, Nieuwborg 1992, Nation 1993), although Hazenberg's research about the vocabulary size required for reading at university level pointed at a much higher threshold of minimally 10,000 base words (Hazenberg 1994). From the viewpoint of the language learner himself, Laufer (1998) remarks that they associate progress in language learning often with an increase in the number of words they know.

Vocabulary researchers believe that measures of vocabulary size could shed light on the relationship between vocabulary growth and different input conditions so that it becomes clear at what stage to prefer comprehension-based rather than production-oriented instruction. Such information could also help to fathom the similarities and differences between the development of passive and active vocabularies.

Notwithstanding the arguments that can be made about the limited nature of vocabulary size testing – learners' proficiency in a foreign language is not solely determined by their vocabulary size, they need to be able to draw on that knowledge in a communicative situation, which reiterates Meara's

preoccupation with the accessibility of the L2 lexicon - it is in any case an important aspect of the lexical development of all language learners.

2.6 Construct definition

The first question to ask when testing vocabulary, according to Nation (1990), is whether you wish to test recognition or recall of vocabulary. In recall tests we are interested in the learners' production of a word in the target language. In recognition tests we want to see if the learners know the meaning of the word after they see or hear it. There are several ways in which the test can elicit learners' recognition of word knowledge. They can be asked to translate the word into their L1, or to provide a synonym or definition of the word in the target language, or to tick the word when they think they know it, or to choose from a set of pictures, L1 words, or synonyms and definitions in the target language (Nation 1990).

The distinction between recognition and recall is what is often referred to as receptive versus productive knowledge. As we have already mentioned in Section 2.3, it is generally assumed that words are known receptively first and only later become available for productive use, which is why it is most useful to think in terms of a receptive to productive continuum, representing increasing degrees of knowledge of a word. This continuous aspect that is inherent to many language abilities illustrates the importance of defining the construct when designing a language test. The term construct refers to the particular kind of knowledge or ability that a test is designed to measure. In the case of vocabulary size tests, the process of clarifying what is meant by receptive vocabulary is an exercise in theory-based construct definition. We need to define what specific learner ability "receptive vocabulary knowledge" refers to. For Nation (1990) knowing a word receptively involves being able to recognize it, being able to distinguish it from words with a similar form, being able to judge if the word form sounds right or looks right, having an expectation of what grammatical pattern the word will occur in, having some expectation of the words it collocates with, and being able to recall its meaning when it is met. However, different test formats could address different construct definitions of receptive vocabulary size, as will become clear throughout this study.

2.7 Characteristics of standardized vocabulary measures

The multiple choice format has long been - and probably still is - the most widely used procedure in standardized vocabulary testing. Anderson and Freebody (1981) pointed out that the distracters in a multiple choice format cannot avoid constraining the participant's response. If the purpose of the test is to provide data on relative performance only, not on absolute level of performance, then the distracters are chosen to maximize the discriminating power of the item. Anderson and Freebody concluded that if one is interested

in vocabulary size, this policy will not do and one should turn to other tasks and formats.

In order to assess vocabulary size in a valid and reliable way, vocabulary size tests must consist of many items. This, in turn, calls for a non-time-consuming administration procedure, which entails that the test task has to be fairly simple. This is why the instruments that have been proposed to date are discrete and context independent in nature (Read 2000). One of the most well-known of these discrete vocabulary measures is Nation's (1983, 1990) Vocabulary Levels Test, which Meara (1996: 38) considers as "the nearest thing we have to a standard test in vocabulary". This test samples words from the 2,000, 3,000, 5,000 and 10,000-word frequency levels, and from an academic register known as the University Word List. It samples recognition knowledge of 18 words sampled from each of the five frequency levels. The test task requires test takers to match a word with its definition, presented in multiple choice format in the form of a synonym or a short phrase. With only 18 items at each of the five levels, the test is compact and usable in classroom conditions.

A second well-known standard vocabulary size test is Meara's Yes/No Vocabulary Test (Meara and Buxton 1987) which is under discussion in this study. It makes an estimate of learners' vocabulary size using a sample of words covering several frequency levels. It is a checklist test consisting of words and non-words and the learners have to tick the words they know the meaning of. It has been turned into a computer application, the Eurocentres Vocabulary Size Test (Meara and Jones 1990) and the format has also been selected as vocabulary test within the European DIALANG system. It will be described in detail in Chapter 4.

Several objections can be made to this kind of discrete vocabulary tests. For one, the test items cannot give a precise indication of what a learner knows about a word. They only capture a partial amount of learner's potential knowledge which is often limited to the meaning of the word in question, or even less. "Meaning" and "word form" are measured on a harsh "knows/does not know" scale. A second objection concerns the discrepancies in vocabulary size estimates that could arise as a result of the sampling procedures. The procedure that consists of taking a sample of words from a dictionary immediately raises the question what is to be counted as a word, and if morphological derivations of a base word should be counted as separate items or not. A "liberal" policy, selecting also derivative and compound forms, will lead to large estimates of vocabulary size. One also has to decide whether proper names, acronyms, technical terms, archaic words, slang and compounds will count as separate words. Researchers have adopted different approaches to these questions, with predictably different results.

When the aim is to test the learners' success in acquiring the vocabulary of a particular course, it seems evident that the words to be tested

should be selected from the course materials. However, vocabulary size tests for diagnostic or placement purposes should be sampled from a more general range of words. Especially in circumstances where the learners have different language backgrounds (different schools and different language teaching methodologies or different L1s), it is commendable to select the words from word-frequency lists. Meara (1996) notes that estimating the total number of words that make up the vocabulary is probably the critical problem in constructing a test of vocabulary size. If the result of a test suggests that the learner knows 25% of the target vocabulary, then the estimated size of the target vocabulary is important, for 25% of 4,000 words is much less than 25% of 20,000 words.

Finally, and this may be the most fundamental objection to these tests, both vocabulary size measures are of course decontextualized, which means that vocabulary knowledge is taken as a distinct construct, separated from other components of language. This discrete-knowledge approach does not coincide with the widely held view that being skilled in a foreign language is not just a matter of possessing a particular knowledge component of language ability (vocabulary, grammar, etc) but being able to apply that knowledge for communicative purposes (Read 2000).

Although vocabulary should preferably be assessed in contextualized language use (where it interacts with other components of language knowledge), it is useful to develop discrete tests that measure whether learners know the meaning of a set of words for placement aims or even diagnostic aims. Moreover, with reference to the context-independent nature of vocabulary size tests, Cameron (2002) argues that decontextualized presentation of a word in a test does not imply that the learner makes sense of the test word in a decontextualized mental void. The recognition process may activate recall of previous encounters and their contexts and it is therefore useful to see how much vocabulary can be recognized without extended linguistic or textual context. Even Read (2000), one of the strongest defenders of the view that vocabulary should always be assessed in context, admits that research on the cloze test has shown that the more the assessment of vocabulary is contextualized, the less clear it may be to what extent it is vocabulary knowledge that is influencing the test-takers' performance.

The position that has been taken in this study – and from which this research project has originated – is that discrete vocabulary measures remain a useful tool for the language teacher and researcher but, as Read (2000) emphasizes, new tests should be thoroughly underpinned and analyzed according to contemporary standards of test design and validation. It is within this context that this book is meant to make a contribution for the case of the Yes/No Vocabulary Test.

Chapter 3

The Yes/No Vocabulary Test

In recent years, the Yes/No Vocabulary Test has been used for research purposes (Abels 1994; Vives Boix 1995; Huibregtse & Admiraal1999; Van de Walle 1999; Shillaw 1996; Hermans 2000; Hommersom 2003) and as a placement test because of its reported merits. It is easy to construct, administer and score, which means it exemplifies an approach that makes efficient use of examiner and examinee time.

After a description of the format in Section 3.1, the history of the test format's development will be sketched (Section 3.2) and the scoring method will be introduced (Section 3.3). Then, a brief summary of the validation evidence is presented (Section 3.4) and followed by a report of the problems encountered in the Yes/No literature (Section 3.5). Finally, several considerations are put forward that relate to the construct validity of the format (Section 3.6).

3.1 Format

The Yes/No vocabulary test is a test format that intends to measure learners' receptive vocabulary size by presenting them with a sample of words in the target language covering certain frequency levels and asking them to indicate the words they know the meaning of (for an example of the test, see Appendix 1). This means that the test aims to measure receptive vocabulary size through word recognition. If a student recognizes a word and ticks it, he or she is supposed to "know" it. Clearly, there is much more to knowing a word than just recognizing it. But, the test is not out to measure deep lexical knowledge (this would include spelling, word associations, grammatical information and multiple meanings of the target words). And, as Cameron (2002) points out, although such a word recognition measure only taps into a small part of the complexity of the vocabulary knowledge of any given language learner, a word recognition count can be a useful indication of the outer limits of the learner's vocabulary knowledge, for presumably the words which a learner understands or uses with any depth of meaning will also be recognized in the Yes/No format.

Like the Vocabulary Levels Test (Nation 1983, 1990), the Yes/No Vocabulary Test is based on the hypothesis that there is a direct relationship between the frequency of a word in a language and the probability that a learner will know it. Clearly, the test will not give an accurate estimate of vocabulary size if the learner's knowledge of words is different from the frequency profile

that is assumed. According to Meara and Jones' (1990) findings, most learners fit this pattern quite closely.

The test task appears to set minimal demands on the testee as far as strategic knowledge is concerned. In fact, Read (2000) highlights the great attraction of the simplicity of the task, as a result of which a large number of words can be covered within the testing time available so that the required sample size necessary for making reliable estimates can easily be achieved. Anderson and Freebody (1983) used the method to estimate the vocabulary size of children's L1 and commented on the advantage of not needing trained item writers or a secure item pool since the test items are not embedded in a complex context of distracters. On top of that, they found that recognition of over twice as many words can be tested in the same time span as in a multiple choice test. Most importantly, they concluded that "[…] a score on a yes/no test provides a much more valid indicator of whether an examinee actually knows the meaning of the tested words than a score on a standardized multiple choice test" (Anderson and Freebody 1983: 269).

With regard to the use of the Yes/No Vocabulary Test as a measure of vocabulary size in a second or foreign language, Meara (1996) states that the test works well across a wide range of proficiency levels. Unlike many standard test formats it seems to be equally suitable for use with beginners as with advanced learners. Moreover, the profiles rendered by the test are sufficiently sensitive to measure vocabulary growth over relatively short periods of time as the repeated use of the test allows tracking the rate at which learners acquire new words (Meara 1993).

3.2 History

The Yes/No Vocabulary Test is derived from a simple format known as the "checklist", which presents the learners with a set of words and instructs them to mark the words of which they know the meaning. This format was originally used in L1 research (Sims 1929; Tilley 1936; Zimmerman, Broder, Shaughnessy and Underwood 1977). Unfortunately, learners' self-report of whether or not they know a word appeared to be a poor guide to their actual knowledge of vocabulary (Nation 1990; Read 1997a). The big question about the Yes/No method has always been obvious: what is to prevent people from overstating their vocabulary knowledge, checking "Yes" for words they do not actually know? Therefore, Anderson and Freebody (1983) decided to add pseudowords to the list in order to take into consideration the possibility that certain learners might be using too lenient a standard in judging whether they "knew" a word. Claiming knowledge of the pseudowords leads to adjusting the score downwards to provide a better estimate of the knowledge of the real words.

Meara and Buxton (1987) applied this adjusted Yes/No format to L2 learners in a first attempt to establish if this test design was workable. They

developed a Yes/No test with 60 real words and 40 pseudowords. Students were asked to indicate if they knew the meaning of the words. Meara and Jones (1988, 1990) developed a computerised checklist, the Eurocentres Vocabulary Size Test (Meara and Jones 1990). For this test, the frequency statistics of Thorndike and Lorge's (1944) list were used in order to sample words from ten frequency bands. It starts with the first thousand words and continues up to the tenth thousand words. The computer programme presents the test taker with a random sample of 20 words of each 1,000-word frequency band. An estimate of the individual's vocabulary size is made up to a ceiling level of 10,000 words. The same basic methodology was used in a book of paper-and-pencil Yes/No tests called the EFL Vocabulary Test (Meara, 1992), in which some changes to the scoring mechanism were introduced, an issue to which we will return in Section 3.3.

Recently, the Yes/No Vocabulary Test has been incorporated into the European DIALANG project, a computerised test battery for assessing language proficiency in 14 European languages (http://www.dialang.org). Within this learner-oriented framework, any testee around the world can arrive at a profile of his/her receptive vocabulary size in a given target language.

3.3 Scoring method

Establishing a representative score for the Yes/No test is not as easy as may appear at first. The introduction of pseudowords in the test format has important implications for the calculation of the test score. As there are two different kinds of items the learner is exposed to and two possible responses, four resulting combinations are possible for each item (see Figure 3.1):

- Hit: a "Yes" response to real word
- False alarm: a "Yes" response to a pseudoword
- Miss: a "No" response to a real word
- Correct rejection: a "No" response to a pseudoword.

This terminology finds its origin in Signal Detection Theory (SDT) which provides a theoretical framework to allow for a description of participants' decision behaviour in a detection task (Green and Swets, 1966).

The raw data matrix (see Figure 3.1) has to be transformed into a test score. This transformation of the learner's response behaviour into a test score is an intricate procedure. The most straightforward way of generating a global test result would be to consider the rate of correct responses (the diagonal hits-correct rejections represents the correctly answered items and the diagonal false alarms-misses stands for the incorrectly answered items, see Fig. 3.1). However, among the numerous scoring methods that have been proposed in the past, this has never been considered (Beeckmans et al 2001).

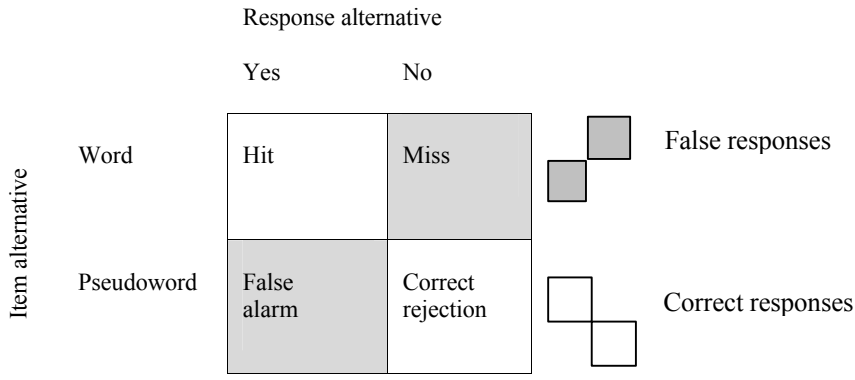


Figure 3.1: The item-response matrix of the Yes/No Vocabulary Test

All formulae encountered in the literature adhere to the same principle: among participants with the same hit rate, those with higher false alarm rates will end up with a lower test score. This illustrates the importance of the false alarm rate when calculating a test score: the pseudowords were introduced in the test design in order to prevent an overestimation of the participant's knowledge and ticking them leads to a negative adjustment of the test score.

Shillaw (1996) argues that the use of pseudowords detracts from the measurement quality of a Yes/No test. He analysed the scores he obtained with Japanese university students learning English by using the Rasch Model, a widely used method of test-item analysis. His results show that a checklist containing a suitable set of real words produces a highly reliable measure of vocabulary knowledge without any need for pseudowords. The words form a series of items that fit very well along a single measurement scale. In addition, Rasch analysis provides a way of identifying learners who may be overestimating their vocabulary knowledge because their responses tend not to fit the overall pattern of item difficulty. However, by excluding the pseudowords from the format, Shillaw has created a different test, and he has returned to the idea of preliminary item analysis, whereas Meara (1990) mentions as an important advantage of Yes/No tests that they do not require the complex standardization of other test formats.

It is clear that in the Yes/No format as it is known and used today, the pseudowords play an important role since they intervene so clearly in the scoring. Being able to distinguish between words and pseudowords is at the heart of the task of the Yes/No Vocabulary Test and of the measured construct. However, as we will illustrate in Chapter 5, when the raw test data exhibit considerable false alarm rates, the scoring of the test becomes unmanageable from a psychometric point of view. Test takers may end up with negative scores and the use of one or the other correction formulae may result

in different rank orders of the test taker population. In our data, this problem manifested itself as soon as a boundary of 15% false alarms was surpassed (Eyckmans, Beeckmans and Van de Velde 2001, Eyckmans, Beeckmans and Van de Velde 2002).

3.4 Validation

Attempts to validate the Yes/No Vocabulary Test have mostly involved using correlational procedures. Generally this has meant: correlating one vocabulary measure with another or considering one test a criterion measure and judging others to be valid according to how highly they correlate with the criterion.

In the early checklist literature – when the format was still devoid of pseudowords and when it was exclusively used as a measure of vocabulary size in L1- there is mention of construct validity problems. Sims (1929) compared four types of tests to measure vocabulary size in L1: (1) a multiple choice test, (2) a constructed answer-test in which the participant attempts to give a definition, a synonym, an illustration or uses the word in a sentence, (3) a Yes/No test and (4) a matching exercise where the participant pairs off words with their synonyms. He concluded that, although the checking method was as reliable as the others, it did not seem to offer acceptable construct validity. This led him to conclude that "the relative simplicity of such a measure, the ease of preparation and administration should not blind one to its invalidity" (Sims 1929: 96). Chall and Dale (1950, see Anderson and Freebody 1981: 108) reported that the average tendency to overestimate word knowledge amounted to about 11% in their L1 research and that the Yes/No test produced inflated estimates of vocabulary size and correlated poorly with other measures. Maybe this ought to be no real surprise in view of the fact that in these test uses the Yes/No format remained uncorrected for overestimation of word knowledge.

Anderson and Freebody (1981) compared the validity of a Yes/No Test (with pseudowords included in the list of items) with the popular format of a multiple choice test. The correlation between the multiple choice scores and the corrected Yes/No scores was .84. To determine which of the two measures gave the most valid assessment of vocabulary knowledge, they interviewed their participants. The children concerned had to read the words and define them or use them in a sentence. The interview scores appeared to correlate much better with the results of the Yes/No than the multiple choice scores which led them to conclude that the Yes/No Test gives a better estimate of true word knowledge than the performance on the standardized multiple choice test.

In L2 research Meara (1996) found that the Yes/No test correlated moderately well with other vocabulary tests and with tests of other linguistic skills, particularly integrative tests like the cloze, listening comprehension and

reading comprehension, where you would expect vocabulary knowledge to make an important contribution.

3.5 Reported problems in the literature

There are two important methodological objections to the use of the Yes/No Vocabulary Test that need to be addressed. The first of these has already been mentioned as a more general remonstrance of discrete measures in Section 2.2.4 and concerns the fact that the words in this test format are presented to learners in isolation, without supporting linguistic context. Current views emphasize that language tests should replicate situations of language use or learning (Bachman and Palmer, 1996) and that the presentation of isolated words may reinforce a simplistic view of what " knowing a word " entails. Contextualized words provide a much richer environment and may enhance the learner's awareness of the usage of these words (Read, 1997a). Cameron (2002) counters that after sufficient contextualized encounters, a word will be recognized regardless of whether it is met in a new context or in isolation. She considers recognizing words out of context an important type of knowledge intimately linked with reading, for vocabulary size test results have long been found to correlate with reading comprehension test results (Read, 1997b). Stanovich (1980) demonstrated that skilled readers use word recognition skills to understand text, only turning to contextual information when word recognition fails.

A second objection concerns the few demands the Yes/No task makes on the testee. The only skill that is measured is the testee's ability to recognize whether an item is a word or not. It does not measure whether testees can actually use the words they claim to know. However, Meara (1996) makes a case for the hypothesis that measuring vocabulary size with a Yes/No test does more than give you a rough measure of how many words a testee can recognize. He argues that it would be very unusual to find somebody with an L1 vocabulary of 10,000 words who did not know that "child" is a common word, used in slightly formal situations, that it is a noun, makes it plural with "-ren", and is associated with "boy, "girl", "parent", and so on. He claims that the circumstances which lead people to develop moderately large vocabularies in their L1 also allow them to acquire other types of information about the words. He assumes there is a similar link for L2 learners who acquire words from exposure to the target language and he states that during this acquisition proces they will inevitably learn more than just recognition of the form. In his words, "a learner with a huge vocabulary and nothing else is a possibility, but something of a freak" (Meara 1996: 44). Therefore, he concludes that measures of vocabulary size are more powerful than they may appear.

Apart from these methodological issues, there are the practical problems and empirical phenomena researchers and language teachers have

stumbled upon and reported. In L1 research, Anderson and Freebody (1983) were confronted with a phenomenon they called "mock" hits. With this term they refer to "yes" answers to unknown words as the result of having transformed them into known ones. They noticed that a word like "sham" was interpreted as "shame". They found the test scores to be inflated as a result of this phenomenon.

One of the particular problems of the format in L2 use, concerns the presence of cognates in the test material and the testees' response behaviour to them. This matter arises when the test is administered in situations where there is a strong lexical resemblance between the target language and the learner's mother tongue. Meara and Buxton (1987) report that particular pseudowords seem to be more attractive to speakers of some languages than others. The form "observement" for example resembles a real word in French or Italian but not in German. Thus, it should be easier for a German speaker to reject it than for a speaker of a Romance language. A a consequence of this response behaviour, the participants' scores got severely reduced and therefore presented an underestimation of the learner vocabulary size, but it was never concluded that the Yes/No test is an unreliable instrument in these cases or that word selection or pseudoword formation should be altered in the case of French participants. Meara and Jones (1990) were confronted with this cognate effect in administering an English Yes/No vocabulary test to speakers of French. For these testees, the Yes/No tests seemed to correlate much less well with other linguistic skills than was the case for testees with other L1s. Meara attributes this to the exceptionally close relationship between the lexicons of English and French. However, this finding is contradicted by another research experiment with francophone learners in Montreal. Meara, Lightbown and Halter (1994) investigated the hypothesis that Yes/No Vocabulary Tests which include a substantial number of cognates in the learners' L1 could lead to an overestimation of those learners' proficiency and, conversely, that the exclusion of cognates could lead to an underestimation of the learners' real vocabulary size. They concluded that tests in which the number of cognates is close to the proportion actually occurring in the language, do not compromise the validity and continue to correlate highly with other measures of language skills.

Cobb (2000) reports that the test is known to function poorly with Arabic-speaking learners, who indicate a very large proportion of non-words as known (Al-Hazemi 1993; Ryan 1997). An explanation for this phenomenon is that vowels are not normally written in Arabic script but rather supplied by the reader following a contextual interpretation (Abu Rabia and Seigel 1995). With cognitive process transfer, Arabic speakers reading English are often blind to vowel-based distinctions between words, especially words out of context. Thus, they are likely to judge "tilt" and" toilet", or "mascarate" and "miscreate" as the same word (Ryan and Meara 1991).

Finally, the most worrying pitfall to note is that the Yes/No format was found to perform less well with low-level learners, who respond unpredictably to the pseudowords. Certain learners obtain very low scores as a result of their overwillingness to claim knowledge of the pseudowords (Meara 1996).

3.6 Considerations relating to the construct validity of the format

It was during our first test construction and experience of the Yes/No Test that we ourselves were confronted with some of the format's shortcomings. What seemed an uncomplicated format at first turned out to be an intricate measure whose effectiveness may be doubted. Below we list some of the key issues that need to be reconsidered and resolved if the test is going to be used as a standardized vocabulary measure, all of which will be addressed in the course of this study.

3.6.1 The format

The Yes/No format itself is not clearly defined. The name suggests a format with an explicit distinction in choosing "Yes" or "No" (see Appendix 2) as was the case in Meara (1992). However, some studies, e.g. the first paper on the use of the Yes/No vocabulary test in SLA by Meara and Buxton (1987) used formats where the participants had to tick the words they claimed to know (see Appendix 1), which does not allow the identification of possible omitted responses[2]. The ambiguity in the format between "No"-responses and omitted responses could cause confusion when interpreting the test results. Meara's intention however (personal communication) was that the Yes/No test should be a forced choice test, where the possibility of non-responses is explicitly ruled out.

3.6.2 The Yes/No task

The simplicity of the assessment task is a big concern in tests of vocabulary size because the simpler the task, the larger the number of words that can be covered within the testing time available. The inevitable trade-off is that a simpler task reduces the quality of information elicited about the testee's knowledge of each word. In the case of the Yes/No format, the task with which the learner is confronted is not a test strictly speaking. It is situated in-

[2] From a terminological perspective, we would like to note that Lord (1980) distinguishes "omitted responses" (i.e. items that the participant read and decided not to answer) from "not-reached responses " (i.e. items at the end of the test that the participant did not reach due to lack of time), but we will not make this distinction and we use the term "omitted responses" to refer to both cases in this study.

between a conventional language test (i.e. characterised by verifiable responses) and self-assessment. A conventional test elicits answers to particular language tasks which are defined a priori and accordingly corrected. Self-assessment, however, is concerned with how learners judge their own ability in a particular skill (Oscarson 1997). The status of correct/false responses clearly differs between both situations. The fact that the Yes/No test cannot be seen as one or the other causes an ambiguity that taints the interpretation of the outcome of the test.

This ambiguity is perhaps most easily illustrated with an example from simple arithmetic: when someone is asked what the result is of seven multiplied by eight, the response will be either right (i.e. 56) or wrong (i.e. all other numbers). However, when someone is asked to indicate with "Yes" or "No" if he or she knows the result of seven multiplied by eight, this "Yes" or "No" response is not verifiable. Moreover, in order to answer the Yes/No question, there is the possibility that the participant's response will not only be based on his knowledge of mental arithmetic but that personal, cognitive and social factors come into play. For in the end, the participant has to make a decision, rather than give a correct response. The Yes/No test clearly has a decision criterion at the heart of the task which may endanger the format's validity.

3.6.3 Word selection

As far as word selection is concerned, it is not clear if one should favour certain word categories or leave out others. In a truly random selection, the Yes/No test will not only consist of verbs and nouns but also of numerals, conjunctions, prepositions etc. , the latter categories being harder to recognize since their meaning may depend more strongly on contextual clues.

3.6.4 Length of the test

The required length of the test in order to attain a representative estimate of the vocabulary size within a certain frequency range has perhaps been underestimated in the past. Meara (personal communication) suggests on the basis of his early work that 60 real words is too small a sample to be workable and currently recommends 180 words versus 120 pseudowords. On the basis of data Meara (personal communication) has obtained from the DIALANG project, 100 words versus 50 pseudowords seems to be a good compromise. However, the DIALANG Yes/No Test in its current form consists of only 50 words and 25 pseudowords.

3.6.5 Proportion words/pseudowords

The proportion of words and pseudowords in the test varies from one study to another. Meara and Buxton (1987) and Abels (1994) used 60 words and 40

pseudowords, Meara (1992) used 40 words and 20 pseudowords per frequency range, Hacquebord (1999) used 60 words and 30 pseudowords. In most of our experiments we have worked with the 60/40 proportion from the original Meara and Buxton study (1987), but the ideal or optimal proportion is still unclear.

3.6.6 The pseudowords

There are no clear guidelines for the construction of pseudowords. There seems to be a general consensus that the pseudowords should respect the phonotactic and morphological rules of the target language. This is why we prefer the term " pseudowords " to " non-words " (Read 1997) or " imaginary words " (Meara and Buxton 1987). Anderson and Freebody (1983: 236) were the first to use pseudowords in the Yes/No format and they created them according to two principles:
(1) changing one or two letters in a real word (e.g. "flirt" becomes "flort" and "perfume" becomes "porfame")
(2) forming uncoventional base plus affix combinations (e.g. "observement", "adjustion") which they call pseudoderivatives.
However, the extent to which pseudowords should differ from existing words remains unclear and the suggestion (Abels 1994) of changing more than one letter in a word in order to prevent that test takers would misread the pseudoword for the actual word is not "waterproof" since changing two or three letters in an actual word could create a pseudoword that differs only in one letter from another actual word. For instance, the Dutch verb "koken" (to cook) could be changed into the pseudoword "karen", which differs only one letter from the Dutch verb "varen" (to sail) or the Dutch noun "koren" (corn).

There are also sound objections to be made to the second pseudoword-formation principle. Anderson and Freebody (1983) found that almost all of the false alarms of their best scoring participants were pseudoderivatives. Apparently, the children that took part in the study were applying the word-formation rules of English to infer meanings for unfamiliar letter strings. One could argue that finding fault with L1 or L2 learners for accepting pseudoderivatives as existing words expresses a rather narrow view on the nature of language and its considerable generative morphological power. For L2 learners in particular it needs to be noted that nowadays learners are often encouraged to make use of knowledge of word building processes to relate unfamiliar words to known words or to known prefixes and suffixes. This is thought of as a creative activity that can help students learn hundreds of words in the target language. Learners are invited to see meaning patterns that lie behind the use of word parts and to take risks, which Anglin (1993) calls "morphological problem solving". Finally, we would like to note that the

characteristics of the pseudowords might also render the test format problematic for participants suffering from – even slight forms of – dyslexia.

3.6.7 The instruction

Little attention has been paid to the test instruction and its implications on the learner's choices. Several authors have pointed out that there are several levels to " knowing a word " (Richards 1976; Nation 1990; Read 1993). The issue of defining the precise construct of the Yes/No measure is of the utmost importance. Does recognizing a word as belonging to the Dutch language mean that you know this word? Is it possible to know a word but not being able to say what it means? Even if one assumes that the Yes/No test taps into a kind of fundamental knowledge of a word, this does not rule out the possibility of complex interaction between different test instructions and several levels of knowing a word. Whatever the definition of the construct that is intended to be measured in this test, the relationship between these levels and different test instructions should be examined.

3.6.8 Correction formulae

A problem shows up with the formulae used to calculate the test scores. The critical problem is how to get a precise estimate of vocabulary knowledge separate from the tendency to over- or underestimate this knowledge. The several formulae that have been proposed so far have been adapted either from the standard correction for guessing formula or from Signal Detection Theory. Although the general principle of reducing the test score according to the size of the false alarm rate remains the same in both cases, the precise way in which this reduction is executed (i.e. the way in which the response bias effects are dealt with), varies greatly from one approach to another. Consequently, different formulae applied to the same data may lead to very different results. The question remains which formula will result in the most meaningful test score.

In short, literature research and initial dealings with the Yes/No vocabulary test lead us to conclude that although the Yes/No vocabulary test has obvious attractions for vocabulary assessment in SLA and for school and classroom use, there are several design and analysis issues which need to be addressed if this type of test is to be considered a valid measure of second language vocabulary knowledge. Few of the drawbacks listed above have so far been investigated, especially from a measurement perspective. The problem of establishing an adequate scoring method is more than just one relevant issue among others, it constitutes a prerequisite in order to be able to address many of the aforementioned properties.

Chapter 4

Research context and research design

The research presented within the scope of this study arose from a need to select a receptive vocabulary test to include in the placement test procedure of the language centre of the Université Libre de Bruxelles. Gathering data within the field of applied linguistics is quite a challenge. A clean experimental design with participants who volunteer to take tests, would have made it easier to test hypotheses but such a set-up can never replicate a real language testing situation. In an experimental design with volunteers, the tests bear no consequences for the participants involved, whereas tests that are part of a language curriculum exert an important influence on the way the participants approach the test taking procedure in terms of motivation or fear of failure. Apart from Experiment 6 (see Chapter 8), in which we approached native speakers of Dutch, the experiments in this study were executed in a realistic language learning and language testing context. As we explained in Section 1.4 of the Introduction, the pragmatism that lies behind this research has its consequences for the way the experiments were organized and for how the data were collected.

In Section 4.1, we will briefly describe the testing tradition as it exists in the language centre where the data of this study were collected, then we will sketch the profile of the students we have worked with and we will throw light on the placement procedure. In Section 4.2, we will turn to a description of the research design that arose as a consequence of the research context. An extensive overview of all the reported experiments in this study is presented, together with a schematic inventory that can be consulted as a guideline throughout this study.

4.1 The ins and outs of the Brussels language centre

Contrary to secondary education in Belgium, which is highly regulated, colleges and universities can set their own standards of learning and evaluating. They can decide on their curriculum and quality control autonomously. This is also the case at the Université Libre de Bruxelles where this research was carried out. At the language centre of the university, language courses are organized for several faculties (political sciences, psychology, economics, business administration, etc) and the course contents are decided in agreement with the faculty concerned.

The centre's line of policy concerning the methods for language learning and language evaluation is that they should be in tune with current

developments in applied linguistics. All syllabus materials and all language tests are created by the centre's staff, and they are monitored regularly and altered when necessary. This procedure allows for a much more flexible organisation of the course content and method than if a specific manual were used. Moreover, it permits a nice fit with the content demands of the faculty or the needs demonstrated by the student population.

4.1.1 Focus on the evaluation of language skills

At the centre, a lot of attention goes to evaluation of language skills. The evaluation assignment at the language centre is threefold:

1) Each year hundreds of students have to be placed in the appropriate language courses.
2) Students that are taking courses need to be evaluated throughout the year.
3) Students have to pass language exams in order to be allowed to enter the next year of their university education.

The challenges presented by this enormous task have inspired the teachers and researchers to take an active interest in language evaluation techniques. New developments in language testing were followed up and the process of evaluation was regularly adjusted so as to include or try out new test formats which were presented or reported in scientific journals. This has created a tradition in which pragmatic test use often generated a line of research in a particular test format. As a result, the growing expertise in the centre in language teaching methods as well as language evaluation is strongly empirically based.

It is evident that such an approach can only be realized if the necessary (financial) support is provided. Resources have to be appropriated in order to centralize all examination data. In the center, a psychometrist is appointed to supervise all examination procedures and the quality of the test data for all language departments. Through years of experience this has amounted to a considerable expertise in the field of language testing. Gradually, the separate language departments have got imbued with the necessity of consistent evaluation. In consultation with the psychometrist, procedures were installed in order to enhance inter-rater reliability. For instance, teachers teaching parallel groups of a particular course have to agree on the same examination procedure; tests have to be corrected by one and the same teacher irrespective of the classes to which the test takers belong; interviews are always conducted by a teacher and a second assessor where it is the teacher's task to engage in the interview and the second assessor's task to construct a linguistic profile of the test taker's language use by means of a carefully constructed assessment sheet; etc.

4.1.2 Testing tradition

Although discrete-point tests (i.e. tests to assess whether learners have knowledge of particular structural elements of the target language) are still used in the placement procedure, the centre has gradually demonstrated a preference for what we would like to call "global proficiency testing", generally referred to as integrative tests in the literature. We use this term to refer to embedded language measures (as opposed to discrete measures) that contribute to the assessment of a larger construct instead of evaluating particular structural items of the language. This shift in the centre's testing policy is consistent with a more general trend in the language testing domain of the past thirty years. To assess proficiency discrete tests are abandoned in favour of performance-based tests where the students have to perform more holistic and authentic tasks. Bachman and Palmer (1996) decribe the task as the basic element in contemporary test design. The profiency tests that are used at the centre are, amongst others, the c-test, the cloze, the rational cloze, the transcription of authentic spoken text, etc.

One of the advantages of the testing tradition is that we have obtained a good notion of our students' response behaviour on the various formats they have been subjected to, not only in terms of their concrete performance on the tests, but also in terms of how they experience the tests and if they feel the tests succeed in reflecting their language competence. Usually, when a test is merciless in portraying their lack of proficiency (when it has a good discriminating power), they demonstrate a strong dislike for it. These reactions should not be ignored for they can play an important role in terms of motivation (and students should be presented the opportunity to give their best performance on a test) and they should be weighed against the psychometric qualities of the test. Some students can be considered to be manipulative when it comes to taking a test. Their extrinsic motivation may certainly influence their performance on tests. Shohamy (2001) argues that the power of tests originates from their capability of causing a change in behaviour in the test takers. She relates this phenomenon to relationships observed in economic models where producers and consumers take steps to maximize their profits and refers to Bourdieu's model "the economy of practice" (1999, cited in Shohamy 2001). In this model Bourdieu explains that various situations which may not be governed by strictly economic logic may none the less concur with a logic that is economic in a broader sense, because the individuals involved are oriented towards the acquisition of some kind of capital (e.g. cultural or symbolic capital) or the increase of some kind of symbolic 'profit' (e.g. honour or prestige). Following this train of thought, Shohamy puts forward that the test takers' desire to maximize scores on tests obeys an economic logic because it is their wish to maximize their scores in view of better job opportunities, increased salaries or gains in terms of recognition by teachers, parents and

peers, or getting the prestige and honour of being the best in the group (Shohamy 2001: 105,106).

For every course that is organized in the language center, the teacher decides in consultation with the psychometrist which test or combination of tests would be most appropriate. This decision is taken not only in function of the course goals and the practical restraints of the testing situation, but also the empirical experience we have gathered with a particular test format when we have used it with our student population. Language testers have reached the consensus that a test has to be chosen and implemented according to the goals, characteristics and specificities of the learning context (Bachman 1990, Read 2000). It goes without saying that in our testing practice at the centre, as well as in any testing practice around the world, there is an inevitable trade-off between the test's characteristics in terms of reliability and validity and the practical constraints of the testing situation.

4.1.3 Student population

The subjects whose data are reported in this dissertation are all Belgian French-speaking university students of Economics and Business Administration taking compulsory Dutch language courses as part of their curricula. They all share a history of learning Dutch as a compulsory L2 in primary and/or secondary school but the number of course hours they took and the levels they obtained vary greatly. This is partly due to the complex Belgian language policy concerning the language education in different parts of the country. Due to changes in the legislation[3] schools that are situated in the southern part of Belgium (la région Wallonne) are exempted from the obligation to organise Dutch courses. The local school authorities can decide to give priority to courses in English or German as a second language instead of Dutch. In Brussels, however, Dutch remains the compulsory second language course. Since the Université Libre de Bruxelles attracts students from all parts of the country, it will not take long before we are confronted with students who never attended Dutch courses and are in fact absolute beginners.

The students have a fixed curriculum and their most important courses are mathematics, statistics, economics and finance. We would describe them as "non-specialists" what their language skills are concerned because they certainly did not choose their university education in function of the language ingredient in their curriculum. However, some of them do realize the importance of languages for their future careers. They are generally highly competitive and rather extrinsically motivated (for a certain number of students it is more important to obtain their degree than to become proficient in a particular language). Nevertheless, the students face high demands on the part of their

[3] Décret portant sur l'organisation de l'enseignement maternel et primaire ordinaire et modifiant la réglementation de l'enseignement; 13 juillet 1998

faculty and the job market as far as their Dutch skills are concerned. Belgium is after all a multi-lingual country. Although the university is situated in Brussels, which is officially a bilingual city (French and Dutch), obtaining an advanced level in Dutch proves to be quite a challenge. A majority of the students never speaks Dutch outside the classroom, they consult Francophone media and their social network is almost exclusively monolingual. Due to political and historical reasons there may even be an attitude of contempt towards the Dutch language[4] (Van Hout and Knops 1988, Morelli et al. 1998).

4.1.4 The placement procedure

Our students' levels range from weak to advanced and a placement test is required to place them into homogeneous groups for their Dutch course in the second year of their studies. At the Université Libre de Bruxelles, the courses of Dutch and English for the faculties of Economics and Business Administration are organized from the students' second year on. This is a cost-cutting measure. Of all the students enrolled in the first year of their university study, about 50% drop out in the course of the year or fail their exams at the end of the year. Therefore, the university chooses to furnish the language courses in the second year since paying teachers to teach small classes (about 20 participants) is a costly affair. The placement tests for the Dutch and English language courses, however, are already administered to the first-year students in the beginning of the academic year. Their results serve to inform them on the possible lack of knowledge in certain language domains and they are strongly advised to brush up their English or Dutch before entering the subsequent year's language courses. The English and Dutch courses are lower intermediate courses and the students should see to it that they have attained a minimal level before entering the classes. Since most of them have been studying English and Dutch for about six years in secondary school, it seems to be legitimate not to organize beginner courses. The obtained scores on the placement test serve two purposes:

> (1) they give the student feedback on his or her level of the target language and when this is unsatisfactory, they are told how to improve their Dutch by the next year.
>
> (2) on the basis of the scores the students will be placed in the appropriate classes the subsequent year. This is a necessary measure to ensure the smooth operation of the language courses since the students' levels are so enormously divergent. Students that have sought tutoring or have worked in order to polish up their Dutch are offered the opportunity to take the placement test again before entering the course.

[4] Recent research seems to indicate that there is currently an evolution towards a more positive attitude with regard to the Dutch language (Mettewie 2003).

Contrary to the embedded approach we take towards language testing within the course context, the placement test procedure for Dutch consists of discrete language tests. This is justified in view of the number of students that have to be evaluated, the large differences in their levels of language ability, the time that is allocated for the placement procedure (about 45 minutes), the fact that the scoring needs to be swift and easy and the relatively low stakes of the testing situation.

At first the placement test for Dutch consisted of a grammar test which used to be a True/False test. Since 1999, the True/False format has been replaced by a 4 alternatives M.C. format whose items and distracters have been thoroughly analyzed. The items included in the test are neither too easy nor too difficult and have good discriminating power. The test has a very high reliability and correlates highly with other language tests.

In recent years, the teachers indicated that they were faced with an increasing lack of vocabulary knowledge in the Dutch courses which hampered the activities in class. Lexis was also one of the students' self-reported areas of weakness. Therefore it was reasonable that the placement procedure should include a vocabulary component as a means to more accurate placement. It was decided that receptive knowledge of high-frequency words of Dutch is a prerequisite in order to deal with the course's reading materials. Our aim was to measure whether our students knew the high frequency words that they are most likely to encounter and need.

Language testers agree that the purpose of the assessment has to serve as a guide to the selection of the appropriate features for the design of the test format. Whatever the purpose may be, a trade-off is always eminent in vocabulary test design. In our case, with a view to measuring vocabulary size within a placement procedure (about 500 students to be evaluated in a short time), a large sample of words needs to be covered within the testing time available in order to obtain a satisfactory estimate of the number of words known. This has as a consequence that the test task had better be simple, the words presented in isolation and it may be necessary to rely heavily on self-report. All these elements have consequences for the quality of information that one gathers with such an assessment tool. The test we were looking for should also be easy to administer and mark. The Yes/No Vocabulary Test seemed to match all these requirements. We used the format not so much to define what a student's vocabulary size is in absolute terms but to check whether a student knows the core vocabulary of Dutch (about 3700 words according to the corpus we have used, see Dieltjens et al., 1995, Dieltjens et al., 1997). This coincides with Chapelle's view that one should not just seek to measure vocabulary size in an absolute sense, but rather in relation to particular contexts of use (Chapelle 1994).

Similar to the Multiple Choice test, the Yes/No Vocabulary Test is a discrete test because it takes vocabulary as a distinct construct, separated from

other components of language competence. We began using the test at the Université Libre de Bruxelles in 1999.

4.2 Research design

The research design of this study was strongly determined by the language centre's daily functioning. Because of the markedly "applied" character of our research, some allowances had to be made when it came to collecting the data. In Experiment 1 and 5, the data were obtained through a placement test procedure. This had as a consequence that in these cases the Yes/No Tests could not be validated (time constraint) and that they could not be administered by computer (insufficient number of computers).

4.2.1 Short description of the experiments

The seven experiments that are reported in this study are summed up below. They are characterized according to their aim, design and number of participants. This description is intended to give the reader a bird's eye view of how the research was organized and it mentions the chapters in which the more detailed reports of the particular experiments can be found. The different materials or language samples that were used in the Yes/No tests of the respective experiments are presented in Appendices 3 to 8.

In order to provide information concerning the participants' level of proficiency in the Dutch language, we will use the same classification as the language centre's. Before the placement test is administered, no level is assigned and the group of students is heterogeneous. It consists of beginners, intermediate and advanced learners, and even native speakers. After taking the placement test, students are assigned to fairly homogeneous groups to take their first university courses of Dutch, which are called "cours du premier degrée" (first level courses). Native speakers are granted exemption from the Dutch courses and higher intermediate and advanced learners are allowed to skip the first year course. The first level course can therefore be labeled as a lower intermediate course. When the students pass their language exam at the end of the year, they can enter the second level course (cours du deuxième degree), which can be considered an intermediate course. Subsequently, they will enter the third level courses (cours du troisième degree) that can be labeled advanced courses. These third level courses are in fact organized into different thematic modules, from which the students can make their choice.

Experiment 1:
Experiment 1 centres around the first use of the Yes/No Vocabulary Test as part of the placement procedure at the Université Libre de Bruxelles. The test was administered together with a grammar multiple choice test to 488 French-speaking first-year students of Economics and Business administration. Their levels varied from "beginner" to "advanced learner". The test was intended to measure if the students knew the Dutch core vocabulary. The results of this first test use are described in Chapter 5 and the discussion that follows from it focuses on establishing a test score on the basis of the students' responses. The different correction formulae that have been proposed in the literature are illustrated and the reliability of the format is reconsidered on theoretical as well as on empirical grounds. The response bias revealed by the students (their willingness to accept a lot of pseudowords as "known" Dutch words) invited a series of experiments to attempt to validate the Yes/No Vocabulary Test. This is reported in the subsequent chapters.

Experiment 2:
This experiment was set up to collect empirical evidence concerning the validity of the Yes/No format. The experiment was conducted on computer and consisted of administering a Yes/No Vocabulary Test followed by a Translation task that contained the same words as the Yes/No Test. The central aim of the experiment was to examine the influence of different correction formulae on the correlation between the Yes/No Test results and the results of the Translation task. Test results were collected from 161 French-speaking university students of Economics and Business Administration, with language levels ranging from lower intermediate to advanced learner (language courses from the first, second and third level). The results are discussed in Chapter 6.

Experiment 3:
Since the previous experiments had shown that the Yes/No Test suffered from a response bias, which caused the reliabilities to be misleading because they reflected a consistent measure of the bias, a third experiment was designed with a view to reducing or eliminating the response bias in the data through modification of the instruction. Chapter 7 reports an experimental design in which a rather vague instruction is contrasted with a rigorous instruction while using identical test content. The Yes/No Test was administered in a paper-and-pencil format to 179 French-speaking university students of Economics and Business administration with lower intermediate levels (first level language course). After taking the test, both the control and the experimental group were presented a Translation task in order to verify if the Yes/No Test that was accompanied by the rigorous instruction resulted in better concurrent validity.

Experiment 4:
Experiment 4 centered on the possible influence of the computer interface design on the response behaviour of the participants. Two radically different computer interfaces were programmed and administered to 125 French-speaking university students of Economics and Business administration with lower intermediate levels (first level language course). Computer application A was designed to resemble the paper-and-pencil version of the Yes/No Vocabulary Test whereas Computer application B was set up to fully exploit the computer's potential to provide a controlled environment. The experiment tests the hypothesis that a more controlled environment could possibly reduce the response bias observed in the participants. Evidence of concurrent validity was obtained by means of a Translation task again. The experiment and its results are presented in the second part of Chapter 7.

Experiment 5:
In this experiment the role and quality of the test content is targeted. In order to rule out the possible hypothesis or criticism that the "homemade" language content of the Yes/No Tests used in the previous experiments might have tainted the data, the Yes/No format was infused with the content of the Yes/No test for Dutch from the European DIALANG test battery. Like in the first experiment, the Yes/No test was part of the placement test procedure which had as a consequence that there was no time to add a Translation task in order to obtain information concerning concurrent validity. However, the placement test also contained a grammar multiple choice test which served as a form of indirect validation. The tests were administered on paper to a group of 462 French-speaking university students of Economics and Business Administration with heterogeneous language levels (ranging from beginner to advanced). Their results are discussed in Chapter 8.

Experiment 6:
In order to evaluate the quality of the DIALANG test content for Dutch, a small-scale experiment was set up to collect native speakers' responses to this test. A paper-and-pencil version of the test was administered to 70 Dutch-speaking university students of Linguistics and Literature. Their erratic results revealed the problems they appeared to have in recognizing the existing Dutch words in the test. Item analyses were performed in order to distinguish good from bad test items for words as well as pseudowords. The results of this experiment are reported in Chapter 8.

Experiment 7:
In Experiment 7, two new formats for measuring vocabulary size were introduced - Recognition Based Vocabulary Test I and II - and contrasted in format, task and scoring with the Yes/No Vocabulary Test. The experimental design consisted of the three different tests that were presented to the

participants in three different but equivalent materials. Afterwards, all three formats were validated by means of a Translation task. The research questions centered around which of these formats gave the most accurate reflection of the participants' vocabulary knowledge and the average time span it took the participants to complete the tests.

The different tests and the Translation task were administered on computer to 177 French-speaking university students of Economics and Business Administration with lower intermediate levels (first level course). The discussion of the empirical data and the resulting conclusion concerning the most appropriate test for measuring receptive vocabulary size are presented in Chapter 9.

4.1.2 Schematic inventory

Exp.	**1**	**2**	**3**	**4**
Chapter	Chapter 5: Calculating test scores	Chapter 6: Concurrent validity	Chapter 7: Reducing the response bias	Chapter 7: Reducing the response bias
Aim	Comparing and discussing the different correction formulae of the Yes/No Test that are presented in the literature.	Establishing concurrent validity of the Yes/No Test by comparing the results with the participants' performance on a Translation task.	Investigating the influence of the instruction on the response behaviour of the test takers.	Investigating the influence of the interface design on the response behaviour of the participants.
N	488	161	179	125
Level	No level assigned yet, hetero-geneous group	First, second and third level students	First level students	First level students
Year	1999	1999	2000	2001
Format	Paper-and-pencil	Computer	Paper-and-pencil	Computer
Material	see Appendix 3	see Appendix 3	see Appendix 3	see Appendix 4
Validation	No data available	Translation	Translation	Translation

Exp.	**5**	**6**	**7**
Chapter	Chapter 8 : DIALANG	Chapter 8 : DIALANG	Chapter 9 : The Recognition Based Vocabulary Test.
Aim	Targeting the issue of test content with reference to the encountered response bias problem.	Investigating the quality of the DIALANG test content from a native speaker perspective.	Reporting on an experimental design in which a new vocabulary test and its characteristics are compared to those of the Yes/No Test.
N	450	70	177
Level	No level assigned yet, heterogeneous group	Native speakers	First level students
Year	2001	2001	2002
Format	Paper-and-pencil	Paper-and-pencil	Computer
Material	see Appendix 5	see Appendix 5	see Appendix 8
Validation	Grammar MC	No data available	Translation

Chapter 5

Calculating test scores

In this chapter we will report the first use we made of the Yes/No Vocabulary Test as part of a placement test procedure for French-speaking learners of Dutch. In particular, we will deal with the scoring problems that arise when using the Yes/No format as a measure of receptive vocabulary knowledge. The reliability and validity of the Yes/No format are re-assessed both by considering its theoretical grounds and by examining experimental data.

In Section 5.1 a study is presented in which a Yes/No test is used as the vocabulary section of a placement test, aimed at estimating how many high-frequency words of Dutch are known by French-speaking university students. The results of this test have led to an in-depth study of the currently proposed correction formulae in Section 5.2. In Section 5.3, we reach the conclusion that the reliability of the Yes/No test is overestimated because the test scores are contaminated by a response bias. It is not yet clear which correction formula is best suited to deal with this bias.

5.1 Experiment 1: First use of the Yes/No Vocabulary Test

5.1.1 Aim

With this first experiment, we aimed to evaluate the usefulness and surplus value of the Yes/No Test as part of the placement test procedure for Dutch. As already mentioned in Chapter 4, approximately 500 students have to be placed in the appropriate Dutch course each year at the language centre of the Université Libre de Bruxelles. Up until 1999, this placement test procedure consisted solely of a multiple choice grammar test. Because adequate knowledge of high-frequency words of Dutch is a prerequisite for dealing with the Dutch course's reading materials, a Yes/No Vocabulary Test was added. Although the Yes/No Vocabulary Test offers the possibility of selecting words according to frequency ranges and making inferences to the size of the learners' global receptive vocabulary knowledge (Meara 1992, Shillaw 1996), it was decided to use the test in relation to a more modest and well-specified aim: measuring the students' knowledge of the Dutch core vocabulary (approximately 3700 words, see Dieltjens et al., 1995, Dieltjens et al., 1997).

5.1.2 Method

Participants
The participants were Belgian French-speaking university students of Economics and Business Administration. Their levels ranged from "beginner" to "advanced learner" and a placement test was required to place them into homogeneous groups for the compulsory Dutch language course in their second year of university. The placement test was administered to 488 participants.

Placement test materials
The grammar test was composed from a large number of 4 alternatives M.C. items that were extensively used within the framework of the CALL-facilities (Computer Assisted Language Learning) of the language centre. On the basis of the automatically recorded difficulty index, 78 items were selected in order to obtain a suitable M.C. grammar test as part of the placement test. A Cronbach's alpha of about .90 was considered a threshold level. The test was administered in three forms (A, B, C), differing in item order only.

The Yes/No Vocabulary Test consisted of 60 words and 40 pseudowords, following the ratio of the original Yes/No test (Meara and Buxton 1987). All words, including those transformed into pseudowords, were taken from Woorden in Context (Dieltjens et al, 1995, Dieltjens et al. 1997), a standard work which contains 3,700 Dutch words selected on the basis of frequency and utility. All words were selected at random and therefore contained verbs, nouns and adjectives, as well as conjunctions, prepositions and numerals. Two parallel versions (I and II) of the test were created (see Appendix 3). Each test version (I and II) contained :

-25 words from the 1,000 word level
-25 words from the 1,000 up to the 2,000 word level
-50 words from the 2,000 up to the 3,700 word level.

The pseudowords were created according to the same word alteration principles as described by Anderson and Freebody (1983) and applied by Abels (1994) and Van De Walle (1999) in their respective uses of the Yes/No test for Dutch:

-The first procedure consists in changing the affixes of an existing word: 11 pseudowords (i.e. the number of words in the sample which permitted this kind of change) were created like this. Example: *prettig* (fun) gets turned into *pretachtig*.

-The second principle is the substitution of one or two graphemes without breaching the phonotactic and morphological rules for word formation in Dutch: the remaining 29 pseudowords were created according to this procedure. Example: *timmerman* (carpenter) gets turned into *tommerman*.

Occasionally, applying one word formation procedure can result in a pseudoword that could also have been obtained by using an alternative procedure. It should be noted that, in order to preserve the universal properties of the format, every language teacher (native speaker or not) should be able to create a Yes/No test in the target language by following a few simple rules.

In order to control for sequence effects and to eliminate the possibility of cheating, three forms (A,B,C), differing in item order only, were created for each sample. The following assignment was given in the participants' L1: *Indiquez à l'aide d'une croix les mots que vous connaissez. Certains mots repris dans la liste n'existent pas en néerlandais!* (Tick the words you know. Certain words figuring in the list do not exist in Dutch.). All students completed the paper-and-pencil test in less than 10 minutes.

5.1.3 Results

In Table 5.1, it can be observed that the reliability scores of the MC grammar test seemed satisfactory (around .90). The mean scores of the MC were consistent with those of previous years. The individual scores varied greatly (which is to be expected in placement testing) as can be concluded from the considerably large standard deviations.

Table 5.1: Descriptive statistics of the results on the Grammar M.C.

Grammar M.C. 78 items (4 alternatives)					
Order	N (488)	Scores	Mean	SD	Reliability
A	166	Raw	32.93	12.08	.890
		Corrected	18.93	15.36	.882
B	162	Raw	32.66	12.67	.901
		Corrected	18.35	16.01	.890
C	160	Raw	35.09	12.26	.892
		Corrected	21.44	15.81	.886

Notes: A, B and C are three different item orders. Means and standard deviations are presented for both raw and corrected scores. Raw scores are the number of correct responses, corrected scores are calculated with the classic correction for blind guessing-formula (cfbg) and the reliability is calculated with Cronbach's alpha.

Although the reliabilities of the two versions of the Yes/No Vocabulary Test were sufficiently high (see Table 5.2), there were some unsettling phenomena in the data which caused the transformation from raw scores into corrected scores to be problematic:

1) Despite the explicit warning in the instruction about the presence of pseudowords in the test, many participants displayed a high rate of false alarms in their responses (20%, which means that the average participant claimed to know 8 out of 40 pseudowords) (see Figure 5.1). This created problems for the scoring of the test since such high false alarm scores reduce the test scores dramatically (see Table 5.2).

Table 5.2: Descriptive statistics of the results on the Yes/No Vocabulary Test

Yes/No Vocabulary Test		(100 items: 60 words – 40 pseudowords)				
Version	Order	N (488)	Scores	Mean	SD	Reliability
I	A	78	Raw	71.19	9.24	.818
			Corr.	42.38	18.47	.818
	B	79	Raw	71.96	7.97	.771
			Corr.	43.92	15.95	.771
	C	78	Raw	74.60	9.34	.848
			Corr.	49.21	18.68	.848
II	A	89	Raw	73.06	8.77	.826
			Corr.	46.11	17.54	.826
	B	82	Raw	71.90	10.68	.877
			Corr.	43.80	21.35	.877
	C	82	Raw	71.70	9.40	.841
			Corr.	43.39	18.81	.841

Notes: I and II are two different versions of the Yes/No test i.e. they are made up of different items. A, B and C are three different orders of any given test version. Raw scores are the number of correct responses, corrected scores are calculated with the classic correction for blind guessing (cfbg) and the reliability is calculated with Cronbach's alpha.

Item alternative	Response alternative: Yes	Response alternative: No
Word	Hit 68.0%	Miss 32.0%
Pseudoword	False alarm 20.5%	Correct rejection 79.5%

Figure 5.1: The item-response matrix of the Yes/No test in Experiment 1. Percentages are calculated within each item alternative.

It should be remarked that the published data on the acceptable false alarm rate are rather conspicuous. Meara and Jones (1990) report that a good percentage of their test takers do not claim to know any of the pseudowords, which is certainly not the case in this experiment. On the other hand, in the guidelines furnished with the paper-and-pencil EFL Yes/No Vocabulary Test (Meara 1992), a level of 10 false alarms out of 20 pseudowords was given as a boundary beyond which the test results become unreliable. With less than 25% false alarms in the data, most of the test takers in the experiment under question did

not exceed this boundary. In figure 5.2, the distribution of false alarms in Experiment 1 is illustrated. Almost half of the population (48%) claims to know the meaning of 7 or more pseudowords in the test.

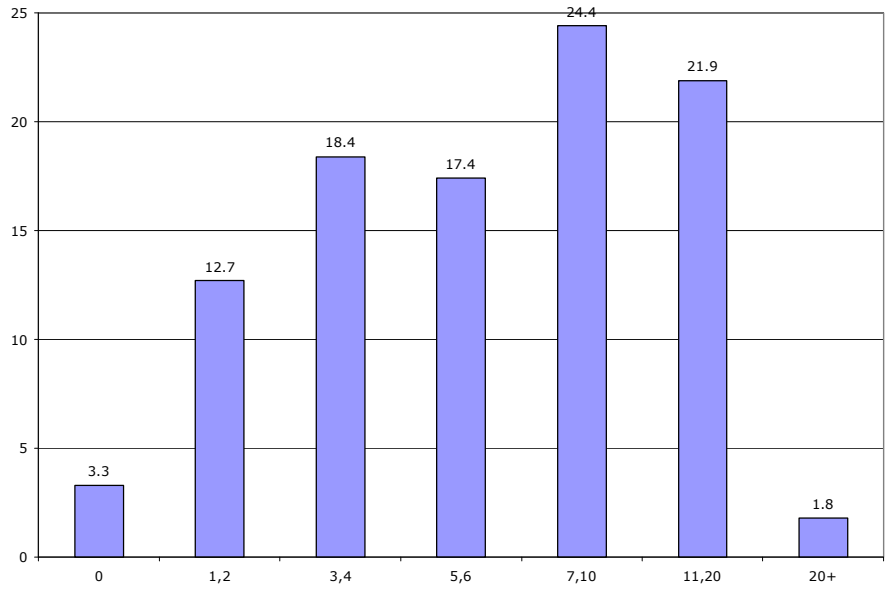


Figure 5.2: The distribution of the false alarms in Experiment 1. The X-axis represents the number of false alarms (on a total of 40 pseudowords) divided into 7 groups. The Y-axis represents the number of participants in percentages.

To our knowledge, the distribution of the false alarms with reference to the level of the participants is not looked into in the literature. But when confronted with high false alarm rates like these, one might presume that weaker students will have had more problems rejecting the pseudowords than more proficient ones. Moreover, the formulae proposed in the literature to calculate a score for the Yes/No Test are clearly based on the principle that being able to distinguish between words and pseudowords is at the heart of the Yes/No Task. The high false alarm rates we encountered in this experiment were not confined to weak students. This means that strong participants did not do a better job at distinguishing words from pseudowords than their weaker peers.

2) There appeared to be a negative correlation (-.37 for Test Version I and -.26 for Test Version II) between the measure of the performance on words and the measure of the performance on pseudowords (see Table 5.3). This means that there was an inverse relationship between the ability to identify words and the ability to reject pseudowords. We found this very disturbing.

Table 5.3: Correlation between scores on words versus pseudowords of the Yes/No Vocabulary Test

Version	Order	Scores /100	Reliability	Correlation w/pw
I	A (N=78)	Raw	.818	-.409***
	B (N=79)	Raw	.771	-.414***
	C (N=78)	Raw	.848	-.353**
	Total (N=235)	Raw	.820	-.373***
II	A (N=89)	Raw	.826	-.397***
	B (N=82)	Raw	.877	-.089
	C (N=82)	Raw	.841	-.287**
	Total (N=253)	Raw	.850	-.264***

Notes: Raw scores are the number of correct responses on the 100 items, words and pseudowords and test reliability is calculated with Cronbach's alpha. Significant correlations are marked with * ($p<.05$), ** ($p<.01$) and *** ($p<.001$).

As far as we know, the correlation between the measure of the performance on words and the measure of the performance on pseudowords was never mentioned in previous reports of Yes/No test use. There are three possible correlational relationships between the scores on the words and the scores on the pseudowords. In view of the original intention of the Yes/No format and the purpose the pseudowords are supposed to fulfill within the format, the correlation between them should be zero. This would mean that there is no relation between the participants' scores on the words and their scores on the pseudowords. A positive correlation between both measures would mean that participants with high scores on the identification of words, reject most or all of the pseudowords. However, a negative correlation between both measures (high scores on the identification of words and low scores on rejecting the pseudowords) can only be explained by the presence of a response bias in the data. We define response bias as the tendency to prefer one particular response in case of doubt as a result of other factors than vocabulary proficiency (cognitive make-up, personal profile, etc.). When a participant has a tendency to respond "Yes", this will have a positive effect on the score for words and a negative effect on the score for pseudowords. When a participant has a tendency to respond "No", this will have a negative effect on the score for words and a positive effect on the score for the pseudowords. When these tendencies are cumulated within a group of participants, this will result in a negative correlation between the measure of the score for words and the measure of the score for pseudowords.

Both findings (the high false alarm rate and the presence of a response bias) have led to a careful examination of the impact of different correction formulae on the psychometric qualities of the Yes/No test because we feared that the contamination of the test results by a response bias might have caused

an overestimation of the global test reliability[5]. In the following section we will explain the difference between formulae based on either a discrete or a continuous model and we will illustrate how extracting the response bias from the raw score results in a severe drop in reliability.

5.2 An Investigation of Correction Formulae

A detailed comparative review of the different correction formulae proposed in the literature for transforming raw Yes/No scores has been made by Huibregtse and Admiraal (1999) and Huibregtse, Admiraal and Meara (2002). The use of the different formulae led to corrected scores which, at least theoretically, could lead to dramatic differences. In the approach we will present here the impact of different correction formulae on the results of the Yes/No test will be re-examined.

The empirical data we collected from the participants are transformed into corrected test scores[6] by using formulae based on either discrete or continuous models. In this section, four formulae will be thoroughly examined on both theoretical and empirical grounds:

-cfbg: Correction for blind guessing (discrete model)
-cfg: Correction for guessing (discrete model)
-I_{SDT}: Index based on SDT (continuous model) (Huibregtse & Admiraal, 1999)
-Hcfb: Hits corrected for bias (continuous model) (Beeckmans et al., 2001)

We will not discuss Meara's Δm (Meara 1992) since this has already been done in a very complete and convincing way by Huibregtse and Admiraal (1999). This formula calculates the proportion of hits a participant would have scored if he or she had refrained from responding "Yes" to pseudowords. Huybregtse and Admiraal (1999) have illustrated that the score generated by this formula approaches zero in the case of weak performances. They also report that in the case of few hits, small differences in performance can lead to seriously divergent scores. When, for instance, half of the words elicit a "yes" response and no pseudowords result in a "yes" response, then the score amounts to 0.50. However, when there is one "yes" response to a pseudoword, the score diminished to 0.37. The difference between both scores is more than 10% of the total score range which means that "yes" responses to pseudowords are severely penalized by this formula. It is clear that Meara's Δm may yield underestimated scores with a particular kind of response behaviour. They have also argued that the formula does not correct for individual response style and that although it is based on Signal Detection Theory, it still rests on the all-or-

[5] Because of the presence of a response bias in the data, we refrained from using the Yes/No results as a placement indication. The students were placed into the appropriate language classes on the basis of the results of the Grammar M.C.

[6] The four types of responses of the raw data matrix have to be transformed into one representative test score.

nothing assumptions of discrete models rather than on continuous modelling. Therefore we have opted to calculate the test scores in this study with the I_{SDT} correction formula that was developed by Huibregste and Admiraal (1999) and Huibregtse, Admiraal and Meara (2002), which takes into account guessing as well as personal response style, and is therefore considered an improvement to the Δm formula. They have shown that the Δm formula and the improved index I_{SDT} are linked in a monotonic - but not linear - relation and that the I_{SDT}, contrary to the Δm, produces a representative score for every type of response behaviour.

The discussion below focuses on a comparison between discrete versus continuous models which could be applied to our data. Other test formats (M.C. and True/False tests) and the empirical data they provided for similar student populations will serve to illustrate a discussion in which we will show that correction formulae used for M.C. and True/False tests are not necessarily applicable to scores of a Yes/No test. We will start with a methodological discussion about the distinction between the correction for guessing formula used with classical M.C. tests (which we will prefer to call correction for "blind" guessing, cfbg) and the apparently similar formula used with the Yes/No test (correction for guessing, cfg).

5.2.1 Discrete models

Correction formulae that are based on discrete models rest on two all-or-nothing hypotheses:
Hypothesis 1: The participant either knows or does not know the answer. There is nothing in between (therefore it is called an all-or-nothing or discrete model).
Hypothesis 2: If the participant knows the answer, his choice will evidently be correct. If the participant does not know the answer, he will either refrain from answering or resort to a blind guess. In this case, the participant has a chance of 1/k of hitting the correct answer, k being the total number of choices.

Correction for "blind" guessing (cfbg)

A) Applying the cfbg to the M.C. format

The correction for guessing formula applied to the raw scores of the multiple choice grammar test is widely used in the field of language testing. As will be explained, it consists in a correction for blind guessing (cfbg), which is not the case with other formulae that bear the ambiguous "correction for guessing" label. The aim of this correction is to take into account the fact that participants have a good chance to obtain the correct response by guessing, in which case the accounted credit fails to reflect participants' real knowledge. The final score

will therefore ultimately result in an overestimation of what is intended to be measured. The theoretical model behind the transformation from raw scores (number of correct responses) into corrected scores (number of items really known by the participants) rests on two all-or-nothing hypotheses mentioned above.

These two assumptions allow a corrected score to be inferred unequivocally from the observable data, which may be interpreted as the number of known items. This is illustrated for the particular case of a 4-alternative M.C. in Figure 5.3.

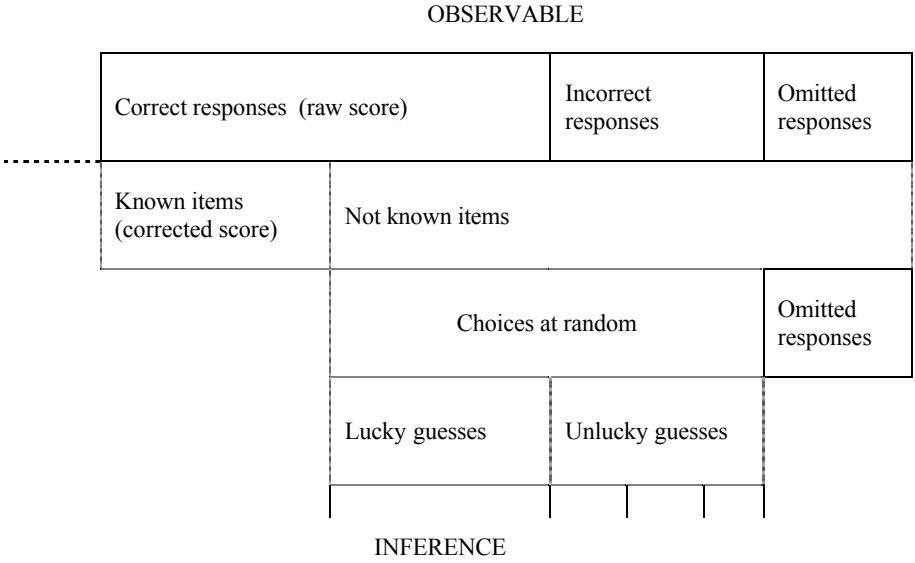


Figure 5.3: Inferring the corrected score from the observable data in the case of a 4 alternatives M.C. test using the cfbg (correction for blind guessing) formula. According to this model, when the participant decides to answer an unknown item, the probability of getting a lucky guess depends only on the number of alternatives.

The observable data collected for one participant can be distributed into three separate categories: the correct responses (the number of items within this class equals the raw score), the incorrect responses and the items that remain unanswered by the participant. In applying the two all-or-nothing assumptions, the data are divided in two different classes: the items which are actually known by the participant and those which are not. The first class corresponds to the corrected score we look for. The second class can be subdivided into two new sub-categories: the items for which the participant made a choice strictly at random (i.e. blind guess) and those the participant left unanswered (i.e. omitted

response). Finally, the first of these sub-categories may again be subdivided into lucky guesses that result in an observable correct response and unlucky guesses that lead to an observable incorrect response. Figure 5.3 illustrates that the number of lucky guesses equals 1/3 (1/k-1 in general, with k representing the number of alternatives) of the number of unlucky guesses. Since the number of unlucky guesses equals the observable number of incorrect items, the corrected score can be computed by simply subtracting 1/3 of the incorrect responses from the raw score.

The use of the cfbg formula leads to qualitatively different results depending on the test format and the test conditions. If, with a computerised version for example, a response to each item is required (forced decision task), the omitted response category is automatically ruled out and the formula is then reduced to a simple linear transformation of the raw score[7]. The rank order of the testees therefore remains unchanged. Whether the test reliability is computed from the raw data or from the corrected scores bears no consequence. The only implication of the transformation is that it provides a different scale in absolute terms which may be of interest only in a criterion-referenced approach. It should be noticed that the weaker the participant, the more this formula will reduce his/her score.

On the other hand, as far as the classical situation of a paper-pencil test is concerned, the presence of a considerable number of unanswered items for several testees makes the use of the formula more imperative and it will result in noticeable differences in the learners' rank order. The larger the individual differences in responding or not responding to the unknown items, the more the testees' rank order will be distorted. A poor correlation between these individual differences and the proficiency level will also increase the discrepancy in rank orders between raw and corrected scores. In other words, taking into account those items which were not answered by the participant is at the heart of the transformation. As can be seen from the results of the grammar M.C., our population shows large between-participant variation in answering behaviour. Table 5.4 shows the distribution of the omitted response category among the learners. About 1/4 of the students did not respond to all the items and this at various degrees. Because of the statistical variability added by the transformation from raw to corrected scores, the risk of a large decrease in the test's reliability cannot be excluded. Comparison between Cronbach's Alpha calculated with raw scores (.894) versus corrected scores (.886), however, shows that this decrease is insignificant for the M.C. (chi square analysis, M-

[7] Generally, the corrected score is a function of both the number of correct responses and the number of omitted responses. Two testees with the same number of correct responses may end up with different corrected scores depending on their respective number of omitted responses. Clearly this will influence the testees' rank order. In the case that the omitted response class is non-existent, the corrected score becomes a function of the number of correct responses (the raw score) solely.

value[8] :.14, df=1, p>.70, see Rietveld and Van Hout 1993: 204). Detailed results (see Table 5.1) confirm this for each of the three forms (A,B,C). As there is no decrease in reliability, corrected scores obtained with the cfbg are to be considered the most appropriate in ranking students who do not answer all items while maintaining a sufficient overall measurement accuracy.

Table 5.4: Frequency distribution of the number of omitted items obtained for the 3 orders (A,B,C) of the Grammar M.C.

Number of omitted items	Number of participants			Participant percentage		
	A (N=166)	B (N=162)	C (N=160)	A	B	C
[0]	129	117	128	77.7	72.2	80.0
[1,2]	8	17	11	4.8	10.5	6.9
[3,10]	10	8	8	6.0	4.9	5.0
[11,20]	6	4	4	3.6	2.5	2.5
[21,30]	6	8	3	3.6	4.9	1.9
[31,40]	6	5	4	3.6	3.1	2.5
[41,50]	0	0	1	0.0	0.0	0.6
[51,78]	1	3	1	0.6	1.9	0.6

B) Applying the cfbg to the True/False format

When the cfbg is applied to the results of a True/False test, all before-mentioned claims remain relevant. The True/False format may be considered as a particular case of a M.C. with two alternatives. However, it should be pointed out that the probability of a blind guess reaches .50. It enlarges the correction factor and it adds a greater statistical variability. Consider, for example, two learners of the same proficiency level (20 known items out of 100): the first one refrains from answering the 80 unknown items, the other one answers all unknown items at random. In the case of a 4-alternative M.C., the difference between both participants' raw scores would be 20 versus 40 (20 + 80/4). In the case of a True/False format, the difference would reach 20 versus 60 (20 + 80/2). However, the variability added becomes larger: the error

[8] The test statistic M is calculated to assess the significance of differences between two or more reliability coefficients. It has an approximate X^2 distribution with df= number of alpha coefficients –1 (Rietveld and Van Hout 1993).

in estimating 80/2 will be twice that of estimating 80/4. This example illustrates that it is even more important to be aware of possible distortions when a True/False Test is concerned but that care must be taken of a possible lack of test reliability when scores are corrected for guessing.

A second point of interest with the True/False format concerns the relation between the performances of participants on the true versus false items. A first question is whether or not the participants exhibit a difference in performance between both kinds of questions. Therefore the correlation between both scores can be examined in comparison with the theoretical value expected under the hypothesis of no difference in behaviour between true versus false items. The Spearman-Brown formula provides a means of calculating the reliability of half a test $alpha_h$ from the entire test's reliability $alpha_e$. The formula in this case is simply : $alpha_h = (alpha_e) / (2\text{-}alpha_e)$

Assuming there is no difference in what is measured by the two parts, $alpha_h$ has been proved to equal the correlation between the scores on both half-tests (Nunnally and Bernstein 1994). An experimental verification can also be carried out by directly comparing the obtained correlation with the average of a set of correlations between scores obtained by randomly splitting the test into half-parts. If the assumption holds (i.e. there is no difference), the correlation should not differ substantially from the theoretical value computed with the formula and neither should it differ from the average of real correlations computed with random split. Analysis of data that we have gathered throughout our use of the True/False format as a grammar test has shown that this is not the case (for a full investigation, see Beeckmans et al. 2001). Participants' performances differ when confronted with true versus false items. It is therefore clear that the all-or-nothing assumptions of the discrete model suffer from a lack of realism.

C) Applying the cfbg to the Yes/No format

When we consider the Yes/No test as a particular case of the True/False format, the use of the cfbg formula raises further specific questions:
1) The proportion of real words versus pseudowords varies from one published study to another. In all cases, the real words are more frequent, which complicates the assumptions related to random guessing. If the participant has the feeling that there are more real words than pseudowords, or if the participant decides to systematically give one response when he or she does not know, the hypothesis of a probability of .50 becomes inadequate. On the other hand, constructing the test with an equal proportion of real words versus pseudowords would make the format less economical because fewer words could be tested in the same time span.
2) In its original form (where the participant is asked to tick the words), the distinction between false response and omitted response is not possible for the

word items: if a participant has not ticked a word, this could either mean that he does not know the word or that he decided not to answer this item. The fact that it is impossible to distinguish between both alternatives undermines the central principle underlying the cfbg formula. Remember that, among the items which are not correct, the boundary between false responses and omitted responses is crucial (Figure 5.3). Moving the boundary to the very left will increase the corrected score. Conversely, moving the boundary towards fewer omitted items will decrease the corrected score. This variation, which is controlled in the case of the True/False test, cannot be controlled in the original Yes/No format.

3) The most important drawback of the classical correction for guessing concerns the nature of the task involved in the Yes/No test in comparison with the True/False test. In the latter case, the presence of a possible bias towards one or the other of the two responses can be considered as being part of the task. In the True/False grammar test, for example, a participant who tends to use only simple structures and tries very hard not to make mistakes, would exhibit a bias in judging many items to be incorrect. One could argue that this bias is part of the task and relevant with reference to the competence that is measured. On the other hand, in a Yes/No vocabulary test, the participant's task is closer to self-assessment than to a real language task. The bias can therefore only be attributed to factors which are beyond the competence of the participant.

In a study with a similar student population, Janssens (1999) showed that the students display a clear tendency of not being able to estimate their language proficiency accurately as far as vocabulary is concerned. The experiment was set up to check whether the students were able to use contextual clues to infer the meaning of words they did not know. First, the participants were presented with a list of target words and were asked to give the French translation (a). Second, the participants received a short text containing the target words and were asked to underline the words they did not know (b). Finally they got the text plus the target words and were asked to translate the words once again (c). Comparing (b) and (c) provides a means of evaluating students' self-assessment. Most students (69%) had a tendency to overestimate their vocabulary knowledge and there were large individual differences in their self-evaluation which were not due to their differences in language competence.

A procedure similar to the one described for the True/False test was applied to the results on the Yes/No test (Table 5.5) in order to gain more insight into the possible existence of a response bias in the data. In addition to splitting the whole test into halves, it was also split into two uneven parts composed of 60 and 40 items (because of the unequal proportion of words versus pseudowords). This was done in order to allow for a meaningful

comparison between the correlation obtained by the random split and the one obtained by dividing the test in 60 words versus 40 pseudowords.

Table 5.5: Correlation between scores on words versus pseudowords of the Yes/No Vocabulary Test compared with theoretical half-test reliability (*), estimated half-test reliability (**) and estimated part-test reliability (***).

Yes/No Test		Scores /100	Test reliability	Half-test reliability		Part-test (60-40) reliability	Correlation w/pw
Vers.	Order		Cronbach's Alpha	Spearman-Brown (*)	Split-half k=50 mean [SD] (**)	Split-part k=50 mean [SD] (***)	
I	A (N=78)	Raw corr.	.818 idem	.692 idem	.723 [.052] idem	.726 [.047] idem	-.409***
	B (N=79)	Raw corr.	.771 idem	.627 idem	.664 [.052] idem	.635 [.054] idem	-.414***
	C (N=78)	Raw corr.	.848 idem	.736 idem	.761 [.030] idem	.757 [.040] idem	-.353**
Total (N=235)		Raw corr.	.820 idem	.695 idem	.722 [.032] idem	.714 [.036] idem	-.373***
II	A (N=89)	Raw corr.	.826 idem	.704 idem	.736 [.033] idem	.720 [.037] idem	-.397***
	B (N=82)	Raw corr.	.877 idem	.781 idem	.797 [.032] idem	.787 [.033] idem	-.089
	C (N=82)	Raw corr.	.841 idem	.726 idem	.746 [.041] idem	.745 [.037] idem	-.287**
Total (N=253)		Raw corr.	.850 idem	.739 idem	.759 [.023] idem	.751 [.023] idem	-.264***

Notes: Raw scores are the number of correct responses on the 100 items, words and pseudowords. Corrected scores are calculated with the cfbg formula. If there were no difference between what is measured by word versus pseudowords, the correlation should equal the half-test reliability. Significant correlations are marked with * ($p<.05$), ** ($p<.01$) and *** ($p<.001$).

Both procedures for inferring correlations between two complementary parts of the entire test (Spearman-Brown formula and random splitting, see Beeckmans et al. 2001 for a description of the method) yielded very similar

results. On average, r = .70 was obtained with Test Version I and r = .74 with Test Version II. The standard deviations of the 50 correlations obtained by random splitting were very low as well. No difference was obtained between the results for both splits (half-test split versus part-test split) and these reliabilities were very close (.02 difference on average) to the corresponding reliabilities computed with the Spearman-Brown formula.

The most revealing result concerns the negative correlations that were systematically obtained between partial scores on word items versus pseudoword items. Only the assumption of a bias can reasonably account for this systematic negative correlation. Again we use the term "bias" in its ordinary sense, i.e. a tendency for a given participant to provide more/fewer responses of one type (true or yes) than of the other (false or no). A difference in discriminability between the two item categories or the fact that the two item classes measure substantially different skills could result in a decrease in the correlation but it could not render it negative. The existence of a bias, however, would automatically lead towards a negative correlation, for the bias has the particularity that it works in opposite directions at the same time. An individual bias towards " Yes " responses will produce an increase in the partial score for the words together with a decrease in the partial score for the pseudowords, and vice versa. Since the correlation is, in fact, negative, there is a substantial possibility that the test will, in our case, measure the response bias itself. As has already been pointed out by Huibregtse and Admiraal (1999), the correction for guessing (what we call cfbg) does not help to eliminate a response bias.

Correction for guessing (cfg)

So far, we have considered the correction for guessing by following the logic of a technique (cfbg) which has been developed in the specific domain of testing. We started with the M.C., moved on to the True/False, and the Yes/No was thus considered as a particular case of these classical tests. Meara's initial approach to the Yes/No vocabulary test, however, has been somewhat different. His goal was to obtain a sensible measure of the proportion of words a participant knows and the pseudowords were added solely with the aim of correcting the obtained proportion of hits. Rather than considering the whole set of data (i.e. the raw score consists of the number of correct responses, both words and pseudowords), the raw score of interest is limited to the number of hits (60 items and not 100) so that the corrected rejections are not included. This score is then corrected by a formula which is unfortunately also called correction for guessing (cfg). This formula resembles the previously discussed cfbg formula to some extent but it differs in some other important respects.

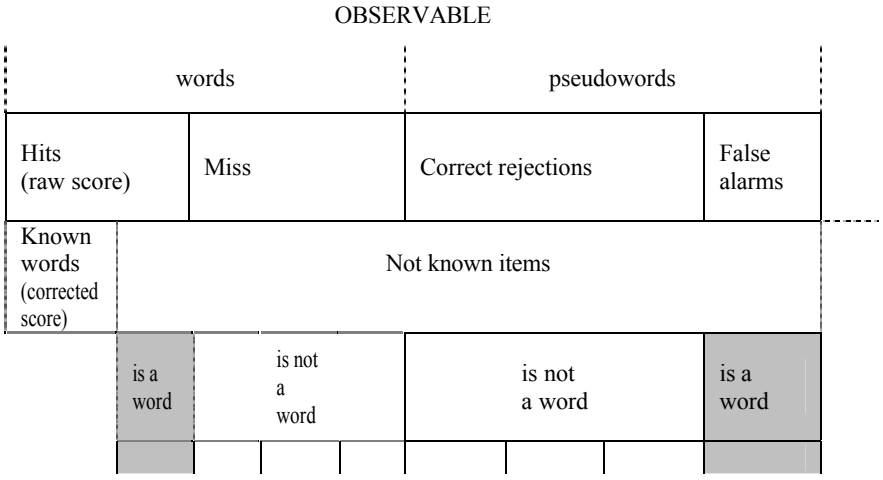
Figure 5.4 illustrates the principle of the cfg and may usefully be compared with Figure 5.3 (cfbg). To simplify the discussion, we will not consider the possibility of the omitted response category, which is irrelevant in

the case of the classical Yes/No, any further. What is common to both cfbg and cfg formulae is the first all-or-nothing assumption stating that the participant knows or does not know the answer, and that there is nothing in between. However, in the case of the cfg model, the possibility of knowing the answer is limited to the category of words only. Knowing that a pseudoword is not a word is ruled out by the model and therefore this possibility is ignored. It follows that the set of items can be subdivided into two categories: the words actually known by the participant and the rest of the items, that is, both the words the participant does not know and the entire set of pseudowords.

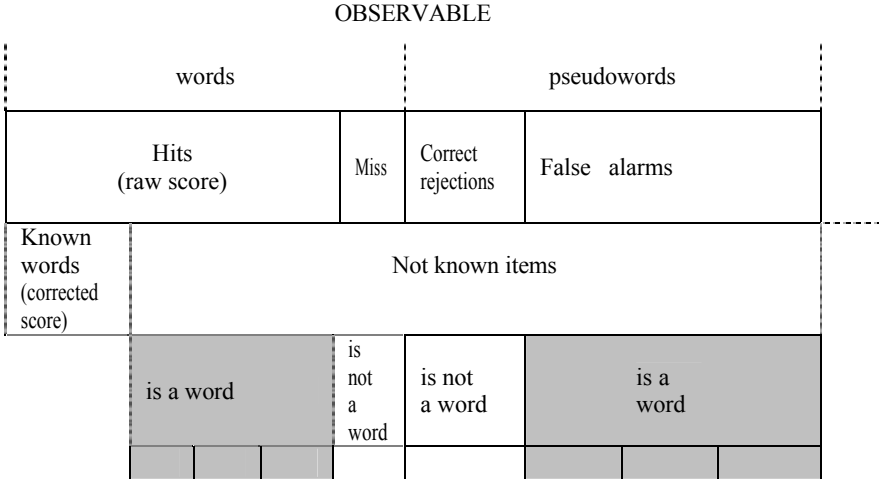
The second assumption also remains the same in its first part: "When the participant knows the answer, his choice is evidently correct" but again restricted to words since " knowing the answer " is now limited to " knowing the word ". The major difference consists in the way of estimating what the data will be in case of guessing. It is important to remember that in both models, cfbg and cfg, when the participant is guessing, nothing about any feature of the item which could be relevant to the measured competence can come into play. In the previous case (cfbg), whatever the participant's strategy when guessing (always responding true or responding alternately true and false, etc.), the usual methodological precautions in designing the test format will ensure that .50 is an unbiased estimate of the probability of getting the correct response. In other words, individual response bias will not contaminate the results. The data of Figure 5.3 do not distinguish between a participant who may have systematically responded " True " and a participant who may have systematically responded " False ", as far as the unknown items are concerned. By contrast, with the Yes/No format, the cfg model will lead to different raw scores for two participants who know the same number of words but who display different decision behaviour in responding to the unknown items. In the examples of Figure 5.4, participant A exhibits a response bias which leads to a rate of 1/4 words responses out of the unknown items, while participant B exhibits a rate of 3/4 words responses, and as can be seen both participants' raw scores are very different.

Figure 5.4: Inferring the corrected score from the observable data in the case of the Yes/No Vocabulary Test using the cfg (correction for guessing) formula. According to this model, when the participant is confronted with items he/she does not know, the participant will guess one of the two alternatives in a certain proportion which is specific for this participant and independent of the nature of the unknown item (word or pseudoword). In this example, this proportion is 1 (is a word) to 3 (is not a word) for participant A and 3 to 1 for participant B. Both participants know the same number of words but differ in their raw scores in accordance with their specific response biases.

Participant A



Participant B



In conclusion, it appears that the cfg model actually does take into account the individual response bias. However, the way in which this bias is evaluated depends largely on the all-or-nothing assumption underlying the model. A comparison with the continuous models that we will describe in Section 5.2.2 will make the theoretical drawbacks of the cfg more apparent.

The presence of large biases in our student population was already assessed by the negative correlation obtained between the participants' performances on the words and pseudowords (Table 5.3). Applying the cfg formula should logically lead to a decrease in reliability. The results presented in Table 5.6 confirm this prediction.

Table 5.6: Effect of the correction for guessing (cfg) on the Yes/No test reliability when the score is limited to the 60 words.

Yes/No test		Scores	Test reliability	Half-test reliability		Part-test reliability (60-40)
Vers.	Order		Cronbach's Alpha	Spearman-Brown	Split-half k=50 mean [SD]	Split-part k=50 mean [SD]
I	A (N=78)	Raw	.910	.835	.841 [.027]	.837 [.024]
		corr. (cfg)	.842	<<	.727 [.055]	.738 [.046]
	B (N=79)	Raw	.884	.792	.802 [.029]	.794 [.036]
		corr. (cfg)	.819	<<	.694 [.048]	.663 [.064]
	C (N=78)	Raw	.920	.852	.858 [.021]	.852 [.026]
		corr. (cfg)	.867	<<	.765 [.034]	.758 [.042]
Total (N=235)		Raw	.906	.828	.835 [.017]	.859 [.021]
		corr. (cfg)	.843	<<	.728 [.033]	.721 [.039]
II	A (N=89)	Raw	.909	.833	.841 [.022]	.827 [.024]
		corr. (cfg)	.845	<<	.732 [.040]	.722 [.044]
	B (N=82)	Raw	.914	.842	.849 [.023]	.841 [.023]
		corr. (cfg)	.875	<<	.777 [.041]	.760 [.039]
	C (N=82)	Raw	.907	.830	.830 [.030]	.827 [.034]
		corr. (cfg)	.848	<<	.736 [.048]	.732 [.045]
Total (N=253)		Raw	.909	.833	.837 [.016]	.829 [.021]
		corr. (cfg)	.855	<<	.747 [.029]	.736 [.029]

Notes: The raw score is the number of hits, i.e. the number of correct words. The corrected score is computed with the cfg formula and in this case, Cronbach's alpha can only be estimated from the split-half reliability by means of the Spearman-Brown formula.

Both estimation procedures, half and part splits, led to comparable results. On average, the decrease in reliability goes from .907 for the raw score (/60 words) to .849 for the corrected (cfg) score. The very high reliability with raw scores is obviously artefactual because the number of correct words also measures the

bias itself. In the case of corrected scores, things are less clear because of the impossibility of dealing with possible omitted responses which were shown to be frequent with the same population in the case of the M.C. and also in our previous experiences with the True/False. It is therefore possible that the reliability of .849 remains overestimated.

In conclusion, controlling the response bias for the Yes/No format in order to be able to eliminate it from the raw data appears to be a central issue. Correction methods based on the discrete model deal with this problem only in a very indirect way.

5.2.2 Continuous models

Response bias is at the heart of the theory based on continuous models because their theoretical foundations clearly distinguish the sensitivity, which is the relevant variable, from the response bias, which has to be identified and ruled out. The theory underlying continuous modelling was initially formulated in military signal detection (SDT) where the task is to detect a signal. Most experimental evidence of the model has also been established in this field. This model is an alternative to the threshold theory. This threshold theory, which also rests on all-or-nothing assumptions, is the counterpart of the correction for guessing formula within detection theory. The advantages of this continuous model have been widely confirmed, first in the field of signal detection, later on in various other domains: experimental psychology, medicine, weather forecasts, etc. Because it is not yet widely used in the domain of language testing we provide a detailed description of the method as it can be applied in the case of the Yes/No vocabulary test (see Figure 5.5).

Contrary to the discrete approach, SDT posits a continuum ranging from " being sure of the presence of a signal/word " to " being sure of the absence of the signal/pseudoword ". In other words, the model can deal with the reality that word knowledge presents itself to participants as a continuum and it recognizes several degrees of certainty when a participant is put on the spot. The middle of the continuum corresponds to maximal doubt. This continuum represents a latent dimension for a particular participant as represented in Figure 5.5a. In the case of a confidence rating of the responses, the dimension as well as the distribution of the items on this dimension can be observed. By contrast, in the case of a Yes/No format, this dimension cannot be observed and has to be inferred by the model. The Yes/No task thus forces the participant to dichotomise the information which the model actually assumes to be continuous.

When an item is proposed to a participant, this item falls somewhere on the participant's latent confidence rating scale. If the participant's proficiency is not zero, a pseudoword will rather fall on the left-hand side (is not a word) of the continuum and a word will rather fall on the right-hand side

(is a word) of the continuum. The whole set of items will split up into two distributions, one for the pseudowords located on the left, one for the words located on the right. The shape of these distributions is crucial for the application of the model since different shapes will lead to differences in the testees' rank orders. The assumption of Gaussian curves with equal variance distribution is discussed extensively in Beeckmans et al (2001).

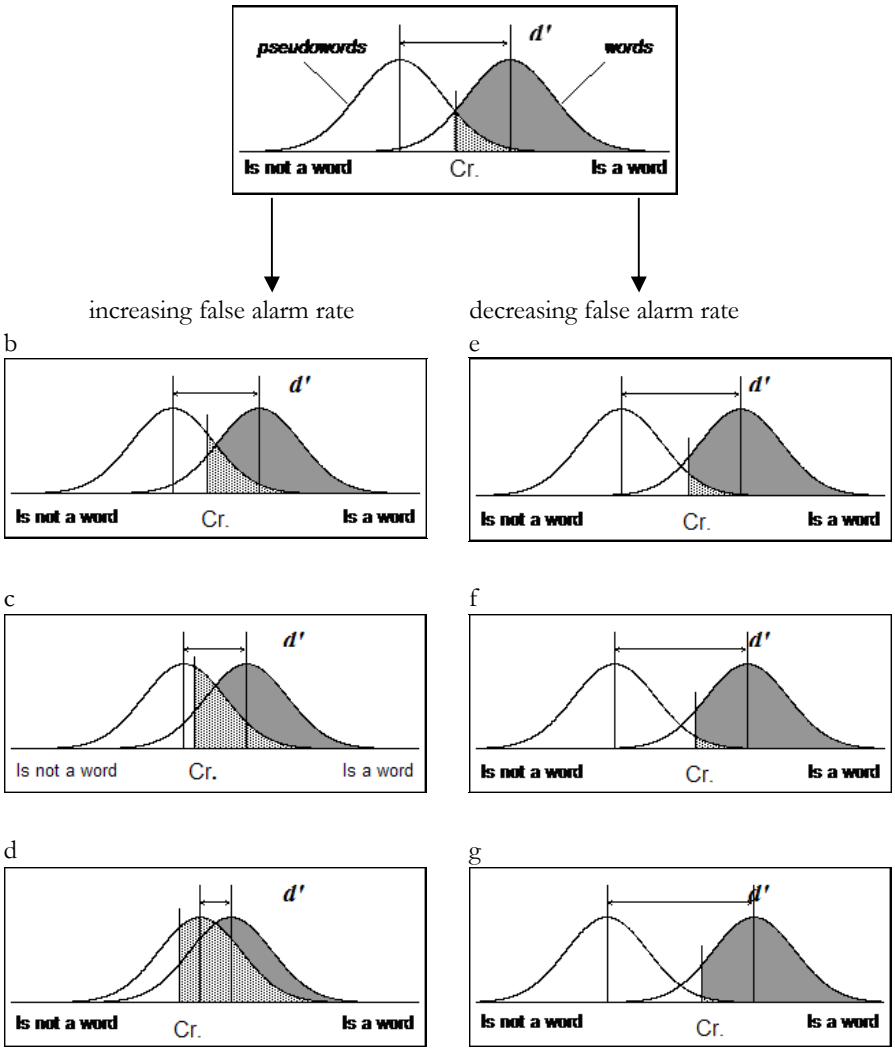


Figure 5.5: Results of several testees for the Yes/No Vocabulary Test as modelled by Signal Detection Theory, d' stands for the sensitivity and Cr. for the criterion.

The distance separating both distributions (the lack of overlapping) for one participant is called the sensitivity (d') for this participant. It is directly linked to the participant's competence. The core hypothesis is that the distributions and thus the competence that one wants to measure are strictly independent of the decision making process. The decision making process has to be clearly distinguished from the proficiency we aim to measure with the Yes/No test since decision making is influenced by the cognitive, psychological, social and cultural make-up of the participant rather than by his vocabulary ability. According to the model, the decision making process itself depends on the position of a criterion (Cr.) that is placed somewhere along the latent dimension. Every item falling at the left of Cr. will result in a "is not a word" decision, every item falling at the right of Cr. will result in a "is a word" decision. The more the criterion is situated versus the pole "is a word", the more the participant tends to answer "is not a word" when he/she hesitates and vice versa.

Figures 5.5 a to g illustrate the way in which the model formalises the distributions for different participants for whom the percentages of correctly detected words (hits in SDT terms) are the same, but who vary across their percentage of responding " is a word " in the case of pseudowords (false alarms in SDT terms). Participant a is special in the sense that his decision criterion Cr. lies at the intersection of the two distributions. In this particular case, the percentage of misses equals the percentage of false alarms. This means that the percentage of mistakes in both ways (not ticking a word versus ticking a pseudoword) is the same. The decision criterion used by this participant is neutral, one could say that there is no response bias in this case. Participants b, c and d have a growing percentage of false alarms. The effect of these increasing false alarms creates a shift of the decision criterion (participant d, for example, has a preference for the answer " is a word " in case of doubt) as well as - and this is the relevant information - a decline of sensitivity (participant d is almost unable to distinguish between words and pseudowords). On the other hand, a decline in the number of false alarms (participants e, f and g) is found with participants who are both more cautious and more competent. The comparison between the d' of two extreme participants d and g illustrates the importance of the false alarms factor in distinguishing students with a low or a high proficiency.

Undoubtedly, the most important merit of the SDT model for the Yes/No test is that it allows us to distinguish the criterion/response bias from the variable that is of real interest to us, namely the sensitivity/proficiency. It is worth repeating that the sensitivity/proficiency is independent of external circumstances (decision making process, overestimation, the task, etc.). These factors can, however, play an important role when it comes to placing the criterion. One could argue that instead of transforming the four raw values of the data matrix into one value of interest, the advantage of SDT is that it

transforms them into two interesting values: the d' and the Criterion, which is in itself worth investigating.

An issue which could question the continuous approach concerns the implicit assumption of what could be called the "homogeneity" of the task. The SDT description implies that the participant applies one and the same decision process when answering all the items. The only difference between one item and another is a supposedly unidimensional confidence scale on which the item is placed. It is possible, however, that when the participant believes the item to be a known word, he adopts one specific strategy which is quite different from when he believes the item not to be a known word. It should be emphasized that the participant's belief is under consideration which is different from presuming different behaviour with words versus pseudowords because this objective distinction is not available to the participant. If two different strategies come into play, none of the models, discrete or continuous, would be justified because the pseudowords would not constitute an adequate control for the response bias. In other words, the bias in the case of items believed to be known words could be different from the bias in case of items believed not to be known words so that inferring the bias with regard to words on the basis of the bias with regard to pseudowords would be spurious.

Huibregtse, Admiraal and Meara's index I_{SDT} (Huibregtse and Admiraal 1999, Huibregtse, Admiraal and Meara 2002) is based on Signal Detection Theory (SDT). When compared to other correction formulae, I_{SDT} is shown to be the only one to meet the three following criteria in a satisfactory way:
(1) taking into account different types of correct and incorrect responses;
(2) taking into account the correction for guessing
(3) neutralising the individual response style (Nunnally and Bernstein 1994)
It was shown that Meara's original ? m and Huibregtse,Admiraal and Meara's I_{SDT} are linked in a monotonic - but not linear - relation. When compared with the correction for guessing formula, however, the relation is no longer monotonic. This results in large differences in participants' rank orders.

Practically speaking, two measures can be regarded as corrected scores: either the d' itself, which takes into account the discriminability between words and pseudowords, or the percentage of correct responses among the words by computing what this percentage would have been if the criterion Cr. had been neutral. Both indices are linked in a monotonic and almost linear relation except for participants who perform very well. The advantage of the last option is that it can be interpreted as a percentage of known words, which is the primary aim of the test. It also provides an operational definition of what is meant by "knowing a word": the participant is considered to make the same number of mistakes in both directions (false alarms rate = 1 - hits rate). However, a last transformation is needed in order to obtain the convenient range of variation (0 to 100%) from the original range of variation (50 to

100%)[9]. The resulting corrected score corresponds to the number of Hits corrected for bias (Hcfb).

The I_{SDT} -formula is based on the geometrical properties of the ROC (Receiver Operating Characteristic) curves (Hodos 1970). However, as argued by Beeckmans et al. (2001), this measure is not truly distribution-free: it departs from the implicit assumption that the underlying distributions are equal variance logistic functions. Both curve families, normal and logistic, have very similar shapes so that the difference in using either Hcfb or I_{SDT} for correcting the raw score should turn out to be fairly small. Nevertheless, experimental evidence about the actual shape of the distributions is still needed and this important question may not be ruled out by deciding to use I_{SDT} instead of Hcfb.

5.2.3 Comparing the formulae based on discrete versus continuous modelling

A comparison between the effects on raw scores by applying cfg on the one hand and the two transformations based on the continuous model on the other hand leads to differences which may be very large, especially when the rate of hits is high. Figure 5.6 illustrates the effect of the three formulae for two different hit rates, .70 and .90.

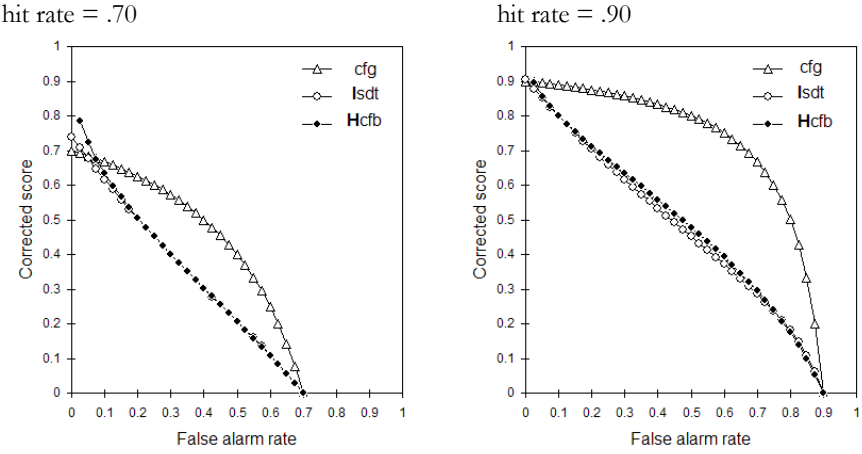


Figure 5.6: Differences in applying different correction formulae for two different hit rates.

[9] In case a participant's competence is zero, the two distributions overlap and the d' = 0. In this case the Cr. is situated at the mean of both distributions and the area on the right side of Cr. is still 50%.

When the false alarm rate increases, both corrections with the continuous model Hcfb and I_{SDT} reduce the score by a comparable amount, much more than does the discrete model. Such large differences are not only theoretically possible, they actually do occur in our data. They are the direct consequence of the differences between both models in taking the response bias into account.

Given the size of these differences, it was decided to examine the effect of the two continuous correction formulae on the test reliability. The method for estimating Cronbach's Alpha was exactly the same as that used for the cfg. The results are shown in Table 5.7 and 5.8, which reproduces the previous values with the raw score (/60 words) and the cfg correction to allow for a complete comparison.

Table 5.7: Effect of the three different corrections on the Yes/No test reliability when the score is limited to the 60 words for test version I.

Yes/No Test		Scores	Test reliability	Half-test reliability		Part-test rel. (60-40)
	Order	/60	Cronbach's Alpha	Spearman-Brown	Split-half k=50 mean [SD]	Split-part k=50 mean [SD]
I	A	Raw	.910	.835	.841 [.027]	.837 [.024]
	(N=78)	corr. (cfg)	.842	<<	.727 [.055]	.738 [.046]
		corr. (I_{SDT})	.780	<<	.639 [.065]	.652 [.054]
		corr. (Hcfb)	.748	<<	.597 [.065]	.738 [.046]
	B	Raw	.884	.792	.802 [.029]	.794 [.036]
	(N=79)	corr. (cfg)	.819	<<	.694 [.048]	.663 [.064]
		corr. (I_{SDT})	.751	<<	.601 [.063]	.568 [.060]
		corr. (Hcfb)	.739	<<	.586 [.067]	.663 [.064]
	C	Raw	.920	.852	.858 [.021]	.852 [.026]
	(N=78)	corr. (cfg)	.867	<<	.765 [.034]	.758 [.042]
		corr. (I_{SDT})	.802	<<	.669 [.036]	.666 [.049]
		corr. (Hcfb)	.766	<<	.621 [.037]	.758 [.042]
	Total	Raw	.906	.828	.835 [.017]	.859 [.021]
	(N=235)	corr. (cfg)	.843	<<	.728 [.033]	.721 [.039]
		corr. (I_{SDT})	.782	<<	.642 [.041]	.636 [.042]
		corr. (Hcfb)	.754	<<	.605 [.041]	.721 [.039]

Notes: The results are split up by order in which the items of the test were presented (A, B, C). The raw score is the number of hits, i.e. the number of correct words. With the three different corrected scores, Cronbach's alpha is estimated from the split-half reliability by means of the Spearman-Brown formula.

Clearly, the decrease in reliability is systematically larger (and also large in absolute terms) with the continuous model than with the cfg which shows intermediate values. On average, the reliabilities are .909 (raw scores), .855 (cfg), .817 (I_{SDT}) and .787 (Hcfb). It would be premature to conclude from these figures that the cfg does a better job at correcting the score than do both other correction formulae. It is possible that part of the bias remains in the cfg scores (remember the large differences shown in Figure 5.6 with large hit rates) which results in an overestimation of the test reliability. Also, the impossibility of identifying potential omitted responses may have come into play.

Table 5.8: Effect of the three different corrections on the Yes/No test reliability when the score is limited to the 60 words for test version II.

Yes/No Test		Scores	Test reliability	Half-test reliability		Part-test rel. (60-40)
	Order	/60	Cronbach's Alpha	Spearman-Brown	Split-half k=50 mean [SD]	Split-part k=50 mean [SD]
II	A	Raw	.909	.833	.841 [.022]	.827 [.024]
	(N=89)	corr. (cfg)	.845	<<	.732 [.040]	.722 [.044]
		corr. (I_{SDT})	.787	<<	.649 [.042]	.635 [.047]
		corr. (Hcfb)	.737	<<	.583 [.046]	.560 [.055]
	B	Raw	.914	.842	.849 [.023]	.841 [.023]
	(N=82)	corr. (cfg)	.875	<<	.777 [.041]	.760 [.039]
		corr. (I_{SDT})	.851	<<	.740 [.049]	.736 [.050]
		corr. (Hcfb)	.842	<<	.727 [.048]	.722 [.051]
	C	Raw	.907	.830	.830 [.030]	.827 [.034]
	(N=82)	corr. (cfg)	.848	<<	.736 [.048]	.732 [.045]
		corr. (I_{SDT})	.805	<<	.673 [.052]	.680 [.046]
		corr. (Hcfb)	.765	<<	.619 [.057]	.631 [.055]

Notes: The results are split up by order in which the items of the test were presented (A, B, C). The raw score is the number of hits, i.e. the number of correct words. With the three different corrected scores, Cronbach's alpha is estimated from the split-half reliability by means of the Spearman-Brown formula.

An important aspect to consider is the influence on the Yes/No test reliability of applying corrections to the raw scores, whatever the ultimate choice between both models, discrete or continuous. To deal with this issue, it is worthwhile to reconsider a major conceptual foundation in psychometric theory, namely the distinction between validity and reliability. It should be noted that reliability and validity are used in the narrow and precise sense they have within the terminology of classical test theory, so that confusions with the

wider concept of validation should be avoided. From a measurement perspective, reliability has, of course, an influence on validity. A weak reliability will never lead to a very high validity. However, the opposite is not true. A very high reliability does not imply a high validity because the reliability is primarily an indication of the accuracy of what is measured, independently of the extent to which the test is actually measuring what it is supposed to measure. This principle has to be kept in mind especially when biases can intervene in what is measured, like the response bias under discussion.

Reliability measures such as Cronbach's alpha are foremostly (if not solely) presented as a measure of internal consistency, i.e., all items measuring essentially the same thing. It is obvious that the higher the internal consistency among items, the more reliable the test will be, other things remaining equal. But it is also true that increasing the number of items will increase the test reliability. However, it would be rather unfair to state that Cronbach's alpha is a measure of the number of items. Reliability coefficients like Cronbach's alpha, KR20, KR21 should be considered primarily as a measure of the accuracy of what is measured by the test (putting the emphasis on whatever it measures) and that internal consistency is one of the factors which will increase the obtained accuracy. It then becomes clear that trying to increase the reliability by only improving the internal consistency may lead to a measure which will become more reliable but less valid.

We should also briefly address the question of test unidimensionality. Although there is no current consensus about this important question, most of the proposed methods are based on a factorial analysis approach. If some of the items measure essentially one thing and other items another thing which correlates poorly with the former, the usual methods will capture the two dimensions involved in the test. However, if each item measures two strictly independent things in a roughly similar amount, then any procedure based on factorial analysis will assess (unduly) test unidimensionality. Moreover, if the accuracy of one or the other measured factor is high, the resulting reliability might be high as well. The results obtained here suggest this is the case when the response bias remains involved in the measurement. The following examples illustrate what has to be considered questionable on the basis of these arguments. The reliability of .91 computed with KR21 obtained by Meara and Buxton (1987) might be overestimated if the response bias was not properly ruled out. In the study by Shillaw (1996), in which it is argued that the non-words are unnecessary in the Yes/No format and that they detract from the measurement quality (Read 2000), the reliabilities show the same trend as obtained here, i.e. a marked increase in reliability when considering the scores on words only. However, this rise of reliability has to be interpreted as the consequence of the accurate measurement of the bias and not as a guarantee for a more accurate measurement of vocabulary knowledge. One way of confirming the artefactual nature of the reliability obtained when the bias

remains embedded in the measurement would be to compare the Yes/No assessment with another vocabulary knowledge assessment in which the response bias can not intervene. This will be at the heart of Chapter 6.

5.2.4 Summarizing the methodological discussion

We can conclude that from a methodological point of view, a distinction between the correction for guessing formula used with classical M.C. tests (cfbg) and the apparently similar formula used with the Yes/No test (cfg) has been shown indispensable. In the first case (cfbg), the correction takes into account true random guessing which can be estimated on the basis of the test characteristics, thus independently of the participant's decision behaviour. As a consequence, a decrease in reliability after correction is observed which is exclusively dependent on the number of omitted responses. This factor can easily be ruled out or controlled. In the second case (cfg), the correction provides a control for the participant's response bias rather than for blind guessing. This bias is estimated on the basis of a discrete model in which the participant's decision rule appears to be far from realistic. The possibility of doubt as well as the possibility of being wrong when judging a pseudoword truthfully are occurrences the model cannot deal with. Moreover, the potential confusion between "is not a word" responses and omitted responses when using the classical format (Meara and Buxton 1987), renders the formula less effective. This raises doubts about the apparently more reliable results that were obtained when using the cfg correction.

When considering the continuous model, SDT provides a theoretical framework which seems appealing for estimating the response bias in order to eliminate it. Truthful mistakes about words or pseudowords and uncertainty in the response, are clearly taken into account by the continuous model. However, for the model to be useful, theoretical assumptions have to be posited. A theoretical model derived from SDT - varying in both the d' and the word variances - could be postulated in order to describe different proficiency levels including that of a native speaker. Such a sophisticated model has to be tested by direct measurements on a confidence rating scale.

5.3 Conclusion

In the first experiment with the Yes/No Vocabulary Test we found evidence for the presence of a substantial response bias contaminating the measurement of vocabulary knowledge. Apparently, several factors relating to the learner's profile (social, cultural, cognitive, etc) interact with the lexical knowledge that is meant to be measured by the test.

The analysis of the different correction formulae proposed in the literature pointed out that: (1) the discrete model uses correction schemes that are derived from classical tests and cause confusion about the difference

between guessing and response bias; (2) the continuous model is better equipped to deal with the response bias than the discrete model but its underlying theoretical assumptions have to be further validated.

A comparative investigation of the formulae applied to the empirical data revealed a severe drop in reliability when an attempt was made to extract the response bias from the raw score. There was also a lack of evidence motivating the choice of the most appropriate correction formula. This last drawback was even more problematic when considering the large differences in the participants' rank orders when applying one or another correction formula. However, it is clear that for both models (discrete or continuous), the reliability will be overestimated when a bias contaminates the score. Therefore, a motivated choice of formula for calculating a meaningful test score for the Yes/No Vocabulary Test cannot be based on a high reliability value but has to exceed the boundaries of the reliability criterion. Empirical evidence concerning the format's validity is required in order to solve this dilemma and this will be presented in Chapter 6.

Chapter 6

Concurrent validation

In Chapter 5, test reliability was shown to be a misleading guide for selecting the most adequate correction formula since contamination of the data by the precise measurement of the bias improved the overall test reliability. In this chapter we will attempt to solve the dilemma of choosing the most appropriate correction formula by collecting empirical evidence concerning the validity of the Yes/No Vocabulary Test. An experiment was set up to examine the influence of using different correction formulae on the correlation between Yes/No test results and the results of a translation task of the same words.

The first section of this chapter considers the validation process both in a general sense and for the particular case of the Yes/No Vocabulary Test. In Section 6.2 the aims of the present study are discussed in detail before moving on to the description of the experiment in which we tried to establish concurrent validity of the Yes/No Vocabulary Test by means of a translation task. Finally, in Section 6.3, the poor correlations that were obtained between the results of the Yes/No Test and the translation task, irrespective of the correction formula that was used, are discussed in relation with different variables relevant for the validation process of this test format. It is concluded (Section 6.4) that additional empirical evidence has to be collected in order to assess valid interpretation of Yes/No Vocabulary Test results.

6.1 The validation process

The current view of the validation process in educational measurement emphasizes the concept of construct validity as a unitary principle (Messick 1989) and this view has been largely adopted in L2 research (Bachman 2000). The construct validity of score interpretations is the cornerstone of the process upon which considerations of values, uses and consequences of tests are based. Changes over time in the concept of validity have been important and complex but one major trend has been the evolution from a narrow conception -which roughly corresponds to criterion-related validity[10] - towards a broader conception which takes into account an increasing number of considerations including the traditional types of validity (content, face, criterion-related, predictive, concurrent, construct) but also the outcome of the test on ethical and sociological grounds. The construct validation as a unified though multi-

[10] Guilford (1946), cited by Messick (1998), claimed that "in a very general sense, a test is valid for anything with which it correlates ".

faceted concept including or linked to all these aspects is now considered as a central issue in testing. It requires a definition of the construct to be measured and a thorough scrutiny of the different complementary facets of the validation evidence.

6.1.1 Recognizing the specificities of a particular testing situation

Within the aforementioned theoretical framework, it should be kept in mind that the relevance of the different facets may vary widely along with the specific testing situation. A methodological approach exemplifying such a validation process can be found in Chapelle's study of the C-Test format (1994). The C-Test format is constructed by deleting the second half of every other word in some sentences of several texts. Initially, the test was intended to provide a measure of overall language proficiency in reaction to test design that focused on testing isolated discrete points of language. It was also presented as an improvement of the well known Cloze test and the format gained a certain popularity, as had the Cloze test in its time. Both formats had been widely used in the 80s at our university and it appeared that most of the teachers considered the testee's task " interesting " even if, at the same time, the central question " but what does the Cloze/C-test actually measure? " was inevitably mentioned. In the literature, current work still continues to address the same question (e.g. Sasaki 2000). In Chapelle's study, the construct of vocabulary ability is defined and the justification of interpreting performance on a particular test format (C-Test) as indicative of this construct is addressed on the basis of a broad spectrum of various pieces of evidence that are not considered as alternatives but rather as complementing each other. Studies like that by Chapelle are therefore very useful when the nature of the assessed construct is at the heart of the debate. It is worthwhile to compare Chapelle's approach with what will be presented here. The differences in methodology between both studies are linked to differences in the overall context in which both test formats were created and not to diverging theoretical conceptions of the validation process.

6.1.2 The particular case of the Yes/No Vocabulary Test

Unlike integrative tests, the construct of the Yes/No vocabulary test is explicit and in some way integrated into the task so that questioning the construct validation in the literature was centred on the usefulness of this construct rather than on evidence that the Yes/No format actually did what it was claimed to do (measuring the size of the receptive vocabulary knowledge of the learner).

An attractive aspect of integrative tests is the expectation (although unfulfilled, Anckaert and Beeckmans 1990) that with some limited investment at the start (defining the genre of texts, providing ways of assessing the level of difficulty of the texts) anyone without any testing expertise could set up a satisfactory new test by following some general guidelines. Such an apparent

facility in constructing the test is a quality which also makes the Yes/No Test attractive. Moreover, the Yes/No Test also exhibits an apparent clarity in the intended measurement, namely receptive vocabulary knowledge. The consequence however is that the construct validity of the Yes/No has not been studied as extensively as is the case for integrative tests.

Our claim is that this apparent clarity in the intended measurement is misleading. The presence of words and pseudowords in the test material gives rise to questions about the construct validity. When results appear to be unsatisfactory (i.e. when the test data are characterized by a high false alarm rate), the first idea that comes to mind is the presence of two different constructs in the test results. The first would be the ability to recognise a word, which is seen as the valid construct. The other would be the ability to discard a pseudoword, which could be seen as distinct from the aimed construct. As a consequence, a simple count of correct responses is not a satisfactory alternative to the count of hits, for this approach would not be tenable when the false alarm rate differs significantly from zero. Because it is impossible to know a priori whether or not a participant actually knows the word, the objective distinction word versus pseudoword cannot have its corresponding counterpart within the data. In other words, the distinction between recognising a word versus discarding a pseudoword simply cannot be operationalized within a Yes/No Test.

6.1.3 Validity within the SDT-framework

When considering the validity issue within the theoretical framework of Signal Detection Theory, the information present in the data matrix is also split into two components, but these are different from the previous ones (recognising a word / rejecting a pseudoword). The d' becomes the part that taps the intended construct whereas the criterion position is clearly independent of the intended construct. Applying the SDT does not only result in a dramatic change of the corrected scores, it also assumes two distinct parts in the data :

- the two parts are not apparent in the raw data,
- the two parts do not correspond to responses to words versus pseudowords,
- the first part is related to the construct, the second to the response bias.

The situation is like a rotation after a factorial analysis in order to reach interpretable dimensions : before rotation, the two axes would be responses to words and responses to pseudowords, both of them being contaminated by the response bias. After rotation the two axes would be the valid part of the construct (d') and the extraneous variable (response bias). Because SDT is not usually used in language testing, we refer to Figure 6.1 (and to Figure 5.5 of Chapter 5), which presents an illustration of how presented stimuli may be positioned on a participant's internal confidence rating scale according to the

basic assumption of SDT. Figure 6.1 gives an example for 4 hypothetical stimuli.

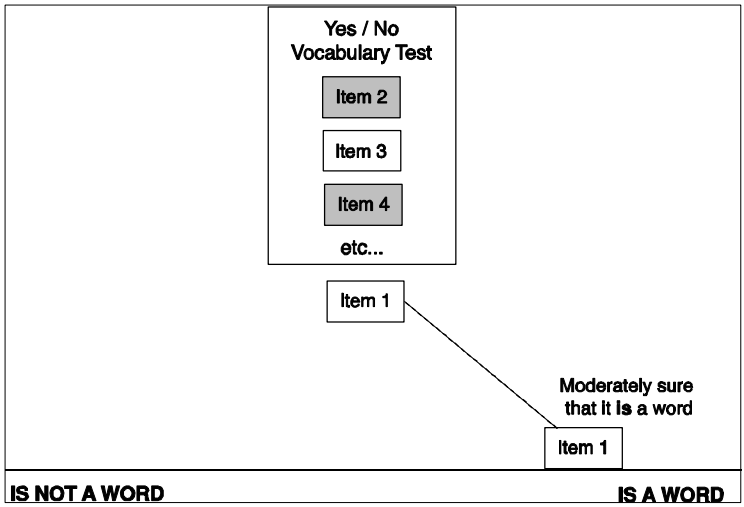
-Stimulus 1 is a word and the participant is reasonably certain that it is a word. As a result, this stimulus is positioned towards the "is a word" side of the scale.

-Stimulus 2 is a pseudoword and the participant is reasonably certain that it is not a word. In consequence, this stimulus is positioned towards the "is not a word" side of the scale, its position being symmetrical to that of stimulus 1.

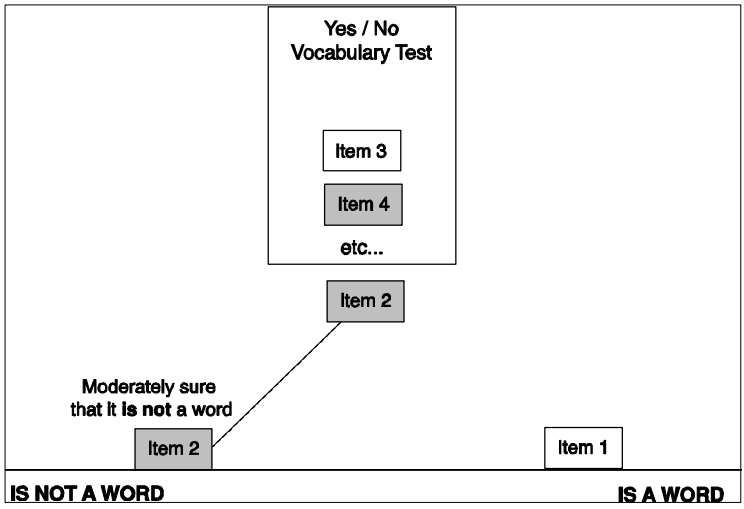
-Stimulus 3 is a word but the participant is in maximal doubt, so that it is placed at the middle of the scale.

-Stimulus 4 is a pseudoword but the participant is inclined to say that it is a word. This stimulus is therefore positioned towards the « is a word » side of the scale, but to a lesser extent than for stimulus 1, where the participant was more confident of his decision.

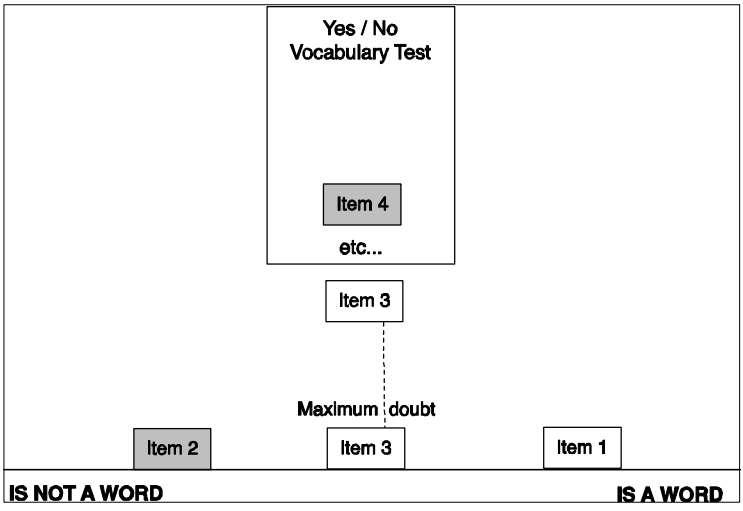
Stimulus 1



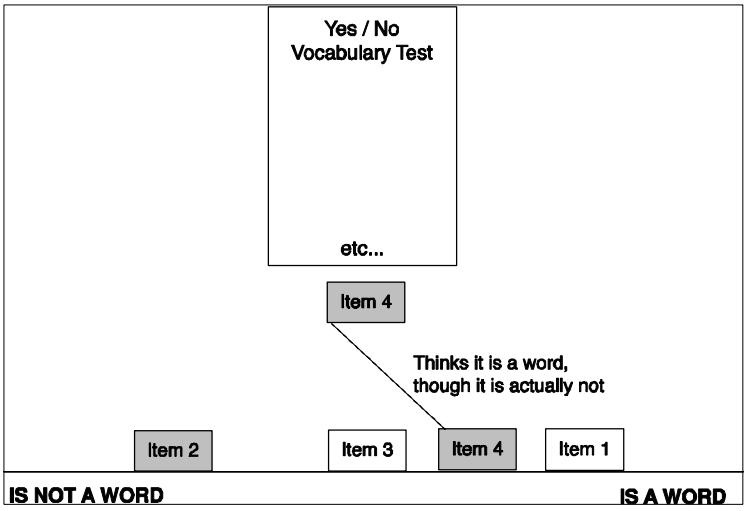
Stimulus 2



Stimulus 3



Stimulus 4



Result

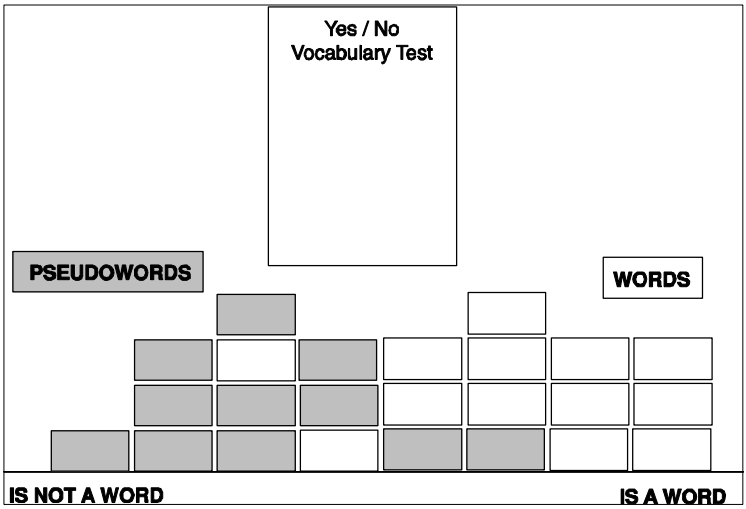


Figure 6.1: Hypothetical stimuli that are positioned according to the SDT-framework.

6.1.4 How to establish concurrent validity in the experiment

In the experiment we have opted for a validation in a very narrow sense. It does not address the relevance of assessing the learner's receptive vocabulary knowledge for definite purposes, nor does it make any assumption about the nature of vocabulary and vocabulary knowledge in a broad view of its role in language use. The aim of concurrent validation will be limited here to gathering evidence about the most suitable scoring method for a Yes/No Vocabulary Test (since reliability proved to be misleading for this purpose). As the high reliability obtained in the former experiment has been argued to be strongly dependent on the accurate measurement of the response bias and not of vocabulary knowledge, an external criterion appears to be necessary to provide insight into this issue. This external criterion should:

(1) be as similar as possible in content to the Yes/No test and
(2) be as independent as possible of the response bias

Similarity in content permits the avoidance of a lack of correlation between the two assessments, test and criterion, due to sampling problems. More technically, it prevents the attenuation effect due to the inferential error affecting both the criterion and the test. We therefore decided to ask the same participants to translate into L1 the words – not the pseudowords[11] - presented in the Yes/No Vocabulary Test. In doing this, we assumed that the task could not suffer the response bias present in the Yes/No format.

The central question of the experiment is : what scoring method will allow the construct "receptive vocabulary knowledge" as measured by the Yes/No Vocabulary Test design to be the same as when we ask the test taker to translate the same words into his or her L1? This is not to say that more general considerations about the validation process of Yes/No vocabulary results are less imperative than with other test designs, and some of them will be proposed for further research in Section 6.2.4.

6.2 Experiment 2 : Validating the Yes/No Vocabulary Test by means of a translation task

6.2.1 Aim

The central aim of this experiment was to gather additional empirical evidence about concurrent validity which could help in selecting the most appropriate correction method for the Yes/No Vocabulary Test. In view of the bias we encountered in former data (Experiment 1), it seemed very important to compare the Yes/No assessment with another assessment of receptive vocabulary knowledge in which the response bias cannot intervene. Therefore

[11] My colleagues and myself shared the opinion that, although interesting, it would not be very pedagogical to present the learners with a translation task that included pseudowords.

we opted for a translation task targeting the same words as the Yes/No Vocabulary Test. At the same time the experiment served to verify the main trends obtained with the previous application of a Yes/No Vocabulary Test - which was a paper-and-pencil test. These main trends were

(1) an unusual high rate of false alarms which was not limited to weak students;

(2) a negative correlation between the performance on words versus the performance on pseudowords;

(3) a contamination of the test results by the response bias.

Of these, the first trend probably had the others as concomitants.

6.2.2 Method

Participants

The participants were first, second and third level French-speaking university students of Economics and Business Administration. A total of 161 participants took the test within the framework of the CALL-facilities (Computer Assisted Language Learning) which is part of the Dutch language course curriculum. They were informed that the test scores served to provide them with an indication of their knowledge of the Dutch core vocabulary.

Specificities of the Yes/No Vocabulary Test

In order to improve the properties of the test and to maximize the chances for good concurrent validity, two alterations were made with reference to Experiment 1:

(1) The same Yes/No Test was used as in Experiment 1 but it was administered on computer. We opted for a computerised version of the Yes/No Vocabulary Test on the assumption that certain constructional aspects that seemed problematic in Experiment 1 (which was a paper-and-pencil experiment) might be better dealt with in a computer application because of the more controlled environment it provides. For instance, the computerised form of the test allowed to record when responses were omitted, which made it possible to distinguish between omitted responses on the one hand and correct rejections or misses on the other hand. This distinction is of great importance when it comes to calculating a test score, as was shown in Chapter 5. The computer application might also underline the forced decision character of the task. It was hoped that the more controlled environment would result in a less biased response behaviour and consequently a decrease in the false alarm rate.

(2) Although the instruction was the same as in Experiment 1 - *Indiquez à l'aide d'une croix les mots que vous connaissez. Certains mots repris dans la liste n'existent pas en néerlandais!* (Tick the words you know. Certain words figuring in the list do not exist in Dutch) – special care was taken to ensure if it was properly interpreted. The students were asked if they understood what was meant by this instruction. To be certain, it was explained that the task was not to decide if

they had ever encountered the respective words, but to determine if they knew what these words meant.

Since the number of participants participating in this experiment was smaller, we decided to use only one of both parallel versions of the test (Version I). As described in Chapter 5, it was composed of 60 words and 40 pseudowords (see Appendix 3). In view of the aim of seeking confirmation or generalisation of the results we obtained previously, we will often refer to the results of this paper-and-pencil experiment and we will compare the analyses and findings to the current one.

The Translation task

The second part of the experiment served as a control measure both to verify the truthfulness of the learners' responses to the Yes/No Vocabulary Test and to provide evidence of concurrent validity of the format. This control measure consisted of asking the same participants to give the French translation of the existing words of the Yes/No Vocabulary Test.

We opted for a translation because we assumed that asking the participants to provide mother-tongue equivalents of target language words was the most univocal way of verifying recognition. Nation (2001) also holds the opinion that the use of the first language to convey and test word meaning is very efficient and that explaining the meaning of target words through translation is much easier for language learners than through multiple choice items or providing definitions. With reference to the latter option, he argues that "the difficulties caused by no exact correspondence between meanings in L1 and L2 are probably less than the difficulties caused by the lack of correspondence between L2 definitions and the meaning they are trying to convey" (Nation 2001: 351).

The translation task was also administered by computer. The sixty existing words of the Yes/No Vocabulary Test were presented on the screen one by one. The participants were informed that all words existed in Dutch. After entering their translation, they had to confirm it by clicking a button which in turn presented them with the next item. This translation test was administered one week after the Yes/No test. Participants were not informed beforehand about this second test. In fact, special care was taken to ensure their ignorance about it in order to avoid a possible taint on their performance on the Yes/No Vocabulary Test.

6.2.3 Results

The Yes/No Vocabulary Test

In Table 6.1 the results on the Yes/No Test are presented for the raw and corrected scores which were discussed in Chapter 5. When compared to the results of the same material in Experiment 1 (see Chapter 5), the mean was

higher and the standard deviation lower. Both differences were expected since the participants in Experiment 2 were more advanced in their L2 curriculum.

Table 6.1: Results on the Yes/No Vocabulary Test in Experiment 1 and 2 (Test Version I) with the different methods of scoring

	Score /100				Correl. w/pw	Score /60			
		Mean	SD	%			Mean	SD	%
Exp. 1 (N=235)	Raw	72.58	8.95	72.6	-.373***	Raw	39.82	9.46	66.4
	Corr. (cfbg)	43.41	17.32	43.4		Corr. (cfg)	34.65	10.79	57.8
						Corr. (I_{SDT})	28.81	9.39	48.0
						Corr. (Hcfb)	29.77	9.95	49.6
Exp. 2 (N=161)	Raw	79.77	5.82	79.8	-.364***	Raw	48.45	5.00	80.8
	Corr. (cfbg)	57.54	11.65	57.5		Corr. (cfg)	44.67	6.39	74.5
						Corr. (I_{SDT})	34.64	7.36	57.7
						Corr. (Hcfb)	35.32	7.60	58.9

Notes: The raw score is either the number of correct responses to the words and pseudowords (/100), or the number of hits (/60), i.e. the number of correct responses to the words. Corrected scores are based either on the all-or-nothing model (cfbg and cfg) or on the continuous model (I_{SDT} and Hcfb). Significant correlations are marked with * ($p<.05$), ** ($p<.01$) and *** ($p<.001$).

In Experiment 1 many participants displayed a high rate of false alarms in their response behaviour, which cast serious doubts on the confidence which could be placed in the hit responses. Furthermore, the high false alarm rate was not restricted to weaker participants. In Experiment 2, the participants formed a much more homogeneous group in terms of Dutch language skills and on the whole their course results indicated that they were significantly better at Dutch than the participants of Experiment 1. However, as shown in Figure 6.2, the results showed that the false alarm rate was high. In fact, it had not diminished with this more proficient population. The ability to identify real words logically increased but the ability to reject pseudowords did not. In fact, it even decreased (20,5% false alarms in Experiment 1 versus 24,2% false alarms in Experiment 2). The hypothesis of the presence of an important response bias (and the problems this caused for establishing a valid test score) was therefore reinforced. Again the results of the experiment contradicted the assertion that only weak students would display an overwillingness to claim knowledge of the pseudowords. Furthermore, it appeared that neither the computerized form of the test nor the explicit reinforcement of the task had led to a decrease in the false alarm rate.

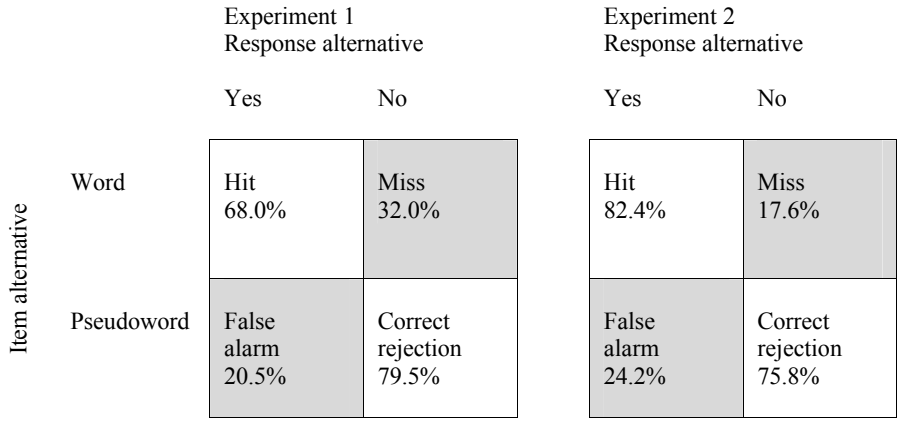


Figure 6.2: The item-response matrix of the Yes/No Vocabulary Test in Experiment 1 and 2. Percentages are calculated within each item alternative.

As a result of the computer application of the test the distinction between omitted responses and correct rejections or misses was controlled (which was not the case in the classical paper-and-pencil test of Experiment 1). We had hoped to encounter fewer omitted responses and this was clearly the case: only 134 responses were missing (n=16,100), which is less than 1%. These few omitted responses appeared to be scattered throughout the student population. Therefore, the missing responses were considered incorrect in further statistical analyses. Having the assurance that the problem of confusion between omitted response and correct rejection was actually discarded, it became possible to verify one of the main consequences of the large number of false alarms in Experiment 1: the negative correlation between the performances on words versus the performance on pseudowords. As could be expected from the high false alarm rate, this result was confirmed in Experiment 2: a similar correlation between performances on words and performances on pseudowords was obtained (Experiment 2, $r = -.364$, Experiment 1 $r = -.373$, see Table 6.1). Again, this weak but significant negative correlation between the ability to identify words and the ability to reject pseudowords is evidence for the presence of a substantial response bias contaminating the measurement of the vocabulary knowledge of the participants.

Another issue that needed to be verified in view of the high false alarm rate and the aforementioned negative correlation concerned the reliability of the test when applying the different correction formulae (discussed in Chapter 5). The reliabilities were established in the same way as in Experiment 1, i.e., the calculation of Cronbach's alpha in the case of the raw scores and its estimation

when scores were corrected (Beeckmans et al 2001). As can be seen from Table 6.2, the Alpha calculated from the raw scores in Experiment 2 was lower than the Alpha for Experiment 1 (.787 versus .906). This can be accounted for by the difference already mentioned in the range of proficiency of both student groups. The population in Experiment 1 contained both very weak (almost true beginners) and very proficient participants which is not the case in Experiment 2[12].

Table 6.2: Effect of the three different corrections on the Yes/No Vocabulary Test reliability when the score is limited to the 60 words.

Yes/No Version I	Scores /60	Test rel. Cr. Alpha	Half-test rel. Sp.-Brown	Split-half k=50 Mean	SD	Part-test rel. Split-part k=50 Mean	SD
Exp. 1	Raw	.906	.828	.835	.017	.859	.021
(N=235)	corr. (cfg)	.843	<<	.728	.033	.721	.039
	corr. (I_{SDT})	.782	<<	.642	.041	.636	.042
	corr.(Hcfb)	.754	<<	.605	.041	.721	.039
Exp. 2	Raw	.787	.649	.654	.041	.651	.036
(N=161)	corr. (cfg)	.711	<<	.551	.060	.537	.057
	corr. (I_{SDT})	.711	<<	.551	.053	.535	.049
	corr.(Hcfb)	.689	<<	.526	.053	.510	.046

Notes: The raw score is the number of hits, i.e. the number of correct responses to words. With the three different corrected scores, Cronbach's Alpha is estimated from the split-half reliability by means of the Spearman-Brown formula.

The observed difference in score variance between groups induced a difference in reliabilities. Although this raw score reliability was not as high as in Experiment 1, the several corrections, once again, reduced it further as was the case in Experiment 1. However, the effect of using the cfg formula was not the same. In Experiment 1 the use of the cfg formula led to reliability values which were systematically intermediate between those obtained with the raw scores and those obtained with the other corrected scores. Here, by contrast, the value appeared to be closer to those obtained with the other two correction formulae. The lack of control of the omitted response category in the case of Experiment 1 could account for these different results.

In short, we can conclude that as far as the results of the Yes/No Test are concerned, all main results of Experiment 1 about the internal qualities of the Yes/No Vocabulary Test were confirmed:
-a high rate of false alarms;

[12] Many weak students did not pass their exams and were consequently not admitted to the next course degree. Near-native students are exempted from the first level courses.

-a negative correlation between performances on words versus performances on pseudowords;
-a decrease in reliability when using the different correction formulae and this time of similar size for both models (discrete and continuous);
-a lack of experimental evidence in favour of one particular correction formula.

The Translation task
Since the words were presented to the participants in isolation, a large variation in the translations per word was expected (synonyms, far-fetched or particular uses of words, mistakes in grammatical category, mistakes in number, spelling errors, etc). Therefore a fairly straightforward taxonomy was made which resulted in the following categories:
1 : correct translation
2 : correct translation but wrongly spelled or typed
3: mistakes due to grammatical category (for example: the Dutch noun "*godsdienst*" [religion] gets translated into the French adjective "*religieux*" [religious]
4 : undoubtedly incorrect translation or no response

Two different correction schemes were constructed on the basis of this taxonomy. The Automatic Correction Scheme only accepts category 1 as correct. The Lenient Correction Scheme accepts all categories except for category 4. Both correction schemes demand a kind of "human-assisted" computer scoring for even the Automatic Correction Scheme is difficult to programme in advance since you have to consider all possible correct uses of the words. The difference between these two schemes was 5% of the total number of translated words (520 of 9660)[13]. Table 6.3 summarises the results for the translation with both correction schemes.

Table 6.3: Descriptive statistics of the results on the Translation test.

	Automatic correction			Lenient correction		
N=161	Mean (/60)	SD	Reliability	Mean (/60)	SD	Reliability
	33.35	6.23	.808	36.58	5.68	.798

The differences between both schemes were small. The test reliability calculated with Cronbach's alpha was not very high when considering both the number of items (60) and the fact that a production test should be more reliable than a corresponding multiple-choice test. The test sample could account for this moderate reliability. The procedure of selecting the words randomly in a certain frequency range is probably not the most efficient way to obtain well-

[13] It is worth noting that 495 out of the 520 could be attributed to errors of spelling and typing (category 2).

performing items in a translation task, but this was not what we aimed at with the present test. The only aim of the translation experiment was to assess whether the results on the Yes/No Vocabulary Test - when corrected in an adequate way - gave accurate information about the participant's knowledge of the actual 60 words, and not of vocabulary in general.

Usually, concurrent validation is based on the correlation between two measures differing in their formats and their content. The lack of reliability which is linked to inferential factors is then of major importance. In our case, things were different because we had the unusual opportunity of using formats with the same content, thus avoiding the inferential problems. This also implies that, in our case, the correlation should be very high in order to obtain good evidence of concurrent validity since the negative effect of the lack of reliability due to the inference factor is ruled out. Table 6.4 shows the obtained correlations between the Yes/No Vocabulary Test and the translation results, for the several correction schemes and formulae under consideration. The overall correlations were weak, which undermines the concurrent validity that was hoped for and none of the correction formulae appear to perform better than the others.

*Table 6.4: Correlation between the Yes/No Vocabulary Test and the translation test. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

Yes/No Test score	Translation of the 60 words used in the Yes/No Test	
(N=161)	Automatic correction	Lenient correction
Raw	r= .389***	r= .406***
Corr. (cfg)	r = .537***	r = .596***
Corr. (I_{SDT})	r = .432***	r = .452***
Corr. (Hcfb)	r = .464***	r = .491***

Using the cfg formula led to a slightly stronger correlation than the corrections based on the continuous model. However, this advantage was counterbalanced by the fact that the cfg did worse than the other corrections in absolute terms. Figure 6.3 shows a comparison between cfg and Hcfb. Clearly, the points are less spread with the cfg while they are on average closer to the diagonal with the Hcfb, which means that the cfg formula succeeds less in correcting the participants' overestimation of their vocabulary knowledge.

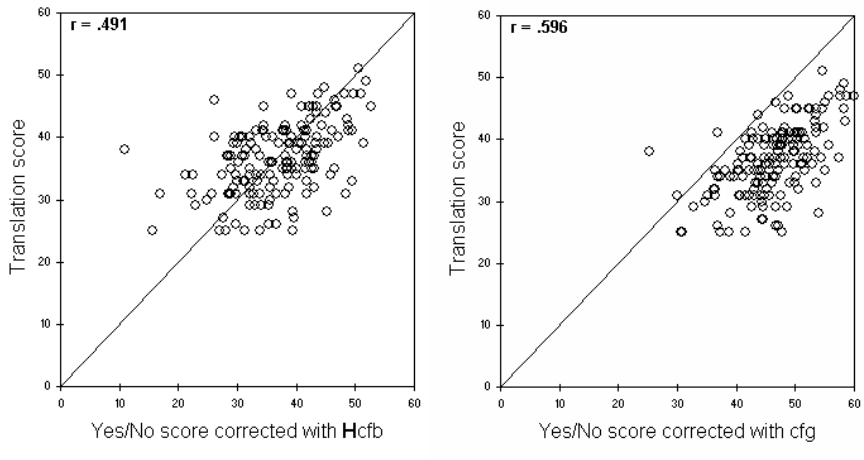


Figure 6.3: Correlations between the number of correct translations (lenient system) and the score on the Yes/No Vocabulary Test corrected with the continuous model (Hcfb) and with the discrete model (cfg).

In view of the high false alarm rate in the Yes/No Test and the doubt this casts on the confidence that is to be attributed to the "Hits" in this experiment, an item analysis was carried out in which the responses to words in the Yes/No Test were matched with the translations that were given for these items. Table 6.5 shows the four possible patterns that result from this match[14]. Of the two possible responses to words in the Yes/No Test ("Yes" and "No"), the participants could have produced either a correct or an incorrect translation of the item in the translation task[15].

Table 6.5: The four possible patterns that result from the match between the responses to words in the Yes/No Vocabulary Test and the Translation task.

Responses to words in the Yes/No Test		Translated words	
Response	%	Correct	Incorrect
Yes	82.4	69.42 %	30.58 %
No	16.9	25.44 %	74.56 %

The number of words that evoked a "Yes" response in the Yes/No Test and that were translated incorrectly in the translation task amounted to 30.58%. This means that within the 82,4% responses "Yes, I know the meaning of this

14 For this match we have used the Lenient Correction Scheme.

15To be exhaustive, we also need to mention the categories "omitted response/correct translation" (0.28%) and the category "omitted response/incorrect translation (0.43%).

word" to word-items, almost one out of three appeared to be the result of defective self-assessment. These results coincided with the high false alarm rate we encountered in the data of the experiment and reinforced the conclusion that the participants were overestimating their vocabulary knowledge.

Another pattern that seemed of importance concerned the percentage of words that elicited a "No"-response and yet resulted in a correct translation. This pattern was expected to be non-existent or in any case negligible. Still, the data show that one out of four word-items that elicited a "No" response, fell into this pattern. Since it appears odd to reject a word-item and translate it correctly afterwards, it was decided to look more closely at the items that induced this kind of response pattern. In Table 6.6 the most attested word-items are listed according to the number of times they elicited the "No" response + correct translation pattern.

Table 6.6: Items that exhibit the "No"-response + correct translation pattern (for more than 10% of the participants), in decreasing order. The values are to be considered on a total number of 161 instances.

Word	"No"- response + correct translation	
	/161	%
sok	52	32
opereren	42	26
moskee	31	19
militair	30	18
humor	28	17
fractie	26	16
verbod	23	14

Apart from the words "sok" and "verbod" for which we cannot come up with any satisfactory explanation as to why they were rejected in the Yes/No Test and yet correctly translated, these items were all cognates. The participants appeared to display uncertain response behaviour towards these cognates. When they encountered them in the Yes/No Test, they rejected them, probably because they suspected that they were being tricked into accepting words that have a strong resemblance with words in their L1. After all, they had been warned in the instruction that the test contained words that do not exist in Dutch. When these words were presented to them in the translation task at a later stage, they realized that these items were legitimate Dutch words and they tried out the French equivalent of the word, which resulted in correct L1 translations in the aforementioned cases. It should be noted that not all cognates of the test induce this kind of response behaviour. Words as "directeur" (headmaster), "chauffeur" (driver), "discriminatie" (discrimination) are also cognates and were rightfully recognized as Dutch words by the participants. Although it has been assumed (e.g. Meara et al 1994) that cognates foremostly lead to "Yes"-responses and might cause an overestimation of the

testee's word knowledge, these data seem to indicate that the cognates can work in both directions. To take the item analysis one step further, Table 6.7 presents the four possible response patterns for word-items when all cognates in the test (14 in this case) were excluded[16].

Table 6.7: The four possible patterns that result from the match between the response in the Yes/No Vocabulary Test and the Translation test, with the exclusion of cognates.

Responses to words in the Yes/No Test		Translated words	
Response	%	Correct	Incorrect
Yes	80.6	61.29 %	38.71 %
No	18.7	14.44 %	85.56 %

It is observed that the "No"-response category is quite similar (16.9% versus 18.7%), but the number of correct translations decreased (from 25.44% to 14.44%). When both tables were compared, the most striking result was that the "Yes"-response + incorrect translation pattern increased when the cognates were excluded. The proportion "Yes"-response + incorrect translation on the total number of "Yes"-responses to words, which was 30.58% for Table 6.5 ran up to 38.71% for Table 6.7. This provided strong evidence for the claim that the cognates in the test were not responsible for the participants' tendency to overestimate their word knowledge. It also supported the tentative conclusion of Chapter 5 that the response bias in the test is not to be attributed to the test content or to construct-relevant variables. It is a bias that has to be eliminated from the data since it clearly works independently of linguistic skills.

An analysis of the data from a participant perspective revealed that 122 out of 161 participants (76%) demonstrated the "Yes"-response + incorrect translation pattern for one third of the word-items. Figure 6.4 illustrates that this tendency to overestimate word knowledge was not at all restricted to a minority of the participant population.

When the cognates were excluded from the sample, 147 out of 161 participants (91%) displayed the "Yes"-response + incorrect translation pattern for one third of the words. Again, it appeared that the cognates were not responsible for the response bias. The overestimation that was displayed by the participants ran through the data irrespective of test content and it was not restricted to a small group of individuals. It rendered the test responses unreliable.

[16] The category "omitted response/correct translation" amounts to 0.15% and the category "omitted response/incorrect translation to 0.54%.

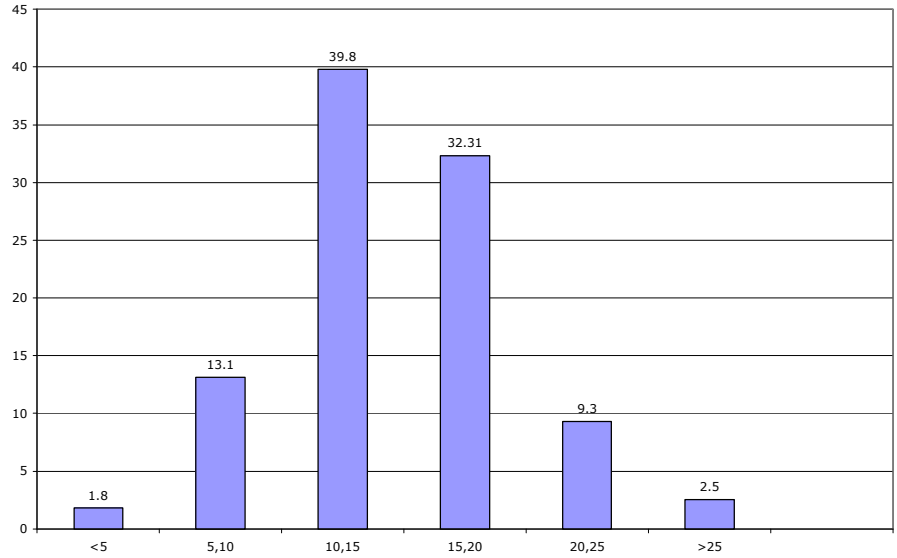


Figure 6.4: Histogram of the number of times the participants displayed the "Yes" response + incorrect translation pattern. On the horizontal axe, the number of instances that "Yes"-responses resulted in incorrect translations, on the vertical axe, the number of participants in percentages (N=161).

6.2.4 Discussion

The results of Experiment 1 were fully confirmed by this study. Sadly, the problems encountered because of the difficulty to handle the response bias were weakened neither by using a more controlled environment nor by a greater emphasis on a clear task description.

With reference to our experimental results, two questions remain to be examined. The first centers around the poor concurrent validity. Is the unconvincing correlation solely the consequence of the lack of reliability of the Yes/No measure or is it also imputable to a more general validity problem? One way of addressing this question would be to try to improve the qualities of the test content, words and pseudowords. Shillaw (1996) has proposed to investigate the qualities of each item on measurement quality consideration in order to obtain better test materials. This option is certainly of great interest in situations where preliminary studies are possible. But, one of the most appealing features of the Yes/No test remains the opportunity to set up a new version by following a few straightforward rules, with the assurance of getting a reliable and valid measurement of a participant's receptive vocabulary without any preliminary analysis (Meara 1990, 1996). It has been mentioned in the Yes/No literature that pseudowords are not linguistically neutral (Read 2000)

and that further investigation on the choice of words and pseudowords is needed. Learners may react differently to words and pseudowords as a consequence of their linguistic background and not as a consequence of their differences in L2 proficiency. However, the item analysis in this experiment has shown that words are not linguistically neutral either, the cognate effect which is thought to lead to an overestimation of vocabulary knowledge as a result of the lexical resemblance between the participants' L1 and the target language, appears to work in both directions.

The second question concerns the generalisation of our findings. Are certain specificities like test instructions or sociocultural factors responsible for some aspects of the present results? The test response behaviour, and more particularly the reactions to pseudowords, may not only be influenced by language background but also by socio-cultural specificities. As we already pointed out in Section 3.6.2 of Chapter 3, the task the learner faces in the Yes/No format lies in-between a traditional language test and self-assessment. This ambiguity is partly responsible for the difficulty in interpreting the test scores. It is evident that sociocultural factors might play a role when self-assessment is concerned and the nature of this influence needs to be investigated (Oscarson 1997). We cannot rule out the possibility that the high false alarm rate may be partly accredited to sociocultural variables. We find evidence for this sociocultural claim in the attitude we observe in our learners throughout the year. Many of them seem to focus entirely on passing their exams rather than on acquiring the competences the particular course is aimed at. In particular, our British colleagues (who taught the same student population) expressed their astonishment at this high level of extrinsic motivation displayed by the learners, and contrasted it with learning attitudes they encounter in other countries.

With reference to these socio-cultural variables, test instructions and their implications on learners' choices have not been sufficiently investigated. In Experiment 1 we opted for an instruction in which we refrained from any in-depth explanation of what we define by " knowing a word ". In the instruction of Experiment 2, however, we insisted on the difference between " judging a word to be Dutch ", " having encountered the word somewhere in the past " and " knowing what the word means ". It was made clear that this third interpretation was the one that should be applied when taking the test. After comparing the results of Experiment 1 and Experiment 2 it could only be concluded that the strong insistence and exemplification of the instruction did not reduce the false alarm rate and did not make the test more reliable or valid. Since it is not clear whether alternative test instructions cause the test to address a different level of the learner's vocabulary knowledge, a new experiment was set up in order to investigate this point further. The results of this experiment will be reported in Chapter 7.

6.4 Conclusion

The data of this experiment have confirmed the main trends obtained with the previous paper-and-pencil application of the Yes/No Vocabulary Test. At the start of the first study (Experiment 1), we claimed that the problem of scoring had to be resolved first in order to address other questions that the format raises. At the end of this second study (Experiment 2), it is clear that the evidence we presented, although illuminating, fails to identify one particular scoring method as the most valid. The scoring problem continues to weigh heavily on the workability of the Yes/No format but several elements of test design together with certain local characteristics of our data (the socio-linguistic specificities of our learners, etc.) prevent us from obtaining conclusive results. Since the scoring dilemma could not be settled, we will continue to present the different corrected scores (cfg, I_{SDT} and Hcfb) in the next chapters.

Improving the test content, the test format or the test task in different ways (i.e. redefining the selection criteria for words, adapting the word frequencies to the proficiency of the population, modifying the test instruction, etc.) could increase the differences among different scoring methods, which in turn should provide more confident statistical properties of the items used and so on. In the subsequent chapters we will therefore reconsider the problem of scoring in parallel with the other aspects we have pointed out, in accordance with the current view of the validation process. In Chapter 7, two experiments are presented in which we have attempted to reduce the response bias through alternative test instructions (Section 7.2) and exercising control on the response behaviour by means of a carefully designed computer application (Section 7.4).

Chapter 7

Reducing the response bias

In Chapter 5 and 6 it was shown that the data gathered with the Yes/No Vocabulary Test suffered from a substantial response bias. The reliabilities yielded by the experiments were misleading because they reflected a measure of the response bias which tainted the estimates of the participants' vocabulary size. It was concluded that the response bias needed to be corrected (by transforming raw scores into corrected scores) or eliminated if the Yes/No test was to be used as a reliable and valid placement test.

In the present chapter two experiments will be described in which we attempted to rule out the possibility of a response bias occurring in our data. In Sections 7.1 and 7.2, the relationship between different test instructions and their influence on the participants' responses is investigated through an experimental design in which a rather vague instruction is contrasted with a rigorous instruction while using identical test content. The experiment described in Sections 7.3 and 7.4 centered around the use of different computer format designs of the Yes/No test and how these influenced the participants' response behaviour.

7.1 Influence of the instruction on the response behaviour

7.1.1 The instruction as part of the test characteristics

The way test takers perform on language tests is affected to an important extent by the characteristics of the tests themselves (Bachman and Palmer 1996). Therefore, test characteristics largely determine how performance on a given language test can be related to language use in non-test situations. In other words, the specificities of a particular test task will determine the validity of inferences made. To provide language testers with a basis for language test development and use, Bachman and Palmer (1996) distinguish five aspects of test characteristics: setting, rubric, input, expected response and the relationship between input and response. (Bachman and Palmer 1996: 48). In this section we will consider the characteristics of the test rubric because they involve the test instructions. Bachman and Palmer define rubric as "(…) those characteristics of the test that provide the structure for particular test tasks and that indicate how test takers are to proceed in accomplishing the tasks" (1996:50).

Because of the need to make inferences on the basis of test performance, the test taker's approach to the testing procedure, more

specifically his intention to fill the test requirements to the best of his abilities, is crucial. The instructions are the first part of the test that test takers encounter. Not only is their primary purpose to ensure that the test takers understand the exact nature of the testing procedure and how they are to respond to the test task, they also bear much of the responsibility for setting the test takers' expectations and appropriately motivating them to take the test conscientiously.

Instructions should not be considered as part of the test itself since they do not represent the items to which responses are solicited. It follows that they may be presented in the test takers' native language as well as in the target language in cases where test takers come from many different first language backgrounds. According to Bachman and Palmer (1996: 190), efficient and effective test instructions have three qualities:

-they are simple enough for test takers to understand,
-they are short enough not to take up too much of the test administration time,
-they are sufficiently detailed for test takers to know exactly what they are expected to do.

7.1.2 Overview of the instructions used in the Yes/No Vocabulary Test

An investigation of Yes/No test administration shows that although different instructions have been used so far, the relationship between the use of a particular instruction and the response bias problem has been neglected. When Meara first developed the Yes/No Vocabulary Test, the following instruction was used:

(A) *Tick the words you know the meaning of, e.g. milk: V* (Meara and Buxton 1987).

In the computerized Eurocentres Test, this is changed into:

(B) *Look through the French words listed below. Cross out words that you do not know well enough to say what they mean. Keep a record of how long it takes you to do the test*. (Meara and Jones 1988: 81)

We can note that the instruction contains no mention of the presence of pseudowords in the test. In an article about the Eurocentres Vocabulary Size Tests, Meara writes "that testees using this test are very cautious in the number of non-words they accept anyway, and this tendency can be enhanced by careful wording of the instructions." (Meara 1990: 110)

The EFL Vocabulary Test (1992), which is an improved version of the Eurocentres Vocabulary Size Test (Meara, personal communication), highlights the forced decision character of the test task:

(C) *Read through the list of words carefully. For each word: if you know what it means, write Y (for Yes) in the box, if you don't know what it means, or if you aren't sure, write N (for No) in the box.*

Presumably this particular instruction was formulated in order to make sure that omitted responses could not be confused with "No"-responses in the analysis of the test responses. Shillaw (1996) retained this instruction when he started using the EFL Vocabulary Test in Japan. In an article about the dimensions of lexical competence in which the growing experience with the Yes/No Vocabulary Test is described and some of the flaws and advantages of the format are discussed, yet a fourth instruction is given:

(D) *Read through the list of words carefully. For each word: if you know what it means, make a mark in the box beside the word. If you don't know what it means, or if you aren't sure, then leave the box empty* (Meara 1996: 43).

This instruction appears to move away from the "Yes/No" decision that was emphasized in the EFL Vocabulary Test and reverts to the initial instructions where items had to be ticked.

When we compare the chronological evolution in these Yes/No test instructions, the following conclusions can be drawn. First, when using instruction (A) or (D), it is impossible to distinguish between "No-responses" and omitted responses in the data analysis. When using instruction (B), it is impossible to distinguish between "Yes-responses" and omitted responses. This matter is resolved in instruction (C) by urging the test taker to make a clear "Yes" or "No" decision for each item and to opt for the "No" response in case of doubt. In fact, instruction (C) is the only one that does the test format's name any credit. Second, in instruction (C) and (D) the possibility of doubt is overtly acknowledged when dealing with the test task and the test taker is told which response to choose when he or she is not completely sure. Third, none of the instructions mentions the presence of pseudowords in the test.

With reference to this last point, we would like to remark that throughout the use we have made of the Yes/No Vocabulary Test we have consistently opted to include a warning about the presence of pseudowords in the test instruction. There are several arguments that lie behind this decision:

1) Information about the presence of pseudowords in the test can be considered as an element of fundamental honesty. Test results always bear certain consequences for the test takers and these consequences may cause the test takers to take certain actions in order to maximize their scores (Shohamy 2001). The tendency for testees to adapt their behaviour in order to gain the benefits associated with high scores is something we encounter frequently with the student population. Therefore, it would not surprise us if some students claimed knowledge of all the items of the Yes/No Test if we withheld the information about the presence of pseudowords in the test. A test designer also needs to bear in mind that test takers (and especially university students who have been subjected to all kinds of test formats during their university career) are aware of the possibility of arriving at a correct response by guessing in a selected response type test like the Yes/No format. Nowadays, renowned language testers recommend encouraging test takers to make informed guesses

on the basis of partial knowledge (Bachman and Palmer 1996). Given the fact that pseudowords are made up of existing syllables of the target language and hence tap into partial knowledge of the target language's morphology, they risk being too attractive to testees to be rejected. In order not to put these test takers on the wrong track, they should be informed of the presence of the pseudowords. In view of the high rate of false alarms we have consistently obtained in the experiments, we assume that this rate would have been even higher if we had refrained from drawing the testees' attention to the presence of pseudowords in the test.

2) Furthermore, if the testees are left in the dark about the inclusion of pseudowords in the test, there is the possibility that some testees may develop an awareness of the presence of pseudowords in the test and others may not. This would create an imbalance in the data and would jeopardize the test's reliability.

3) Abels (1994) reported a Yes/No experiment in which the participants were given the same test twice. The first time the presence of pseudowords in the test was not mentioned, the second time the participants were told that the list contained pseudowords and they were allowed to alter the responses they had given the first time. She reports that the number of "Yes"-responses declined in both the word as the pseudoword category (words and pseudwowords) but there were more changes from a "Yes"- to a "No"-response in the pseudoword category than in the word category. The participants had clearly chosen a more careful response behaviour. Once they realized that the test contained pseudowords they were less prone to overestimating their vocabulary knowledge. Abels compared the Yes/No scores with a MC vocabulary test and a c-test and found that the scores yielded by the second test taking correlated better with both tests than the scores the participants had obtained the first time. She therefore argues that the test instruction should include a warning about the presence of pseudowords in the test.

4) According to Shohamy (2001) tests cannot only create fear or anxiety in testees but also subversion. If the testees have the impression that they were not told everything there is to know about a test, they might feel they have not been able to show the best of their abilities, which might cause frustration.

Finally, we come to the instruction of the most recently construed Yes/No Vocabulary Test that is part of the European DIALANG diagnostic tool, which will undoubtedly be considered the most authoritative Yes/No instrument since its use will be so widespread. On computer the following instruction is given in each of the target languages (here in English):

In the test, you will be presented with a collection of 'words', some of which are real, and some of which are invented. For each word, you must press the "Yes" button if you think the word exists. If you think it is an invented word, press the "No" button. (http://www.dialang.org)

Not only does this instruction contain a clear warning about the presence of pseudowords, it also changes the task significantly from knowing the meaning of a word to knowing whether a word exists in the target language. This instruction may well tap into a different level of vocabulary knowledge and therefore turns the format into a new and different test.

7.1.3 Particularity of the Yes/No test task

Special care should be taken in wording the instructions of a Yes/No Vocabulary Test because of the particularity of the Yes/No task. The specificities of this task have not been sufficiently recognized in the literature.
An important distinction is to be made between tests where the required responses to the items of the test are set a priori and tests that involve a kind of self-assessment. In the first test type the testees' responses either coincide with the required responses or they do not and they are corrected accordingly.
Example:
Translate into English
- *cheval:* …….

A self-assessment test is quite different in nature: the responses carry another status and their correctness cannot be verified at once.
Example:
You will be presented with a number of statements and you have to decide whether each one applies to you or not. Press the YES button if it does, and the NO button if it does not.
- I can ask someone for directions in French:

YES	NO

The task with which the testee is confronted in the Yes/No Vocabulary Test lies somewhere in between both test types. At this point it is important to point out the apparent similarities between the Yes/No test and a True/False test for they are often characterized as essentially the same format (for example: Huibregtse and Admiraal 1999). However, the tasks of the respective tests differ fundamentally and this has important consequences for the participant's response behaviour. A True/False test belongs to the type of tests where the responses are set a priori and can be verified immediately. Example:
Indicate if the following statements are TRUE or FALSE.
The word *fromage* is French for cheese:

TRUE	FALSE

When the testee chooses "TRUE", the response is correct. When the testee goes for "FALSE", the response is incorrect.

However, in a Yes/No test for French this item would turn into: "Do you know the meaning of the word fromage?", and the testee would be required to answer "Yes" or "No" without having to furnish any further information or indication as to prove that he really understands the meaning of the word. Therefore, this response cannot be verified. We have to take the

testee's word for it. This is something completely different than having to say whether a statement is true or not. The testee's Yes/No decision is based on his own evaluation of knowledge of the particular words. This generates the problematic issue of response biases because the decisions or judgement calls people make, are often inspired by how their personalities were formed, their cultural background, their upbringing or the expectancy pattern they hold concerning the outcome of a test. Holec (in Janssen-van Dieten 1992: 44) states that decisions can be prompted by the social learning environment and the way the learner relates to this environment. They can also be influenced by cultural factors like one's views on language or one's ideas on how language should be taught and evaluated. Finally, one should not overlook the importance of psychological parameters since self assessment is inevitably dependent on the characteristics of one's personality. In Cohen's list of personal characteristics that could potentially affect the test performance, he includes age, foreign language aptitude, socio-psychological factors, personality, cognitive style, language use strategies, ethnolinguistic factors and multilingual ability (Cohen 1994: 74).

Statistical analysis of the data of experiments 1 and 2 has demonstrated how the response behaviour of the participants was based not only on their lexical knowledge but also on their individual decision making process. The low correlation between the scores on the Yes/No Test and the scores on the Translation task in Experiment 2 illustrated that the wide range of cognitive and social factors that may have determined the testees' Yes/No decisions do not relate to the construct the Yes/No Vocabulary Test aims to measure.

In view of the particularity of the Yes/No task, more specifically the problems that may arise from the decision-making process, the instruction could be essential in the test design. So far, no attention has been paid to the test's instruction and its implications for the learner's choices. However, it may prove interesting to tailor the instruction appropriately so that the test takers are enabled to perform at their best. A different instruction could influence the validity of the testees' responses, especially in this particular test format in which the learners' responses bear such ambiguous status.

7.2 Experiment 3

7.2.1 Aim

After being confronted with a response bias in the previous experiments and the lack of concurrent validation resulting thereof, we aimed to eliminate the response bias through the test instruction. If a certain instruction could rule out the response bias that is now inherent in the task with which the participants are faced, this will have important consequences for the calculation of the test score and the validity of the test. We hoped to intervene in the participant's decision making process by reinforcing the test's instruction. Therefore an

experiment was set up in which widely divergent instructions were contrasted while using identical test material. Afterwards the participants were asked to translate the words of the Yes/No test so that we could verify the validity of their responses in the Yes/No test.

The central question of the experiment, i.e. whether the instruction could possibly influence or eliminate the response bias revealed by the participants in previous experiments, was broken down into two subquestions:
(1) Does the instruction have an influence on the false alarm rate displayed by the participants? (the false alarm rate is of importance since we will determine the presence of a response bias through an investigation of the correlation between the hits, i.e. Yes-responses to words, and the false alarms, i.e. Yes-responses to pseudowords).
(2) If we obtained a lower false alarm rate (because we have urged the participants to be more careful in their response behaviour and not to overestimate their vocabulary knowledge), would this then result in a higher correlation between the Yes/No Test and the Translation test?

7.2.2 Method

Participants
The participants were French-speaking university students (N=179) of Economics and Business Administration taking Dutch first level language courses. The participants were divided into an experimental group (N=103) and a control group (N=76) according to the group they belonged to (hence the uneven number of participants in the control and the experimental group) and each group was presented with a Yes/No Vocabulary Test. It was a paper-and-pencil test and the participants took the test in less than ten minutes.

Material
The same test material was used as described in Chapter 5 (Experiment1). Both parallel versions (I and II) of the Yes/No vocabulary test were used, each consisting of 60 words and 40 pseudowords (see Appendix 3). In order to control for sequence effects and to reduce the possibility of cheating, both test versions (I and II) were made up into three different item-orders. Previous dealings with the test versions (Chapter 5) had shown them to be equally difficult. Nevertheless, both versions were used in the experimental as well as in the control condition.

Instructions
The instructions were given in the participants' mother tongue. The instruction that was presented to the control group mimicked the instruction of the original Meara and Buxton Yes/No study (1987) apart from the fact that we included a warning about the pseudowords:

Indiquez à l'aide d'une croix les mots que vous connaissez.
Certains mots repris dans la liste n'existent pas en néerlandais!
(Tick the words you know. Some of the words in the list do not exist in Dutch.)

We considered this instruction to be rather minimal because it left much to the imagination of the test taker. The instruction that was presented to the experimental group was set out to be more rigorous:

Indiquez à l'aide d'une croix les mots dont vous connaissez la signification. En cas de doute, ne cochez pas le mot.
Attention ! Certains mots repris dans la liste n'existent pas en néerlandais. Nous vous demanderons par la suite de fournir la traduction de certains mots de la liste.
(Tick the words you know the meaning of. When in doubt, do not tick the item. Notice that some of the words in the list do not exist in Dutch. After completing this test, you will be asked to translate some of the words of the list.)

We considered the second instruction to be more strict or stringent than the first. Not only because we emphasized that they should "know the meaning of the word", which seems to be less open to discussion than simply to "know a word"; but also because we urged them to refrain from ticking a word unless they were sure of their reply. Finally, we announced in the instruction that the validity of their responses would be checked afterwards by means of a Translation test of the same items. We hoped that this would discourage dishonest or uncertain response behaviour and that the false alarm rate would diminish as a result of this.

Translation Test
The participants were presented with the 60 existing Dutch words and were asked to provide a translation for each item in their mother tongue. It was a paper-and-pencil test and the participants performed the task in about 15 minutes.

As in Chapter 6, we assumed the translation to measure a well-defined construct: the extent to which the participants are able to provide an L1 translation of L2 words that belong to the core vocabulary of Dutch. When a participant ticked the word "stoel" (chair) in the Yes/No test and could provide us with a translation of it in his mother tongue afterwards, we interpreted this as a validation of the response in the Yes/No test. When the translation was wrong or lacking, we concluded that the participant had misjudged his knowledge of this particular item.

7.2.3 Results

The reliabilities for the words were presented separately from the reliabilities for the pseudowords (Table 7.1) in order to illustrate that the pseudowords play an important role in the test format. The considerable reliabilities rendered by the pseudoword-items indicated that there was a systematicity to them, which contradicted the initial idea that pseudowords function randomly and are a marginal phenomenon in the test, merely serving to correct a potential overestimation of the participants. The data showed that the reliabilities of the pseudoword-items were nearly as high as the reliabilities of the word-items.

Table 7.1: Scores and test reliability of word- and pseudoword-items for both parallel versions of the Yes/No Vocabulary Test (I and II) for the experimental condition (strict instruction) and the control condition (minimal instruction).

Test Version	Instruction	Words /60				Pseudowords /40				Correl. w/pw
		mean	SD	%	rel.	mean	SD	%	rel.	
I	Minimal (N=33)	33.12	9.55	55.20	.866	34.39	4.26	85.98	.718	-.580***
	Strict (N=59)	38.85	9.18	64.75	.873	36.20	3.44	90.50	.769	-.512***
II	Minimal (N=43)	41.05	7.98	68.42	.880	34.79	4.39	86.98	.816	-.153
	Strict (N=44)	36.75	6.76	61.25	.825	36.98	2.19	92.45	.498	-.333*

Notes: The reliabilities are calculated with Cronbach's alpha. Significant correlations are marked with * ($p<.05$), ** ($p<.01$) and *** ($p<.001$).

In accordance with the findings in Experiment 1 and 2, a negative correlation was observed between the measure of the word-items and the measure of the pseudoword-items, and this for both test versions and for both the experimental as the control condition (see Table 7.1). The data of both test versions still suffered a response bias and while this bias seemed to be slightly reduced in the experimental condition of Test version I (from -.580 to -.512), this was not the case in Test version II where the negative correlation increased from -.153 in the control condition to -.333 in the experimental condition.

One of the most important results in this experiment concerned the influence of the instruction on the false alarm rate displayed by the students. A comparison of the matrices in Figure 7.1 revealed that the participants' response behaviour was particularly influenced with regard to the pseudoword-items and not so much with regard to the word-items. In the experimental condition the false alarm rate dropped to 8.7% (versus 13.5 % in the control condition). This decrease in the false alarm rate was shown to be significant (t-test, $F=13.05$; $df=1,177$; $p=.001$).

Item alternative	Response alternative Minimal instruction: Yes	Response alternative Minimal instruction: No	Response alternative Strict instruction: Yes	Response alternative Strict instruction: No
Word	Hit 62.7%	Miss 37.3%	Hit 63.3%	Miss 36.7%
Pseudoword	False alarm 13.5%	Correct rejection 86.5%	False alarm 8.7%	Correct rejection 91.3%

Figure 7.1: The item-response matrices of the Yes/No Vocabulary Tests for the control condition (minimal instruction) and the experimental condition (strict instruction). Percentages are calculated within each item alternative.

In order to corroborate this result an analysis of variance (ANOVA) was executed with "False alarm rate" as dependent variable and "Instruction" and "Vocabulary knowledge" as independent variables. The "Vocabulary knowledge" variable was fixed by means of the results the participants obtained on the Translation test and was split up into five different levels. This analysis showed the variable "Instruction" to be significant ($F= 12.97$, $df= 1, 169$, $p= .00$). There was no significant effect of the "Vocabulary knowledge" variable on the false alarm rate ($F= .78$, $df = 4, 169$, $p= .54$). The interaction of the variables "Instruction" and "Vocabulary knowledge" was not significant ($F= 2.30$, $df= 4, 169$, $p= .06$). These results allowed us to conclude that the instruction had an important effect on the false alarm rate of the participants, independent of their respective vocabulary knowledge.

An analysis of covariance (ANCOVA) was also carried out with "False alarm rate" as dependent variable and "Instruction" and "Vocabulary knowledge" as independent variables. The "Vocabulary knowledge" variable was again fixed by means of the results the participants obtained on the Translation test and served as a covariate. The analysis showed the interaction of the variables "Instruction" and "Vocabulary knowledge" to be not significant ($F= 1.84$, $df= 1, 175$, $p > .05$). There was, however, a significant effect of the variable "Instruction" on the false alarm rate ($F= 13.84$, $df = 1,$

176, $p < .001$). This reinforced the conclusion that the response behaviour of the participants can be influenced by means of the instruction[17] .

The test scores were calculated according to the formulae discussed in Chapter 5. The group of participants that took Test Version I with the minimal instruction obtained markedly weaker scores than the other groups. If this was the result of a difference in the level of proficiency, it would undoubtedly be confirmed by the results of the translation task.

Table 7.2: Results on the Yes/No Vocabulary Tests with the different methods of scoring. The raw score is the number of hits and the corrected scores are based either on the all-or-nothing model (cfg) or on the continuous model (I_{SDT} and Hcfb).

Test Version	Instruction	Formula	Test score /60		
			Mean	SD	%
I	Minimal	Raw (hits)	33.12	9.55	55.20
	(N=33)	cfg	29.14	9.45	48.57
		I_{SDT}	27.42	6.88	45.70
		Hcfb	29.10	7.84	48.50
I	Strict	Raw (hits)	38.85	9.18	64.75
	(N=59)	cfg	36.88	9.34	61.47
		I_{SDT}	35.50	6.54	59.17
		Hcfb	38.53	6.84	64.21
II	Minimal	Raw (hits)	41.05	7.98	68.42
	(N=43)	cfg	37.95	9.26	63.23
		I_{SDT}	34.60	9.14	57.67
		Hcfb	36.41	10.04	60.68
II	Strict	Raw (hits)	36.75	6.76	61.25
	(N=44)	cfg	34.89	6.95	58.13
		I_{SDT}	34.74	5.47	57.88
		Hcfb	37.95	6.51	63.25

The translation task was carried out on paper and it was corrected with the same degree of leniency as described in Chapter 6 (Section 6.2.3) in order to maximize the chances for concurrent validation. The reliabilities of the translation task were higher than the one obtained in the previous experiment:

[17]The same analysis was carried out with as dependent variable the "Criterion", which is a reflection of the response bias. The analysis confirmed the results. The Criterion amounted to .43 for the control group and .55 for the experimental group. The difference between these values was significant and not dependent of the vocabulary knowledge of the participants: the interaction of the variables "Instruction" and "Vocabulary knowledge" was not significant, $F = .92$, $df = 1, 175$, $p > .05$ and again there was a significant effect of the "Instruction" variable on the false alarm rate., $F = 20.72$, $df = 1, 176$, $p < .001$. The participants of the "strict" group shifted their Criterion towards the pole "Yes, I know the meaning of the word", which means they displayed a more careful response behaviour and opted for the answer "No" when they were in doubt.

.884 for Test version I and .852 for Test version II versus .789 in Chapter 6 (Test Version I). The means amounted to 35.13 for Test Version I and 35.63 for Test Version II, they were slightly beneath the mean of Experiment 2 in Chapter 6 (36.58). With a score of 30.18, it was confirmed that the group of participants that took Test Version I with the minimal instruction had a lower proficiency than their peers in the other groups (37.90, 35.58 and 35.68 respectively).

Table 7.3: Descriptive statistics of the results on the Translation Test. The reliabilities are calculated with Cronbach's alpha.

Test Version	Instruction	Translation Mean /60	SD	Reliability
I	Minimal (N=33)	30.18	5.92	.827
	Strict (N=59)	37.90	6.89	.863
	Across conditions (N=92)	35.13	7.51	.884
II	Minimal (N=43)	35.58	7.86	.888
	Strict (N=44)	35.68	5.32	.784
	Across conditions (N=87)	35.63	6.66	.852

The central hypothesis of this experiment concerned the influence of the instruction on the correlation between the Yes/No Test scores and the results on a Translation task of the same words. It was already established that the false alarm rate diminished significantly as a result of a strict instruction. Regretfully, this did not lead to a better concurrent validity. Table 7.4 shows that the correlations between the scores on the Yes/No Vocabulary Test and the scores on the Translation test were stronger for the group that received the minimal instruction for all formulae, and this for both Test Versions. This ran counter to what we expected. Therefore, the hypothesis that a lower false alarm rate would result in a stronger correlation between the Yes/No test and the translation test was refuted. We noted that the values obtained with the SDT-formula were lower than those of the cfg-formula in either condition. This corroborated our previous conclusion (Chapter 5) that the SDT-model still has to be adjusted to the specificities of the Yes/No task.

Table 7.4: Correlation between the results on the Yes/No Vocabulary Test and the results on the Translation Test.

Correlation Yes/No Test and Translation Test					
Translation	Instruction	Raw (hits)	cfg	I_{SDT}	Hcfb
Test I	Minimal (N=33)	.797	.813	.590	.445
	Strict (N=59)	.768	.747	.574	.348
Test II	Minimal (N=43)	.827	.862	.735	.638
	Strict (N=44)	.613	.654	.661	.518

Item analysis showed that concerning the responses to the word-items, the "No" response + correct translation pattern was about the same for both the control (9%) as the experimental (10%) group. It also revealed that this pattern was present throughout the test material, for cognates as well as non-cognates although some cognates appeared to be relatively susceptible to this tendency. Similarly to what we reported in the previous chapter, some cognates appeared to get rejected because the participants found their resemblance to words in their mothertongue too strong (e.g. "militair", "opereren", "moskee", "organisme", "biljart", "effect"). The participants dismissed these items probably because they could not believe them to be Dutch words. However, when they encountered them in the translation task, they realized that the words actually existed in Dutch and they translated them correctly.

7.2.4 Comparison of the main results with the results in previous experiments

Figure 7.2 sums up the item-response matrices of the experiments that have been discussed so far. The experiments can be compared in their materials (1), instruction (2) and the fact whether a Translation test (3) was included or not.

(1) Experiment 1 and 3 contained the same test material: two parallel versions of a Yes/No Test. Experiment 2 only used one of the test versions (Version I).

(2) Experiment 1 presented the participants with the same instruction that was passed on to the participants in the control condition of Experiment 3. In Experiment 2, the instruction was orally exemplified.

(3) In both Experiment 2 and 3 the Yes/No responses were validated by means of a Translation test.

Although the same material was used, there was a remarkable difference between the false alarm rates of Experiments 1 and 2 (respectively 20.5% and 24.2%) and the false alarm rates of Experiment 3 (13.5% for the control group and 8.7% for the experimental group), which were much lower. The difference in false alarm rate between Experiment 1 and the control condition of Experiment 3 could not be explained through a difference in test instruction since it was identical. This raised the question what factor could have caused the participants in Experiment 3 to display such significantly more careful response behaviour than the participants in previous experiments.

Furthermore, the correlations between the Yes/No Test scores and the results of the Translation test are much higher in Experiment 3 than in Experiment 2 (where it was astonishingly low), as can be seen in Table 7.5. Here also, the question arose as to what might have caused the responses of the participants to be more valid than in previous experiments.

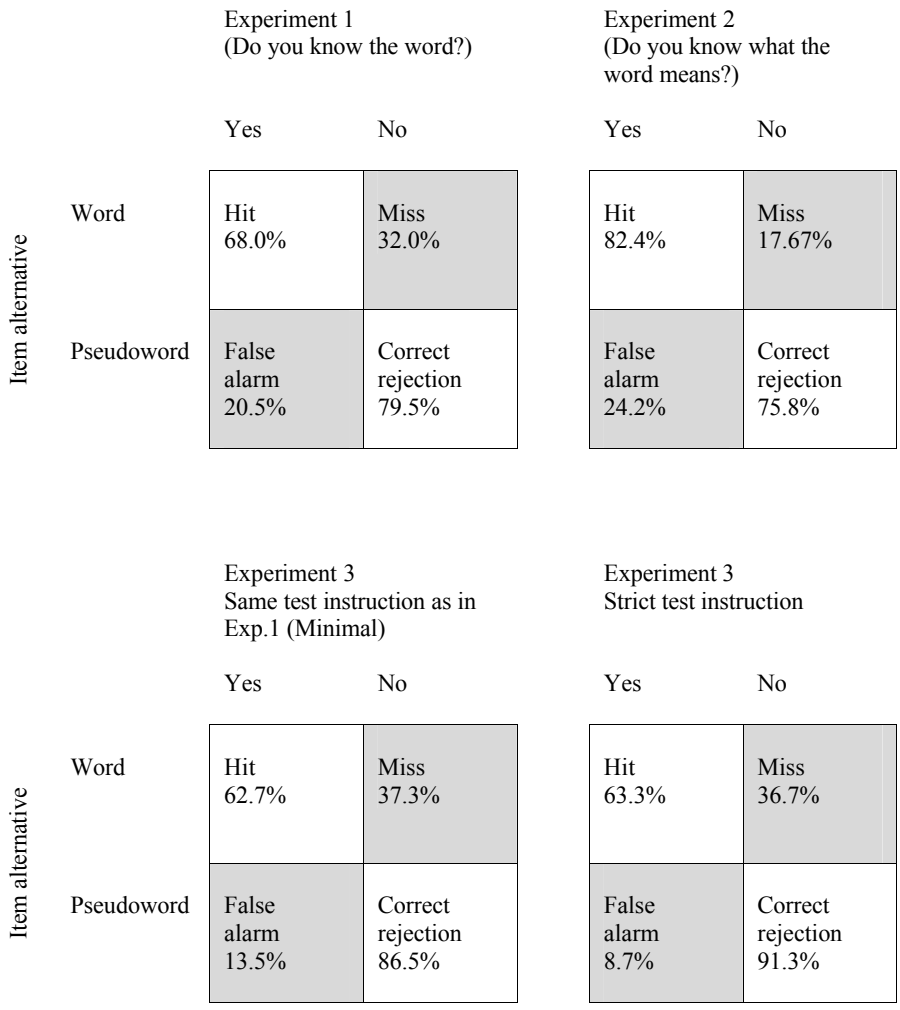


Figure 7.2: The item-response matrices of the Yes/No Vocabulary Tests for Experiments 1, 2 and 3. Percentages are calculated within each item alternative.

Table 7.5: Correlation between the results on the Yes/No Vocabulary Test and the results on the Translation Test for Experiment 2 and Experiment 3 for the raw score.

Translation	Instruction	Raw (hits)
Experiment 2	Minimal	.406
Experiment 3	Minimal	.851
	Strict	.728

The only explanation for these diverging results lies in the test circumstances. Experiment 1 was part of a placement test procedure and took place in a large theatre with anonymous supervisors. In Experiment 2, the Yes/No Test was delivered in a computer room where supervision was not very tight. By contrast, the tests in Experiment 3 were administered in the classroom by the participants' own teachers. We suspect that they may have felt compelled to respond more carefully because they were in small groups and under watch and ward by their proper teacher. This shows how the results of the Yes/No Test are extremely susceptible to conditions and circumstances. The format seems to lack robustness vis-à-vis the many variables that come into play in a testing situation.

Finally, we would like to draw attention to the difference in the "No" response + correct translation pattern for Experiment 2 and 3. The tendency to reject words yet translate them correctly afterwards was more than twice as large in Experiment 3 than in Experiment 2. In Experiment 2, approximately 17% of the existing words elicited a "No" response and one out of four of these was correctly translated afterwards. In Experiment 3, almost 37% of the existing words elicited a "No" response and again one out of four was translated correctly afterwards. This gives the impression that the participants in Experiment 3 were often prone to underestimating their vocabulary knowledge rather than overestimating it. The participants have responded with more caution in this experiment but this has not increased the validity of their responses, for 25% of the rejected word-items were translated correctly. In spite of the fact that the lower false alarm rate seemed to suggest a more careful and therefore more valid response behaviour, the response in Experiment 3 seemed to exhibit a blatant uncertainty within the participants when it came to deciding whether they knew the meaning of a word or not.

7.2.5 Discussion

The empirical data allow us to answer the central research questions of this experiment. First, we have established that the use of a strict instruction can lead to a significant decline of the false alarm rate. To put it in SDT-terms: when using a strict instruction, we observe that the participants shift their Criterion. However, the correlation between the Yes/No test and the translation did not improve as a result of this, quite to the contrary. The hypotheses that a Yes/No test with a rigorous instruction would lead to a higher validity of the participants' responses is countered by the results. A possible explanation for this finding could be that we have added an extra variable to the experiment: the participants' susceptibility to rigorous warnings in the instruction. This susceptibility is clearly independent of the vocabulary knowledge of the participants. Influencing participants' decision behaviour does not result in a more valid Yes/No Test.

We also observe that although the formulae based on continuous models (I_{SDT} and Hcfb) allow us to distinguish the response bias from the receptive vocabulary knowledge, they yielded lower correlations than the other formulae. As the SDT-model is further developed for use with the Yes/No test, researchers will have to verify if the model succeeds in separating the response bias from the vocabulary knowledge whilst maintaining a sufficiently high reliability. Such a refined SDT-model would have to account for distributions that are less spread out in the case of stronger testees.

The experiments we have reported seem to indicate that there is an interplay between the multitude of characteristics that determine the learner (cultural, psychological, sociological, etc.) and the vocabulary knowledge that the Yes/No Test is supposed to be measuring. Of the many factors that can affect test performance (individual characteristics of test takers, unexpected disturbances during the test administration, etc.) the characteristics of the test task are the only ones that are under the control of the test developer. In tests like the Yes/No Vocabulary Test (relying on self-assessment or decision-making) the variables that underly the validity of the responses need to be determined and investigated. In Section 7.3 of this Chapter, we will therefore consider the specific design of the Yes/No Test and how a computer-controlled format design may help make the testees' responses more valid.

7.3 Influence of the computer-controlled format design on the response behaviour

7.3.1 Computer-based testing

Computer-based testing has witnessed rapid growth in the past decade and this trend will undoubtedly persist in the years to come as the internet is increasingly used to deliver tests to users. The advantages computers have to offer to testing in general or language testing in particular are numerous (Chapelle 2001). Test compilation and construction are facilitated, test delivery across the world has become a reality through the internet, the problem of deciphering student handwriting is eliminated and test scores or exam results can be provided instantly. Computers can also be used for storing tests and details of candidates so that accurate and consistent evaluation becomes possible, this serves as an important diagnostic assistance to teachers. Test items can be tried out, calibrated and added to item banks, which is a huge relief to test writers. Rapidly evolving computer software allows swift access to banks of test items and permits teachers to tailor a test to the ability level of examinees (Meunier 1994). With regard to test analysis, experimental findings seem to reveal the superiority of computer-based tests to paper-and-pencil tests in terms of reliability and validity, particularly when relatively few items are administered.

Apart from the practical advantages that computers offer in test administration and analysis, we briefly need to mention the tremendous benefit of computerised corpora construction. Because discrete vocabulary tests serve to evaluate whether learners know the lexical items they need to meet certain learning objectives, identifying or selecting those items on a principled basis becomes a significant issue. Computer corpus software allows us to calculate the frequency of words or count the occurrences of word forms in large sets of texts more efficiently than was possible in the past (e.g. the COBUILD Bank of English totals more than 300 million words). Recurring combinations of words and the contexts in which particular words occur can be identified because concordance programmes can rapidly assemble multiple examples of a particular word or phrase, each in its linguistic context. By means of commercially published corpus programs learners have the opportunity to create their own vocabulary lists for learning.

The development of Computer Adaptive Language Tests (CALT) has also caused an upsurge in the language testing domain (Chapelle 2001). When taking Computer Adaptive Language Tests individual test-takers respond to a set of items selected successively by the computer from an established item bank on the basis of the individual's pattern of responses. Thanks to this procedure the tests are uniquely tailored to each individual and automatically terminated when the examinee's ability level has been determined. The process of developing computer-adaptive tests is thought to help illuminate the relationship between language and cognition because it will provide insight into individual learner differences. They are described as psychometrically sound and unusually efficient testing instruments that are generally much more precise and much shorter than conventional paper-and-pencil tests (Gervais 1997).

The basic item types for discrete vocabulary testing (checklist, multiple choice, matching, blank-filling) are very attractive for computerised presentation, and context-independent items in particular lend themselves well to computer adaptive testing. At present there are, to our knowledge, two computerised versions of the Yes/No Vocabulary Test, the Eurocentres Vocabulary Size Test (Meara 1992) and the Yes/No Vocabulary Test that is part of the DIALANG test battery (http://www.dialang.org).

The Eurocentres Vocabulary Size Test is a vocabulary placement test commissioned by Eurocentres, a network of language schools and it is a computer adaptive language test. It presents the test taker with a sample of words covering numerous frequency levels. If the test-taker achieves a criterion level of performance, the program proceeds to the next level. If not, it is assumed that the testee has reached the upper limit of his vocabulary knowledge and a further set of 50 words from the same frequency level are presented in order to fine tune the learner's vocabulary size estimate. Regretfully, this software package is no longer available today.

The computerized Yes/No Vocabulary Tests that are part of the DIALANG test battery can be accessed by every language learner around the world. These tests are not computer adaptive and they work with a much smaller test sample than the Eurocenters Test (the Eurocentres Test draws on ten frequency bands and presents a random sample of 20 words from each band, DIALANG works with one fixed sample of 75 words for each of the fourteen languages for which it has been developed).

7.3.2 The design of the computer format as part of the test characteristics

Now that we have established that computers have a great potential to contribute to language test construction and assessment, it is important to acknowledge the influence the computer design may have on the characteristics of a particular test. Gervais (1997) argues that the accuracy of a computer-based test versus a traditional paper-and-pencil test can be compared by addressing the advantages of a computer-delivered test in terms of accessibility and speed of results on the one hand, and possible disadvantages in terms of bias against those with no computer familiarity or with negative attitudes to computers on the other hand. However, this is a very narrow conception of the possible effects computer-based language tests may exert. When transferring a paper-and pencil test to a computerised format, one should realize that the particular computer design one programmes can influence the response behaviour of the test taker or even alter the test task altogether. Through computer programming time limits per item can be built in, the possibility of omitted responses can be ruled out, items can be singled out on separate screens as a result of which test takers can be denied an overview of the test, etc. All these factors may affect the individual's test performance. Therefore they too have to be considered as part of the characteristics of the test task.

It is our view that when designing a computerised version of a paper-and-pencil test, one should go beyond the advantages the computer offers in terms of speed and ease of test correction, and also exploit the computer's possibilities to gain more control over the testing situation. Confronted with the multitude of factors (individual characteristics of test takers, temporary changes in their physical or mental condition, etc.) that may impede or complicate the already difficult relation between the testee's test performance and the testee's actual knowledge or skill, attempting to control the test task characteristics by design is the best option any test developer has. A good interface design should reduce the possibility of construct-irrelevant variance, which may threaten the inferences that may be drawn from test scores (Fulcher 2003).

The central question we will deal with in this section is whether a computer-based test can offer any added value over a paper-and-pencil test in

the particular case of the Yes/No Vocabulary Test. If this were to be the outcome of the described experiment, then the value of technology for language learning and in this case vocabulary testing would be all the more significant.

7.3.3 The computer-controlled environment in the particular case of the Yes/No Vocabulary Test

To our knowledge, the Yes/No Vocabulary Test that is part of the DIALANG diagnostic tool, is the only computerized Yes/No test that is readily accessible today. Apart from the Yes/No Vocabulary Tests in the DIALANG test battery and the Eurocentres Vocabulary Size Test, Meara and other scholars have almost exclusively worked with paper-and-pencil tests (Read 2000).

When designing a computerised version of a paper-and-pencil test two approaches seem feasible: one can mimick the paper-and-pencil format as closely as possible or one can make the most of the computer's advantages to control the test characteristics. Within the DIALANG assessment frame the first option has been taken. From the point of view of computer design, the DIALANG Yes/No Vocabulary Test resembles the paper-and-pencil Yes/No tests. All the items are presented on the screen in a list, the test taker chooses in what order to respond to the items, responses can be altered and there is no time limit involved in taking the test. The only control that is built in and that clearly distinguishes the test from a paper-and-pencil version from a structural perspective is the fact that the test taker is forced to respond to all items before quitting the test. This is a control measure that is taken with a view to correcting the test and it guarantees the exclusion of omitted responses (remember that individual differences in responding or not responding to unknown items will distort the testees' rank order, cf. Chapter 5), it turns the test into a forced decision task.

However, there are other ways in which the computer can furnish a more controlled environment for the Yes/No Vocabulary Test:
1)When programming a computer application of the Yes/No format sequential operations can be preferred instead of sticking to the traditional presentation of items in a list. This way, the items appear on the computer screen one by one. This aspect is of greater importance than one would think because it changes the test experience drastically. The testees do not have an overview of the complete test (which is the case in the traditional "list"-presentation where all the items are presented right in front of the test taker). They do not know how many items are still to come, nor how many of them they have already rejected. They cannot alter the choices they have already made and they cannot ponder their choice by deciding to leave a particular item unanswered and get back to it when they have skimmed through the remaining items. All aforementioned aspects might influence the testees' response pattern.

2)A computer application can be designed to present the items to the different test takers in a random order, in order to prevent sequence effects, i.e. differences due to fatigue or boredom when responding to the last items of the test. This is hard to do with a paper and pencil test since one can only work with a limited set of fixed orders.
3)With reference to the problem of omitted responses, two approaches seem possible in a computerized version of the test. Either the test is designed to elicit a response for each item, which means the test taker will not get a test result without having responded to all test items (i.e. omitted response category is ruled out), or the test allows for omitting responses (test takers can leave an item unanswered and move on to the next one) but records when it happens.
4)Computer programming allows imposing a time limit per item. A time limit can serve several goals but the most important one is that it leads to more uniformity because the time variable no longer comes into play. It renders the test more univocal for all test takers.
5)The test instruction can be repeated on each screen in order to remind the test takers of the exact nature of the task they are expected to perform. This explicit reinforcement of the nature of the decision might result in a more consistent decision behaviour.
We presuppose that certain constructional aspects that seemed problematic in earlier paper-and-pencil experiments might be better dealt with in a specifically designed computer application because of the more controlled environment it could provide. We hypothesize that the more controlled environment would result in a less biased response behaviour of the testees and consequently a decrease in the false alarm rate.

7.4 Experiment 4

7.4.1 Aim

In this experiment the influence of the computer test design on the particpants' test performance was investigated. Two computer applications (A and B) of the Yes/No Vocabulary Test were compared. The focus of the experiment was on the role and the influence of these different computer test designs on: (1) the false alarm rate, (2) the correlation between the performance on words versus pseudowords, and (3) the external validation of the Yes/No Vocabulary Test in both cases. After the participants had taken the computerized Yes/No Vocabulary Test, they were asked to translate the words of the Yes/No test so that the validity of their Yes/No responses could be verified.

7.4.2 Method

Participants
Similar to Experiment 3, the participants were French-speaking university students of Economics and Business Administration following Dutch first level language courses. Two groups (a total of 125 participants) were administered a computerized Yes/No Vocabulary Test in order to evaluate their knowledge of the Dutch core vocabulary.

Material
A Yes/No Vocabulary Test was constructed that consisted of 60 words and 40 pseudowords (see Appendix 4). The test sample was a random selection of words from Woorden in Context (Dieltjens et al, 1995, Dieltjens et al. 1997) but afterwards this selection was modified according to the following restrictions:
1) The test sample was restricted to nouns and verbs on the assumption that these grammatical categories generally carry stronger lexical meaning than for instance adverbs or prepositions and should therefore be easier to recognize when encountered in isolation.
2) Cognates were banned from the test material. Although previous item analysis (cf. Exp. 2) has shown that the cognates are not responsible for the overestimation revealed by the participants, it could be argued that the mere presence of cognates in the test material could have elicited an uncertain response behaviour in the participants which has led them to overestimate their vocabulary knowledge when confronted with non-cognate words and pseudowords.
Although the experiment centered around the question of computer design and how a more controlled computer design might provide a pathway to reduce the response bias, questions about how the new test content may have influenced the qualities of the test will be considered when the results are discussed.

The computerized Yes/No Vocabulary Tests
The tests that were constructed for this experiment differed only in their computer design. Computer application A was designed to resemble the paper-and-pencil version of the Yes/No test and had the following characteristics:
-All the items of the test were presented on the screen in a list allowing a complete overview of the test items. If the testee wanted to, he could count the total number of presented items.
-The item order within the list remained the same regardless of how many times the programme was accessed. This had as a consequence that the item order was the same for all testees.
-The participants could scroll up and down the list and they had the opportunity to

change their responses as many times as they wanted. When confronted with a test that consists of a set of items, participants may want to count the number of times they have responded "No" in proportion to the total number of test items, and they may consequently wish to change their responses.
-There was no time limit imposed but the participants' individual test taking times were recorded by the computer. They could not end the test unless all items had been answered. As a consequence there were no omitted responses to be dealt with in the data analysis.
-The instruction for computer application A was the following: Indiquez les mots dans la liste dont vous connaissez la signification. Attention : Certains mots repris dans la liste n'existent pas en néerlandais! (Mark the words in the list of which you know the meaning. Beware : the list also contains words that do not exist in Dutch.)

Computer application B was designed to make the most of the computer's potential to provide a controlled environment. The way in which this control was exerted is listed below :
-The test items were presented to the testees sequentially. The words appeared on the screen one by one. This sequential aspect of operations allows the test developer more control of the test taker's response pattern, as we have described in Section 6.2.3. In short: the testee has no knowledge of the total number of items in the test, and it is practically impossible to keep tally of the number of items one has responded "No" to. Responding to particularly difficult items cannot be postponed until later and decisions cannot be altered in retrospect.
-The items were presented in a different and random order each time the programme was accessed.
-Two buttons were created on the screen: one with the text "je connais ce mot" (I know this word) and one with the text "je ne connais pas ce mot" (I do not know this word). With every item, these buttons re-appeared and the testees had to click one of them.
-There was no time limit. On the one hand a time limit per screen seemed attractive because it can be argued that it should not take these participants long to identify known core vocabulary and a time limit might prevent them from dwelling on their knowledge of the items. On the other hand the pressure of having to respond within a time constraint could lead to biased responses (which is why we decided against it).
-The possibility of omitting responses was excluded by designing the computer programme in a way that testees could not skip items. This turned the test into a forced decision task.
-The instruction for computer application B read: *Cliquez sur le bouton JA si vous connaisez la signification du mot qui apparaîtra à l'écran. Cliquez sur le bouton NEE si vous connaissez pas la signification du mot. Attention: Certains mots repris dans la liste*

n'existent pas en néerlandais! (You will be presented with a set of words. Click the JA button if you know the meaning of the presented word, click the NEE button if you do not know the meaning of the presented word. Beware : the set also includes words that do not exist in Dutch.) The instruction was repeated with each new screen because we hoped that this would reinforce the nature of the decision task, which in turn might result in a more consistent decision behaviour.
-Special care was taken to avoid double jeopardy (inadvertently evaluating not only language but also computer expertise). Before starting the test, the testees were given a warm-up session in order to familiarize them with the computer application.

The computerized Translation Task
After they had finished the Yes/No Vocabulary Test, the testees took a Translation test which had not been announced. The participants were presented with the 60 existing Dutch words of the Yes/No Vocabulary test they had just completed and were asked to provide a translation for each item in their mothertongue. The target words were presented on the screen consecutively.

The computer application offered the advantage that correction became less time consuming. It also suppressed the problems arising from decoding the testees' handwriting. We turned the computer correction into a kind of human-assisted scoring because we considered all the given responses ourselves before feeding the computer the correction key which responses to accept and which to reject.

As before, we assumed the translation to measure a well-defined construct: the extent to which the participants are able to provide an L1 translation of L2 words that belong to the core vocabulary of Dutch.

7.4.3 Results

In accordance with the results of Experiment 3, the reliabilities showed that the pseudowords functioned systematically in the test (see Table 7.6). The mean scores for the word-items (51.14 for computer application A and 52.69 for computer application B) were higher than in Experiment 3 (mean score of 37.44, across conditions) but the mean scores for the pseudoword-items were lower (30.64 for computer application A and 30.31 for computer application B versus 35.59 in Exp. 3, across conditions).

Table 7.6: Test reliability of word- and pseudoword-items of the Yes/No Vocabulary Test for Computer Application A and Computer Application B.

Computer application	Words /60				Pseudowords /40				Correl. w/pw
	mean	SD	%	rel.	mean	SD	%	rel.	
A (n=64)	51.14	5.39	85.23	.789	30.64	4.83	76.60	.779	-.345*
B (n=61)	52.69	4.34	87.82	.719	30.31	4.16	75.78	.696	-.005

Notes: The reliabilities are calculated with Cronbach's alpha. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).

The correlation between the measure of the performances on words and the measure of performances on pseudowords was negative for computer application A (r=-.345*). For computer application B there was no correlation between the measure of the performance on words and the measure of the performance on pseudowords (r=-.005, not significant) (see Table 7.6). In previous experiments, high false alarm rates were always accompanied by a negative correlation between the ability to identify words and the ability to reject pseudowords. The presence of a negative correlation was an indication of the presence of a systematic response bias in the data, the definition of response bias being that participants express a preference for one particular response alternative when taking the test. In the case of the experiments we have described, this has always been a preference for the "Yes" response, which has resulted in high scores on the word-items, low scores on the pseudowords-items and consequently severely corrected global test scores. Although we found no evidence of a negative correlation between the performances on word-items and those on pseudoword-items in the data of computer application B, this did not exclude a possible presence of a response bias, for the raw data were characterized by a low score on the pseudoword-items (see Table 7.7), which means we were again confronted with a high false alarm rate. The reason why the overestimation by the participants did not translate into a systematic negative correlation might be found in the sequential aspect of computer application B (remember that the items were presented one after another and participants could not skip them and return to them later), which might have prevented the participants to keep their Criterion stable throughout the test.

The false alarm rates were high (see Figure 7.3) and exceeded the rates we had obtained in previous experiments. An investigation of the matrices also revealed a much higher hit rate than in the previous experiment (85.2% and 87.8% versus 62.7% and 63.3% in Exp.3). The exclusion of cognates in the test material has not prevented the participants from overestimating their vocabulary knowledge blatantly.

Item alternative		Computer Application A Response alternative Yes	 No	Computer Application B Response alternative Yes	 No
	Word	Hit 85.2%	Miss 14.8%	Hit 87.8%	Miss 12.2%
	Pseudoword	False alarm 23.4%	Correct rejection 76.6%	False alarm 24.2%	Correct rejection 75.8%

Figure 7.3: The item-response matrices of the Yes/No Vocabulary Tests for Computer Application A and Computer Application B. Percentages are calculated within each item alternative.

With an average of 24.2% false alarms the participants of computer application A did not obtain a lower false alarm rate than the participants of computer application B (23.4% false alarms).

Table 7.7: Results on the Yes/No Vocabulary Tests with the different methods of scoring. The raw score is the number of hits and the corrected scores are based either on the all-or-nothing model (cfg) or on the continuous model (I_{SDT} and Hcfb).

Computer Application	Formula	Test score /60		
		mean	SD	%
A	Raw (hits)	51.14	5.39	85.23
(n=64)	cfg	48.56	6.57	80.93
	I_{SDT}	37.93	6.91	63.22
	Hcfb	39.03	7.10	65.05
B	Raw (hits)	52.69	4.34	87.82
(n=61)	cfg	50.19	6.04	83.65
	I_{SDT}	38.83	7.18	64.72
	Hcfb	39.96	7.56	66.60

Compared to Experiment 3, the mean scores were much higher (although they were severely reduced by the correction formulae as a consequence of the high rate of false alarms). This difference in scores between Experiment 3 and Experiment 4 could not be attributed to a difference in proficiency since the participants of both experiments had approximately the

same level of competence (they were all second-year university students taking the Dutch first level course programme). These high scores might be explained by the fact that the test material turned out to be easier than in Experiment 3. The results of the Translation task will provide more information concerning the degree of difficulty of the test material.

The test reliability of the Translation task was calculated with Cronbach's Alpha and amounted to .867 for computer application A (n=64) and .865 for computer application B (n=61). With an average score of 32/60 the participants of computer application B appeared to be slightly stronger in vocabulary than the participants of computer application A who obtained an average score of 28/60. The discrepancies between the scores on the Translation task and those on the Yes/No Test were substantial.

Table 7.8: Descriptive statistics of the results on the Translation task.

Computer Application	Translation score Mean /60	SD	Reliability
A (n=64)	27.91	7.85	.867
B (n=61)	31.77	7.93	.865

When we compared the test scores between Experiment 3 and Experiment 4, we observed that while the participants of Experiment 4 score much higher on the Yes/No Test than the participants of Experiment 3, the reverse was true for the Translation task (27.91 and 31.77 for Experiment 4 versus 35.38 in Experiment 3, across conditions). The test material was not easier than in Experiment 3, quite to the contrary. This raised the question as to what may have caused the participants to accept so many items in the Yes/No Test of Experiment 4.

The correlations between the results on the Yes/No Vocabulary Test and the Translation Test (Table 7.9) were relatively low for Computer Application A (r=.663 [Raw]; r=.737 [cfg]; r=.693 [I_{SDT}]; r= .741 [Hcfb] as well as for Computer Application B (r=.704 [Raw]; r=.731 [cfg]; r=.677 [I_{SDT}]; r= .719 [Hcfb]) and they resembled the correlations that were obtained in Experiment 3. With the exception of the correlation between the raw Yes/No scores (number of hits) and the translation scores, the correlations were weaker for computer application B than for computer application A. This means that the validity of the Yes/No Test that was administered under the experimental condition was not superior to the one that mimicked the paper-and-pencil version of the test.

Table 7.9: Correlation between the results on the Yes/No Vocabulary Test and the results on the Translation Test.

	Correlation Yes/No test and Translation	
Computer Application	Yes/No formula	Correlation with Translation
A (n=64)	Raw	.663
	Cfg	.737
	I_{SDT}	.693
	Hcfb	.741
B (n=61)	Raw	.704
	Cfg	.731
	I_{SDT}	.677
	Hcfb	.719

An item analysis was carried out in which the responses to words in the Yes/No Test for both computer applications were matched with the translations that were given for these items. Table 7.10 reflects the four possible patterns that resulted from these matches. Of the two possible responses to words in the Yes/No Test ("Yes" and "No"), the participants could have rendered either a correct or an incorrect translation of the item in the translation task.

Table 7.10: The four possible patterns that result from the match between the responses to words in the Yes/No Vocabulary Test and the translation task for Computer application A and Computer application B.

	Responses to words in the Yes/No Test		Translated words	
	Response	%	Correct	Incorrect
Computer Application A	Yes	85	53 %	47 %
	No	15	12 %	88 %
Computer Application B	Yes	83	52 %	48 %
	No	17	13 %	87 %

Almost half of the word-items that evoked a "Yes"-response in the Yes/No Test were translated incorrectly. This confirmed earlier suspicions about the amount of trust that could be placed in the "Hit" responses of the Yes/No Test. The defective self-assessment of the participants when it comes to judging a word to be known seemed to run through both computer applications. These results coincided with the comparable false alarm rates we encountered in the data of both computer applications. Again, it seemed that a "cognate-free" test sample does not solve the problem of overestimation of vocabulary knowledge.

7.4.4 Discussion

With reference to the focus of this experiment, we can conclude that the controlled environment of computer application B did not have the desired influence on the participants' response behaviour. Although the raw data were not characterized by a negative correlation between the performance on word-items and the performance on pseudoword-items, the false alarm rate was not diminished. In fact, it was slightly higher for computer application B than for computer application A. The controlled environment of computer application B had not urged the testees to respond more carefully. On the contrary. Concurrent validation was rather weak and not higher in the case of the experimental condition: the correlations between the Yes/No Vocabulary Test and the translation task were unsatisfactory[18] , which is not surprising given the high false alarm rate. None of the correction formulae appeared to perform significantly better than the others.

The sequentially programmed computer Yes/No Test did not offer any added value in this experiment. None of the control measures that were taken seem to have contributed to the validity of the participants' responses and hence the validity of the test.

Finally, it was shown that excluding cognates from the test sample did not improve the qualities of the test. In situations where there is no confusion caused by the lexical resemblance between the participants' mother tongue and the target language, the overestimation remains imminent. These findings provide strong confirmation for the view that the response bias functions independently of lexical skills or linguistic factors.

7.5 Conclusion

At the onset of this Chapter we stated that in order to solve the dilemma of the choice of correction formula and to be able to re-assess the validity of the Yes/No format, ways had to be found to reduce the response bias observed in the participants in previous experiments. Because of the fact that the false alarm rate is essentially the surface phenomenon through which a response bias is revealed, both described experiments in this chapter were aimed at a reduction of the false alarm rate.

The first experiment (Section 7.2) was built on the realization that the proportion of hits and the proportion of false alarms and the relation between these two is not only dependent on the lexical knowledge of the participants, but also on the participants' awareness of the consequences their response behaviour might have. Urging the participants to a more careful or thoughtful

[18] We would like to remind the reader that the content of the translation task was exactly the same as the content of the Yes/No Vocabulary Test, which means that the lack of concurrent validation cannot be attributed to inferential factors.

response behaviour was considered a possible pathway to improve the validity of the responses in the Yes/No Test. The hypothesis was developed that the response behaviour in the Yes/No Vocabulary Test might be influenced by the instruction and that a more rigorous instruction would result in a different proportion between the number of hits and the number of false alarms. This hypothesis was confirmed. The false alarm rate decreased significantly but this did not result in a better concurrent validity. Influencing participants' response behaviour in the Yes/No Vocabulary Test does not automatically result in a more valid measure of vocabulary size. Moreover, this experiment has illustrated that while high false alarm rates can be seen as an indication of a reponse bias, the reverse is not true: a reasonable or low false alarm rate in the data is no guarantee for a valid test.

In the second experiment that was described in Section 7.4 of this Chapter, a different angle was taken. It centred around the hypothesis that a controlled computer design might render the testees' responses more valid because it would prevent the participants from tailoring their response behaviour to certain characteristics of the test such as the number of items of the test, or the number of times they had rejected items, etc. A controlled computer application might hinder the participants' tendency to develop the kind of test-taking expertise through which they try to manipulate the test in order to obtain a better test score. Regretfully, the controlled computer application did not lead to unbiased responses, quite to the contrary. It is possible that when confronted with the succession of isolated "words" in computer application B, the testees decided to be on the safe side and developed an even stronger bias for the "Yes" response than the testees of computer application A. Consequently, the concurrent validation of the Yes/No Vocabulary Test turned out to be weak.

Both experiments have revealed that we have not succeeded in reducing the response bias of the participants. It was shown that it is possible to influence participants' response behaviour by manipulating certain variables of the test but these do not overcome or counterbalance the inherent problem of the format, namely that two dimensions are measured at the same time: the vocabulary size of the participants and their own estimation of their vocabulary knowledge. It was the intention to investigate (and improve) the variables that underly the validity of the testees' Yes/No responses, but at the end of this chapter we have to contemplate the possibility that the fact that the bias lies hidden in the Yes/No task of the format itself is too strong an element to be counterbalanced by test design.

As a last resort, it was decided to infuse the Yes/No placement test with the DIALANG material for Dutch in order to verify if this new content would improve the qualities of the test. The results of this experiment will be reported in Chapter 8.

Chapter 8

DIALANG

At the end of Chapter 7 we reached the conclusion that our efforts to reduce the response bias of the participants were in vain. The question was raised whether the quality of the test content could lie at the source of the problems we have encountered in the several experiments. Is there something we were doing wrong and that others apparently were doing right? Did the content we used cause the high false alarm rate? Was the procedure for making up pseudowords responsible for the response bias in the data? In a last ditch attempt to eliminate or reduce the response bias an experiment was set up in which the content of the DIALANG Yes/No Vocabulary Test for Dutch was used as part of the placement test for Dutch at the beginning of the academic year 2000-2001. In the DIALANG diagnostic language testing system (http://www.dialang.org), the Yes/No Vocabulary Test is used to measure vocabulary size of testees around the world, regardless of their L1. This means that the system also targets the large French-speaking community.

After presenting some general considerations about the selection of the test content and the way in which the DIALANG Yes/No content appears to differ from the content we have used so far (Section 8.1 and 8.2), two experiments will be described in which we have used the DIALANG Yes/No test content. Section 8.3 reports the first experiment (Experiment 5) in which the Yes/No Vocabulary Test was infused with the DIALANG content for Dutch and administered to a French-speaking population. In Section 8.4, a second experiment (Experiment 6) checks the response behaviour of Dutch native speakers on the same material. Finally, it is concluded in Section 8.5 that the DIALANG test content provoked an even larger response bias in the French-speaking population than the material we had used before. Furthermore, native speakers seemed to experience difficulties in recognizing the word-items of the test, which seriously subverts the confidence that can be placed in the DIALANG Yes/No Vocabulary Test for Dutch.

8.1 Test content of the Yes/No Vocabulary Test: a string of decisions

When constructing a Yes/No test, several decisions have to be made, each of which could seriously influence the test's validity and the inferences that can be made on the basis of the test results. In this section, we will briefly consider the ways in which the DIALANG Yes/No test differs from the Yes/No tests we constructed and used in the previously described experiments. Therefore, we will have to consider corpus and sample size, the selection of words as well as

the construction of pseudowords and whether cognates should form part of the test material.

8.1.1 Corpus and sample size

At the onset of this thesis it was pointed out that it is not our purpose to measure global vocabulary size in the target language with the Yes/No Vocabulary Test. We restricted ourselves to measuring the participants' receptive knowledge of the Dutch basic vocabulary (up until 3750 most frequent and useful words according to the corpus we have been using) because we are concerned with assessing whether learners know the lexical items they need to meet their learning objectives once they enter the Dutch classes. According to Meara (1996) scholars have generally found that the Yes/No tests work best when the target vocabulary is fairly tightly defined (for instance: the second thousand most frequent words in English). With a test consisting of 100 items (60 words and 40 pseudowords) for a total of 3750 corpus words, the sample in the tests we have used relates to the totality of the corpus in a proportion of 1/62.

Most of the time, the Yes/No test is used to construct a profile of the testee's global vocabulary size. In the Eurocenters Vocabulary Size Test (Meara and Jones 1990), the testee is presented with a random sample of 20 words for each 1000-word frequency band. This boils down to a sampling proportion of 1/50. Actually, the Eurocentres Vocabulary Size Test does not measure total vocabulary size strictly speaking because it uses lists up to the 10,000 most frequent lemmas of English. But for most intermediate learners of English as a foreign language, there is probably little difference between what the test measures and their total vocabulary size (Nation 2001).

In the DIALANG system, the Yes/No test is also intended to measure the test takers' global vocabulary size. In the piloting tool the test consisted of 150 items. The final DIALANG Yes/No Vocabulary Test as it can be found on the internet today, presents the learner with 75 items, and this is the case for all fourteen target languages. Since there are no frequency lists available for all DIALANG languages, the words were selected from medium-sized bilingual dictionaries (Meara, personal communication).

8.1.2 Selection of target words

In the previously described experiments, random selections had been made within the corpus, irrespective of the grammatical category of the words. This had as a consequence that the samples contained numerals, prepositions or conjunctions, apart from nouns, verbs, adjectives and adverbs. In Experiment 4, the selection of target items was restricted to nouns and verbs because of the strong lexical relevance of both categories. In Meara's EFL Vocabulary Tests (1992) nouns, verbs, adjectives and adverbs make up the different Yes/No

tests. In the DIALANG Yes/No test the item selection is restricted to the grammatical category of verbs. A semi-random set of 1000 lexical verbs (not auxiliaries or modal verbs) was chosen and from these a sub-set of 100 verbs was selected at random (Alderson, personal communication). This was based on the assumption that the selection of 100 verbs would reflect the overall frequency of the verbs in the language (Meara, personal communication). The verbs are all presented in their base form (infinitive). For future test development one might consider selecting words from a particular grammatical category according to the proportion in which this category is represented in the corpus. We will return to this issue in Chapter 9.

8.1.3 Construction of pseudowords

As we have already remarked in Chapter 2, there are no guidelines for the construction of pseudowords. There seems to be a general consensus among scholars that the pseudowords should "share the physical characteristics of the real words (…)" (Meara and Jones 1988: 85) but the extent to which they may differ remains vague. For the construction of the pseudowords in the described experiments we have applied the two principles advocated by Anderson and Freebody (1983) that consist in changing one or two letters in an existing word or altering the stem-affix combination of a word. In the EFL Vocabulary Tests (Meara 1992), the pseudowords consist of syllables of words from the frequency range involved that are put together at random. The resulting pseudowords are judged on their consistency with the English phonological rules by native speakers. When one examines the pseudowords in the DIALANG test, they could have been constructed by either of the above-mentioned techniques. In fact, with reference to the selection or formation of pseudowords, Meara claims that "(…) the choice of non-words may be relatively unimportant within the overall framework of the test" (1990:110).

8.1.4 Inclusion or exclusion of cognates

On the presence of cognates (or pseudowords that contain a cognate-affix) in the test sample and its influence on the participants' responses to words as well as pseudowords, different hypotheses have been formulated. As was already exemplified in Section 3.5 of Chapter 3, it has been argued that cognates artificially enhance test scores as well as that they would depress the test score.

In the literature, ways have been proposed to counterbalance or exploit the cognate effect. Meara (1990) explains how pseudowords can be designed to look as though they are genuine cognates. He gives the example of the Spanish stem form "tarde" (which means late) and combines it with the affixes "re-" en "-imiento" to form the non-word retardemiento*. When a testee responds "Yes" to this item, his false alarm rate will go up but his real hit rate will be adjusted downwards, for testees are penalized for assuming they know what an

item like retardemiento* means. Therefore, Meara concludes that thanks to the presence of pseudowords, the Yes/No test reflects not only the passive knowledge of the test taker, but also "how confident the testee is about his ability to use the words he claims to know" (1990:109). In the Introduction of his book EFL Vocabulary Tests (1992), he argues that cognates are often low frequency words (he refers to cognate words that share the same Greek roots, in English for instance "exclude", "emancipate", etc.) and that one can get round the problem of cognates by including a proportion of imaginary words that are made up of Greek and Latin roots so that testees who are merely guessing the meanings of words based on similarities with their L1 will be corrected. He finds that this correction factor works reasonably well in that it may underestimate the passive vocabulary skills of Romance speakers, but seems to give a quite accurate measure of their active vocabulary knowledge.

In the DIALANG test we observe that,

1) there are no different Yes/No tests for people with a different language background. A native speaker of German will get the same set of words as a native speaker of French when he takes the DIALANG test for English.
2) cognates are not excluded in the DIALANG Yes/No material. The Dutch material for instance contains the words "examineren, camoufleren, hypnotiseren, detineren, stabiliseren, royeren" which can be considered cognates for people with a Romance language background (or with a proficient knowledge of a Romance language as second or third language). None of the pseudowords of the Dutch material resemble cognates, which means that there is no intent of the test developer to counterbalance the 'cognate effect', should it arise.

8.2 Why choose the DIALANG test content?

The reasons why it was considered expedient to infuse the Yes/No Vocabulary Test with the DIALANG test content for Dutch are obvious. The DIALANG project for the development of diagnostic language tests is a strongly subsidized and authoritative body which is carried out with the help of language experts from all over Europe (http://www.dialang.org/english/summary.html). It is predicted that the system will play a major role in language teaching institutions, as an instrument for placement purposes and for diagnosis of learning needs (Alderson and Banerjee 2001). The system uses the Yes/No test as a vocabulary measure and uses the score for two purposes:

to inform the test taker of his lexical ability in the target language
to select the appropriate language tests for this particular testee.

The tests cover all levels, from beginner to advanced, and the approach to assessment is learner-oriented in that it treats self-assessment or self-rating as an integral part of language ability (Luoma and Tarnanen 2003). The tests for each language are anchored in the same scales of proficiency levels and test

specifications and the DIALANG proficiency levels are based on the Council of Europe's scales, which are part of the Council's Common European Framework of reference. On top of that the DIALANG organisation has the means and resources to conduct an ongoing process of item calibration. Since Dutch is one of the fourteen target languages for which tests have been developed and will be used throughout the world, what better test content to use as a point of reference than this one?

At the time of the experiment the DIALANG system was not yet fully operational, which is why we could only use the DIALANG piloting tool for Dutch. Meanwhile the DIALANG team has processed the piloting data they have gathered and some adjustments have been made to the Yes/No Vocabulary Test before the DIALANG system was finalized and put on the world wide web. In the final DIALANG Yes/No Test as it is currently presented to test takers all over the world, the total number of test items was reduced from 150 in the piloting tool to 75 in the finalized tests. These 75 items were selected from the 150 piloting tool items and to our surprise a very rare word as "aanhitsen" and the pseudowords "vandagen" en "vandaagen[19]" (the presence of both items in the same test we thought a fluke) survived the item calibration.

Unlike the Eurocentres Vocabulary Size Test (Meara 1992), which is also a computerized Yes/No test, the DIALANG Yes/No test is not adaptive. If the testee reaches a criterion level of performance in the EVST, the program proceeds to the next level. If not, it is assumed that the testee has reached the upper limit of his vocabulary knowledge and a further set of 50 words from the same frequency level are presented in order to fine tune the learner's vocabulary size estimate. In the DIALANG system, there is only one set of 75 items available, which means that the same items are used for all testees (no matter what their L1 background might be) and if a testee wants to take the test a second time, he is offered the same material as before, in exactly the same order as before. This scant amount of material might explain why it is not turned into a computer adaptive format. In the next section we will report how French-speaking learners of Dutch performed on a Yes/No Test that was infused with the Dialang Yes/No test content.

8.3 Experiment 5: DIALANG with French-speaking participants

8.3.1 Aim

This experiment was set up with the primary aim to answer the question whether the test content we devised and used in the previous experiment was

[19] It has to be remarked that "vandaagen" does not follow the rules for pseudowordformation because it violates Dutch orthography. In Dutch, vowels in open syllables are written with only one grapheme (in this case: with only one "a").

responsible for the participants' response behaviour. In this experiment, the Yes/No Vocabulary Test was infused with the DIALANG content for Dutch and administered to a French-speaking population in order to see if the gathered data would show a decline of the false alarm rate and of the response bias.

8.3.2 Method

In this experiment the DIALANG Yes/No test for Dutch was mimicked in all respects: exactly the same words and pseudowords were used and the proportion in which they occurred (100 words and 50 pseudowords) was respected (see Appendix 5). We also used the same instruction, which read for the French-speaking testees: *"Dans la liste des mots ci-dessous, indiquez les mots qui existent réellement et ceux qui ont été inventés en choisissant "Oui" ou "Non". Tous les "mots" sont des verbes, par exemple, 'parler', 'courir', 'manger', etc. Repondez à toutes les questions."* (Distinguish the words that really exist from those that have been invented by choosing "Yes" or "No" in the list of words below. All word are verbs, for example: 'speak', 'run', 'eat', etc.) With this instruction the DIALANG test developers have replaced the notion of judging a word to be known (which involves clear self-assessment) by judging if a word exists in the target language, which changes the test task fundamentally. In accordance with the DIALANG test, a time limit was not imposed. The experiment was carried out on paper because the Yes/No Test was part of the placement test procedure, which involved assigning 450 students to the appropriate courses and we did not have sufficient computers at our disposal.

Since DIALANG presents the Yes/No items in a list on the screen which testees can scroll up and down and since it allows for altering responses as many times as the testee wishes, the paper-and-pencil test can be considered to be quite similar to the DIALANG computer test. Only, omitted responses are not possible in the DIALANG test (you cannot arrive at a test result unless you have responded to all the items), whereas they remain a possibility in a paper-and-pencil test, even when you strongly advise against it in your instructions.

The participants were also administered the multiple choice grammar test since this test has a central role in the placement procedure at the language institute. Thanks to years of calibration it has turned into a highly reliable instrument. There was no time left to add a Translation task to the testing procedure. We did not want to wear the students out. However, the results on the grammar MC will serve as an indirect form of external validation when we consider the results on the Yes/No Vocabulary Test.

8.3.3 Results

By the time the results of this experiment were analyzed, the DIALANG test battery had been put into operation on the internet[20]. As we have mentioned earlier, the final set of items in the DIALANG Yes/No Vocabulary Test was reduced from 150 to 75 items (50 words and 25 pseudowords). When discussing the results of the described experiment, we have systematically calculated the values for the final set of 75 items in order to verify what the effect of the calibrated test material could have been on the qualities of the test.

The reliability values for the words and the pseudowords in the test (.842 for words and .836 for pseudowords) were higher than in previous experiments, which was expected since there were considerably more items in the test (Table 8.1). In order to weigh the reliability of the 50 pseudowords in relation to the reliability of the 100 words, a split half was calculated. This yielded an alpha of .726 for the pseudowords, which was well below the actual alpha of .837, in other words: the reliability of the pseudowords in the test exceeded the reliability of the words. Again, this finding ran counter to the presumptions that have been made about the role pseudowords should fulfill in this test format.

When the reliability was calculated for the entire set of items (words and pseudowords), it dropped to .720. This was already a strong indication of a possible negative correlation between the performance on the words and the performance on the pseudowords. The presence of this negative correlation was confirmed for all three orders of the test, as can be seen in the last column of Table 8.1.

*Table 8.1: Test reliability of word- and pseudoword-items of the Yes/No Vocabulary Test. A, B and C are the different orders in which the same test material was presented. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

150 items	Words /100			Pseudowords /50				Rel. /150	Correl. w/pw
	mean	SD	Rel.	mean	SD	%	Rel.		
A (N=151)	60.82	9.42	.819	33.38	6.62	66.76	.811	.695	-.396**
B (N=150)	61.41	10.92	.872	34.07	7.16	68.14	.847	.703	-.560**
C (N=149)	63.48	9.39	.823	34.05	7.62	68.10	.853	.751	-.310**
Total (N=450)	61.90	9.98	.842	33.98	7.13	67.96	.836	.720	-.422**

[20] Within the Dialang system the scores are calculated with the Δ_m algorithm. However, the scoring of the test is currently being investigated (Alderson, personal communication).

In Table 8.2 the results are presented for the final set of 75 items that make up the DIALANG Yes/No Vocabulary Test today. The reliability for the retained words of the final set amounted to .733, whereas the rejected word-items attained a reliability of .738. According to these data, it could not be concluded that the "best" word-items have been withheld in the final set. For the pseudowords one would expect to find a reliability in the final set that is well below that of the rejected pseudo items. This was not the case: the reliability of the rejected pseudowords came to .683 while the reliability of the retained pseudowords of the final set added up to .757. Again, the decision to favour certain items and leave out others did not coincide with the findings we obtained with our French-speaking participants.

When the reliability of the entire set of items was calculated (words and pseudowords), a severe drop could be noticed (from .733 for the word-items and .757 for the pseudowords-items to .467 for the 75 items as a whole). With the exception of Test Order C, the correlations between the performance on words and the performance on pseudowords were negative, but these negative values were less outspoken than those in Table 8.1. It could be that the negative correlations were somewhat suppressed by the rather low reliability.

*Table 8.2: Simulated test reliability of word- and pseudoword-items of the Yes/No Vocabulary Test for 75 items of the final DIALANG set. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

DIALANG Final set	Words /50				Pseudowords /25				Rel. /75	Correl. w/pw
(75 items)	mean	SD	%	Rel.	mean	SD	%	Rel.		
A (n=151)	29.29	5.19	58.58	.705	16.42	3.90	65.68	.739	.422	-.136
B (n=150)	29.56	5.68	59.12	.765	17.34	3.82	69.36	.744	.458	-.350**
C (n=149)	30.54	5.35	61.08	.728	16.69	4.40	66.76	.788	.494	.027
Total (n=450)	29.80	5.43	59.60	.733	16.82	4.05	67.28	.757	.467	-.148

The most important result of this experiment concerned the false alarm rate displayed by the participants. As can be seen in Figure 8.1 the false alarm rates reached 32.0% for the pilot set and 32.7% for the final set of items. The difference between both rates was not significant (ANOVA, one way, F=.053, df= 2,499, p= .948) and it showed that the false alarm rate did not appear to decrease with the "improved" DIALANG material. Both false alarm rates were extremely high (the highest values we have obtained throughout the use we have made of the Yes/No Test). The fact that on average every participant accepted one out of three pseudowords put the validity of the test under question once more. The DIALANG test content did not solve the validity problem of the test format. If anything, it rendered the data even more unreliable. It can also be remarked that the new instruction that replaced the

rather ambiguous "*Do you know the meaning of the word*" by "*Does the word exist in Dutch*" did not solve the format's questionable validity.

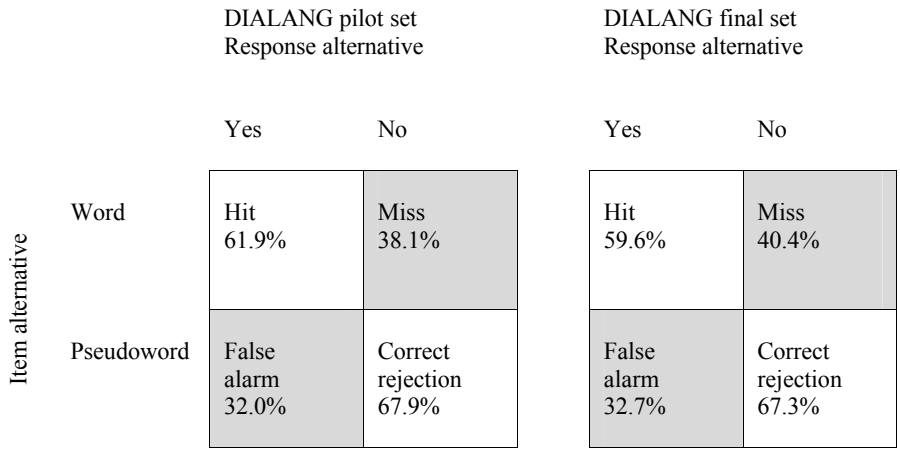


Figure 8.1: Item-response matrix for the DIALANG pilot test (150 items) and a simulated matrix for the 75 retained items that make up the DIALANG Yes/No Vocabulary Test today.

Item analysis revealed that the tendency to accept pseudowords as existing Dutch words ran through the entire pseudoword sample. Table 8.3 shows the number of times a pseudoword received a "Yes"-response for the seven items that were misjudged by more than half of the participants.

Table 8.3: Item analysis of the pseudowords that appeared most attractive for the 450 participants.

Pseudoword	Total number of Yes-responses (n= 450)	%
inzoeken	391	86.50
achterslaan	371	82.08
afbreden	312	69.03
stremen	265	58.63
sloeten	244	53.98
ontlonen	238	52.65
verhekken	226	50.00

When we considered the remaining 25 pseudowords in the final sample of the DIALANG Yes/No Vocabulary Test, it appeared that, with the exception of "achterslaan" and "ontlonen", these extreme distracters survived the item calibration.

Item analysis of the words disclosed that about one third of the word-items of the DIALANG piloting tool sample were not recognized as words by more than 75% of the participants. The most difficult word-items seemed to be "gadeslaan" [to observe], "royeren" [to expel], "hamsteren" [to hoard], "beamen" [to endorse], "prakkezeren" [to muse], "kenschetsen" [to characterize], "detineren" [to detain], "cementen" [to cement], avanceren" [to advance] and "brabbelen" [to babble]. Of these, "gadeslaan", "prakkezeren", "avanceren" and "brabbelen" were rejected in the final word set of the DIALANG Yes/No Vocabulary Test.

The test scores were very low (Table 8.4). This was not surprising given the high false alarm rate. Furthermore, the DIALANG material was not selected from high frequency bands whereas the Yes/No Tests in the previous experiments only used words from the Dutch core vocabulary. The fact that the DIALANG material only consisted of verbs, which might render the test task more univocal for the participants, did not appear to counterbalance the degree of difficulty of the material. The standard deviations were large, which was to be expected in a placement context where the population is usually characterized by extremely varied language levels.

Table 8.4: Results on the DIALANG pilot set with the different methods of scoring. The raw score is the number of hits and the corrected scores are based either on the all-or-nothing model (cfg) or on the continuous model (I_{SDT} and Hcfb).

DIALANG Pilot set 150 items	Formula	Test score /100 mean	SD	%
A (n=151)	Raw (hits)	60.82	9.42	60.82
	cfg	40.63	16.98	40.63
	I_{SDT}	29.50	13.42	29.50
	Hcfb	29.91	14.05	29.91
B (n=150)	Raw (hits)	61.41	10.95	61.41
	cfg	42.56	14.38	42.56
	I_{SDT}	30.95	12.72	30.95
	Hcfb	31.45	13.52	31.45
C (n=149)	Raw (hits)	63.48	9.34	63.48
	cfg	44.49	17.65	44.49
	I_{SDT}	32.90	15.58	32.90
	Hcfb	33.51	16.44	33.51
Total (n=450)	Raw (hits)	61.90	9.98	61.90
	cfg	42.55	16.44	42.55
	I_{SDT}	31.11	13.99	31.11
	Hcfb	31.61	14.76	31.61

When we calculated the test scores for the final set of items of the current DIALANG Yes/No Vocabulary Test (Table 8.5), the decrease in the scores was blatant (from 61.90% [raw], 42.55% [cfg], 31.11% [I_{SDT}], 31.61% [Hcfb] for

the pilot set to 39.71% [raw], 24.29% [cfg], 18.71% [I_{SDT}], 19.19% [Hcfb] for the final set).

Table 8.5: Calculated results for the 75 items of the final DIALANG test with the different methods of scoring.

DIALANG 75 items	Formula	Test score /75 mean	SD	%
A (n=151)	Raw (hits)	29.29	5.19	41.84
	cfg	16.58	12.34	22.11
	I_{SDT}	12.60	9.14	16.80
	Hcfb	12.87	9.63	17.16
B (n=150)	Raw (hits)	29.58	5.70	39.44
	cfg	19.59	9.72	26.12
	I_{SDT}	15.00	8.18	20.00
	Hcfb	15.41	8.89	20.55
C (n=149)	Raw (hits)	30.48	5.32	40.64
	cfg	18.50	12.78	24.67
	I_{SDT}	14.50	10.60	19.33
	Hcfb	14.90	11.24	19.87
Total (n=450)	Raw (hits)	29.78	5.42	39.71
	cfg	18.22	11.73	24.29
	I_{SDT}	14.03	9.39	18.71
	Hcfb	14.39	10.00	19.19

Table 8.6: Calculated results for the 75 items that were not retained in the final DIALANG Test.

Rejected items	Formula	Test score /75 mean	SD	%
A (n=151)	Raw (hits)	31.53	5.19	42.04
	cfg	22.99	8.27	30.65
	I_{SDT}	16.96	6.54	22.61
	Hcfb	17.21	6.79	22.95
B (n=150)	Raw (hits)	31.86	5.97	42.48
	cfg	22.50	8.12	30.00
	I_{SDT}	16.08	6.57	22.44
	Hcfb	16.36	6.89	21.81
C (n=149)	Raw (hits)	33.05	5.17	44.67
	cfg	24.64	10.62	32.85
	I_{SDT}	18.38	7.53	24.51
	Hcfb	18.74	7.86	24.99
Total (n=450)	Raw (hits)	32.14	5.48	42.85
	cfg	23.37	9.10	31.16
	I_{SDT}	17.14	6.94	22.85
	Hcfb	17.43	7.24	23.24

In Table 8.6, a simulation of the mean scores and standard deviations is presented for the rejected material. When we compared the retained material of the final set with the rejected material, it was noticed that the scores for the retained set of items were lower (39.71% [raw], 24.29% [cfg], 18.71% [I_{SDT}], 19.19% [Hcfb] for the retained set versus 42.85% [raw], 31.16% [cfg], 22.85% [I_{SDT}], 23.24% [Hcfb] for the rejected set). On the basis of these data, it seems that for the DIALANG final set the more difficult items have been selected.

The results on the Grammar MC were in line with those on the Yes/No Vocabulary Test: the corrected scores were low and the standard deviations large (Table 8.7).

Table 8.7: Descriptive statistics of the available results on the Grammar MC. The scores are calculated with the classic correction for blind guessing (cfbg) and the reliability is calculated with Cronbach's alpha.

MC item order	MC 78 items (4 alternatives)	Mean /100	[SD]	Reliability
Order 1 (N=151)	Raw	30.89	11.70	.886
	cfbg	16.01	15.00	
	%	20.52	19.23	
Order 2 (N=150)	Raw	32.05	13.56	.914
	cfbg	17.97	16.96	
	%	23.04	21.76	
Order3 (N=149)	Raw	32.11	13.84	.919
	cfbg	17.99	17.22	
	%	23.07	22.07	
Total (N=450)	Raw	31.68	13.05	
	cfbg	17.28	16.39	
	%	22.16	25.02	

Within the constraints of the placement context there was no time to administer a Translation task in order to validate the Yes/No responses. However, the results of the Grammar MC might serve as an indirect validation of the Yes/No Test, and in view of the initial aim to use the Yes/No Vocabulary Test as a complement to the grammar test in the placement procedure, it seemed interesting to examine how these measures coincided. In Table 8.8 the correlations between both measures are presented for the pilot test and the final DIALANG test. The correlations were rather low and the simulated final test scores correlated sightly better with the Grammar MC than the scores on the piloting tool. The formulae based on continuous modeling rendered the highest correlations (.599 [I_{SDT}] and .613 [Hcfb] for the piloting tool).

*Table 8.8: Correlation between the Yes/No Test scores and the results on the Grammar MC for the DIALANG pilot test (150 items) and the final DIALANG Yes/No Vocabulary Test (75 items). Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001). N=450*

Correction formula	Correlation Grammar MC / pilot test (150 items)	Correlation Grammar MC / final test (75 items)
Raw	.244**	.408**
cfg	.462**	.520**
I_{SDT}	.599**	.652**
Hcfb	.613**	.666**

Although the MC test is a measure of grammatical competence rather than of vocabulary knowledge, the modest correlations were discouraging since they raised the question as to what measure should gain precedence when assigning the participants to the appropriate classes? In Figure 8.2 a scatter plot illustrates the distribution of the scores the participants obtained on both tests.

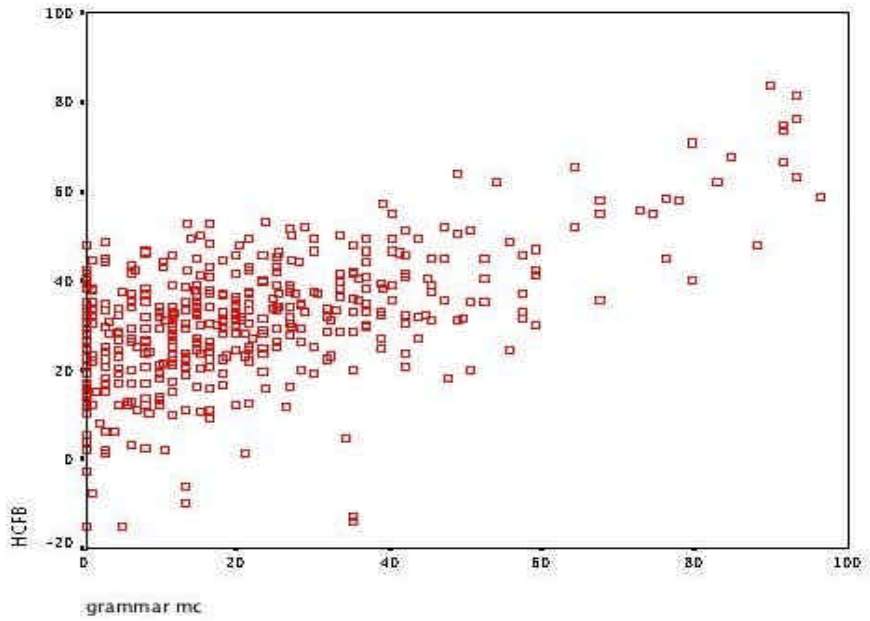


Figure 8.2 : Scatter plot of the participants' scores on the Grammar MC and the Yes/No Vocabulary Test (scores corrected according to the Hcfb-formula).

It can be observed that participants who obtained the same score on the MC, had widely diverging Yes/No scores. The reverse was also true: participants that scored between 20 and 40 on the Yes/No test had quite diverse results on

the MC. The scatter plot elucidated that the decision to place the participants according to the Grammar MC or according to the results on the Yes/No Test, would have resulted in quite a different constitution of the language classes. It is clear that within the placement test procedure, the results of the Yes/No Test cannot be considered supplementary to the results of the Grammar MC.[21]

Summarizing we can state that when we compare the data assembled with the DIALANG test content to the data we gathered with the content we constructed ourselves in the previous experiments, the qualities of the Yes/No Test have not improved under influence of the DIALANG test material. The false alarm rates were the highest we have encountered so far and constitute a clear sign of a validity problem. The negative correlation between the performance on word-items and the performance on the pseudoword-items was again indicative of a response bias in the data. The simulated results for the final set of DIALANG items did not resolve these issues.

8.4 Experiment 6: DIALANG with native speakers

8.4.1 Aim

This experiment was set up because we experienced a lot of difficulty in taking the DIALANG Yes/No test ourselves. We did not succeed in obtaining a perfect or nearly perfect score because we could not always distinguish words from pseudowords in the DIALANG material. We wondered if the test would prove equally challenging for fellow-native speakers. Therefore, a small-scale experiment was set up in which the DIALANG Yes/No Test was given to Dutch-speaking university students.

8.4.2 Method

The DIALANG Yes/No Test was administered to 70 Dutch speaking university students in the first year of their study of Germanic Languages at the Vrije Universiteit Brussel. Surely, one is justified to expect a near perfect score from native speakers when it comes to recognizing words as belonging to their native tongue. Furthermore, given their choice of study these students' aptitude in their own native language should be excellent. The test was exactly the same as in the experiment described above, except that the instruction was translated into Dutch. The students were given the test during a linguistics course and they were told it was part of an experiment that aimed to validate a particular test format. They completed the task in a few minutes.

---

[21] Since the Grammar MC had proven to be a valid placement indicator in the past and since the high false alarm rate of the Yes/No Vocabulary Test has not inspired much confidence, it was decided to place the participants according to the results on the Grammar MC.

8.4.3 Results

The most striking result about the native speakers' performance on the DIALANG material was the poor score on the word-items (mean score of 82.79 on 100 word-items). The fact that the Dutch-speaking students of linguistics did not perform too well confirmed the qualms we had about the selection of words and pseudowords in the DIALANG piloting tool. When the scores that were obtained on the word-items were compared to those compared on the pseudoword-items (Table 8.9), it was observed that the pseudoword-items caused fewer problems for the native speakers than the word-items.

At .750 for the words and .555 for the pseudowords, the reliabilities were higher than we had expected. There was a weak negative correlation between the performance on the words and the performance on the pseudowords.

*Table 8.9: Descriptive statistics for the word- and pseudoword-items of the DIALANG Yes/No Piloting tool. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

150 items	Words /100				Pseudowords /50				Rel. /150	Correl. w/pw
	mean	SD	%	Rel.	mean	SD	%	Rel.		
n=70	82.79	5.21	82.79	.750	47.91	1.98	95.82	.555	.661	-.264*

When the descriptive statistics were simulated for the 75 items that belong to the final set of the DIALANG Yes/No Vocabulary Test (Table 8.10), it was observed that although the scores on the word- and pseudoword-items improved, the score for the words was still by no means what would be expected of native speakers. The final set of items was qualitatively better than the piloting tool, which could also be concluded from the decreasing reliabilities for words and pseudowords and the zero correlation between the performance on words and the performance on pseudowords. The alphas for the words and pseudowords that were rejected from the final set, were .614 (words) and .538 (pseudowords). They were higher than the alphas of the retained words and pseudowords, as can be seen in Table 8.10. On the basis of these findings, it was clear that the material of the final set was an improvement to the material of the piloting tool, but still far from satisfactory.

*Table 8.10: Simulation of the descriptive statistics for the word- and pseudoword-items of the final set of items. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

75 items	Words /50				Pseudowords /25				Rel. /75	Correl. w/pw
	mean	SD	%	Rel.	mean	SD	%	Rel.		
n=70	43.41	2.60	86.82	.549	24.50	.86	98	.365	.491	-.029

The matrix in Figure 8.3 confirmed the above-mentioned results: the hit rate increased in the final item set and the false alarm rate decreased. Still, 13.2% misses on the word-items is too much for a Dutch speaking population, especially when one takes into consideration that they were studying languages and could be considered language specialists.

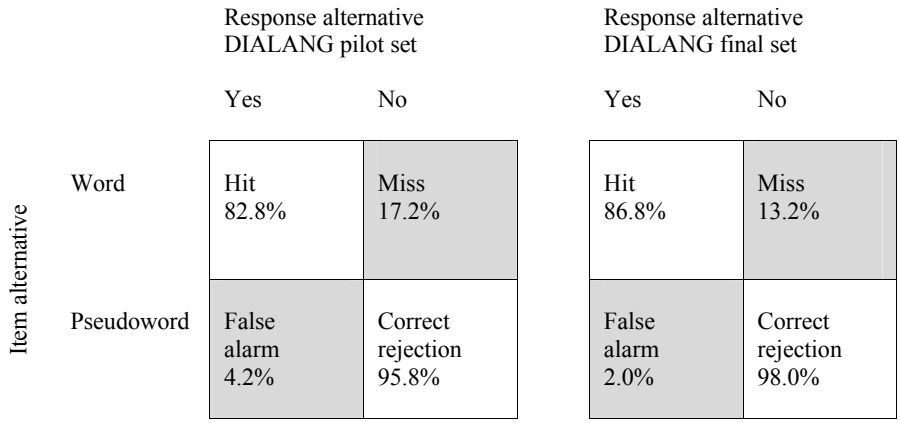


Figure 8.3 : Item-response matrix for the DIALANG piloting tool (150 items) and a simulated matrix for the 75 retained items of the final set.

When the mean score was corrected (Table 8.11), it was obvious that the corrected scores were almost identical to the raw score. The low mean score was not to be attributed to the performance on the pseudowords, it was the consequence of a poor hit rate.

An analysis of the inadequate performance on the word-items of the test is illustrated in Figure 8.4. The grey line illustrates the perfect performance one would expect of native speakers. The triangles show how the native speakers' reactions to the pseudowords were not without fault, there were seven items that seemed to be quite difficult to reject for a Dutch speaker. The deflecting line of circles indicates that the performance on word-items was much worse. More than half of the word-items presented a problem when it

came to recognizing them as existing Dutch words. Among those, there were about 10 word-items that were misjudged by half of the population. An item analysis was performed in order to provide further insight into these items.

Table 8.11: Results on the Yes/No Vocabulary Tests with the different methods of scoring. The raw score is the number of hits and the corrected scores are based either on the all-or-nothing model (cfg) or on the continuous model (I_{SDT} and Hcfb).

N=70	Formula	Test score /100		
		mean	SD	%
Pilot test	Raw (hits)	82.79	5.21	82.79
(150 items)	cfg	82.02	5.34	82.02
	I_{SDT}	79.27	5.14	79.27
	Hcfb	82.97	6.20	82.97

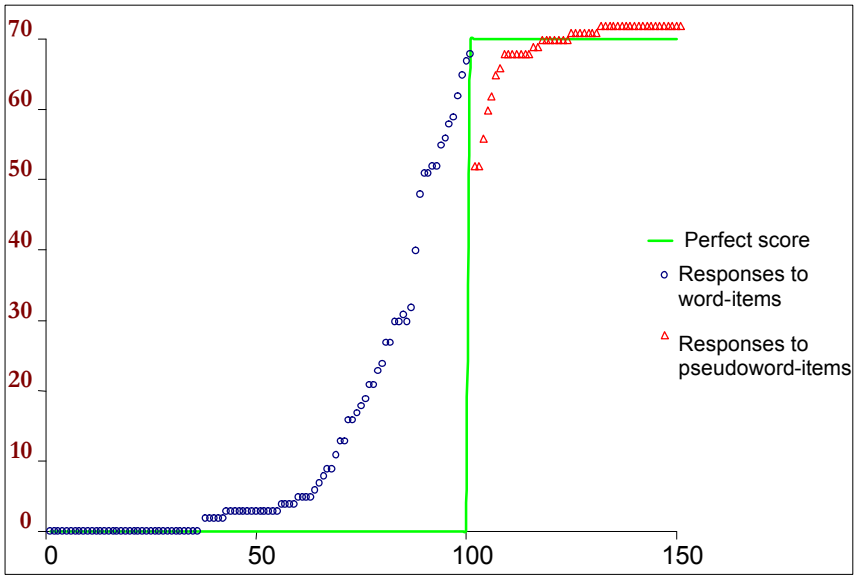


Figure 8.4 : Native speakers' performance (N=70) on the word-items and the pseudoword-items of the DIALANG piloting tool (100 words and 50 pseudowords).

As can be seen in the item analysis in Table 8.12, there were word-items that received a "No" response by more than 90% of the Dutch speaking participants. As said before, we have to admit that we ourselves experienced a lot of difficulty in taking the test. It took us several try outs on the piloting tool to arrive at a perfect test score. Therefore, we were not surprised at the native speakers' responses and for some of these items we heartily admit that we

rejected them ourselves. Not only did we consider some of these words to be quite useless to learners of the Dutch language (e.g. *schoren* [to shore up], *lenzen* [to empty], *gorden* [to gird]), some of them were downright farfetched and hardly ever encountered in contemporary Dutch (e.g. *belommeren* [to give shade], *prakkezeren* [to muse], *aanhitsen* [to incite]). The final set of the DIALANG Yes/No Vocabulary Test, has got rid of most of the word-items of Table 8.12. Only "*aanhitsen* [to incite]", "*verhaasten*" [to speed up] and "*royeren*" [to expel] have survived. The material has become much more univocal from a native speaker point of view which could indicate that the item calibration might have been executed on the basis of native speakers' reactions to the material.

Table 8.12: Item analysis of the word-items that were rejected by more than half the population.

Words	Rejections out of 70	%
schoren [to shore up]	66	94.29
lenzen [to empty]	65	92.86
gorden [to gird]	63	90.00
belommeren [to give shade]	60	85.71
prakkezeren [to muse]	57	81.43
aanhitsen [to incite]	56	80.00
verhaasten [to speed up]	54	77.14
cementen [to cement]	53	75.71
bijmengen [to mix in]	49	70.00
royeren [to expel]	47	67.14
koteren [to pick]	38	54.29

Although the pseudowords did not seem a source of problems for native speakers, an item analysis was performed to identify the most "attractive" pseudowords (Table 8.13). More than one out of four native speakers judged "*veradelen*" and "*kwaadstoken*" to be existing Dutch words. The four pseudowords of Table 8.13 have not been retained in the final set of the DIALANG Yes/No Vocabulary Test.

Table 8.13: Item analysis of the pseudoword-items that were rejected by more than 10 % of the population.

Pseudowords	Number of "Yes" responses (n=70)	%
veradelen	20	28.57
kwaadstoken	20	28.57
achterslaan	16	22.86
winken	12	17.14

When we calculated the native speakers' test scores on the 75 items that were retained in the DIALANG final set and we compared them to those 75 items of the piloting tool that were rejected, we observed that - in contrast with the French-speaking population - the native speakers obtained a better score on the

final set of items, which again proved that the material has become more univocal for a native speaker.

Table 8.14: Calculation of the native speakers' results on the retained and rejected items of the DIALANG piloting tool. The raw score is the number of hits and the corrected scores are based either on the all-or-nothing model (cfg) or on the continuous model (I_{SDT} and Hcfb).

N=70	Formula	Test score /50		
		mean	SD	%
Final set	Raw (hits)	43.41	2.60	86.82
	cfg	43.23	2.67	86.46
	I_{SDT}	42.40	2.80	84.80
	Hcfb	44.29	2.86	88.58
Rejected set	Raw (hits)	39.37	3.12	78.74
	cfg	38.62	3.28	77.24
	I_{SDT}	36.69	3.78	73.38
	Hcfb	38.76	4.82	77.52

8.5 Conclusion

The DIALANG Yes/No Vocabulary Test for Dutch did not produce more reliable or valid results than the Yes/No Tests we had used in previous experiments. The main problems that have been pointed out throughout this thesis cropped up once again. The data revealed a high false alarm rate, a negative correlation between the performance on words and the performance on pseudowords and a discouraging correlation between the Yes/No Test scores and the results on the Grammar MC. Although it is indicated on the DIALANG website that item calibration is an ongoing process, the final set of items that make up the DIALANG Yes/No Test today are not much of an improvement. Furthermore, it has to be remarked that item calibration runs counter to some of the main advantages of the Yes/No format, that is, its user-friendliness and its universal applicability.

The experiment in which the DIALANG Yes/No Test was administered to Dutch native speakers showed that the material for Dutch is questionable. The native speakers' performance on the words and the pseudowords was far from perfect. Meara (1988) reports that some of the pseudowords in his experiments caused native speakers of English "to puzzle for a long time" (1988:86). In our case however, recognizing the words turned out to be more problematic than rejecting the pseudowords. Item analysis confirmed that there were a number of extremely far-fetched words in the DIALANG selection, most of which - but not all - have been removed from the final DIALANG Yes/No Vocabulary Test. Selecting words through dictionary sampling clearly does not constitute a reliable or valid test content for the Yes/No Vocabulary Test. It generates problems because the word

selection is bound to comprise far-fetched words or even regional variants, that do not get recognized as words by native speakers themselves.

Despite all our efforts to adapt the Yes/No Vocabulary Test for use with a French-speaking population, we have not succeeded in turning it into a sufficiently reliable or valid instrument. The tests have consistently yielded high false alarm rates and substantial response biases, which made the issue of choosing the appropriate correction formula very complex. The presence of a high false alarm rate endangers the overall validity of the test since finding a solution for the statistical treatment of the false alarm responses remains problematic to this day (Beeckmans et al. 2001). Time and time again, we have been confronted with a low concurrent validity, as shown by the weak correlations between the Yes/No Vocabulary Tests and translation tasks, indicating the lack of confidence that could be placed in the "hit"-responses. The fact that decision making is so central to the Yes/No task blurs the proficiency we aim to measure. The decision making process is influenced by the cognitive, psychological, social and cultural make-up of the participants. The experiments have shown that neither the instruction, the computer design nor the DIALANG test content could eliminate the response bias that is introduced by the format's Yes/No dichotomy. In the next chapter, we will therefore present and investigate an alternative test format to measure learners' vocabulary size.

Chapter 9

The Recognition Based Vocabulary Test

In this Chapter a new vocabulary test is introduced that retains the general principles and attractive features of the Yes/No Vocabulary Test but is designed to circumvent the pitfalls we have reported throughout this study. The most problematic drawback is that the Yes/No task prompts the participants to exhibit a response bias whereas formats should not in themselves adversely affect performance (Weir 1993). It is clear that in our use of the Yes/No Vocabulary Test the measurement of the vocabulary trait is contaminated by the method employed and this explains the low concurrent validity we have always obtained. The correction formulae that are based on the continuous model seem to be able to extract the bias from the raw data but the results they produce, are not consistent. Therefore, the construction of a new test format seems imperative.

In Section 9.1, a new test format will be described in terms of how it is designed to sidestep the response bias problem we were confronted with when using the Yes/No Vocabulary Test. In Section 9.2, an experiment is reported that compares two variants of this new format with the Yes/No Vocabulary Test. Finally, Section 9.3 evaluates the merits and shortcomings of the new test.

9.1 A new test format

In the previously reported experiments, one particular correction formula performed better than another depending on the circumstances (characteristics of the population, high versus low stake test context, administration of the test under teacher supervision or not, etc). This means that none of the proposed correction formulae are robust with regard to the different variables that play a part in the test. Time and again the test user has to determine which formula would provide the most reliable and representative test score in a given situation. This is unacceptable for a standardized format. On the basis of the assembled data we can therefore only conclude that the Yes/No Vocabulary Test is not so fit for use throughout different languages and consequently different language testing contexts.

When contemplating a new test for measuring vocabulary size, we sought to retain the advantages of the Yes/No Vocabulary Test (easy to construct, covering many words in little time, etc) but at the same time do without the Yes/No dichotomy that invokes response biases in participants.

Therefore it was decided to replace the detection task that is intrinsic to the Yes/No format by a discrimination task. In the new test format, the testees are presented with 60 pairs of words and pseudowords and they are asked to distinguish the existing word in the item-pair. The new test is called Recognition Based Vocabulary Test (RBVT) because the task the learner has to engage in is essentially one of recognizing the existing word in a pair. The format steers clear of the problems we encountered with the Yes/No Test for the following reasons:

1) Because false alarms are no longer possible in the new format, the test avoids all the problems concerning the correction formulae and the discussion of how to calculate the test score. One straightforward formula can be used to yield a univocal and representative score. The formula corrects for guessing and since we are now dealing with classical blind guessing, the probability of a correct guess is determined by the number of choices and not by the inclination of a participant to prefer the "Yes" or "No" response.

2) The task changes drastically. Although it is often stated in the literature that the Yes/No task is clear and not very demanding, the reported experiments have shown the task to be deceptive and open to interpretation. This is in fact an inherent characteristic of any yes/no decision and it is probably the primary cause of the response bias we encountered in the data of our experiments. The new task is stripped of the ambiguity of a yes/no decision because it replaces the detection task by a discrimination task. The yes/no decision had to be made in a "void", which caused the participants' self-assessment to become strongly influenced by meta-cognitive or sociocultural factors. As a consequence, the decision did not reflect the participant's lexical knowledge nor was it the result of a blind guess. In the new test a word opposes a pseudoword in every item, therefore both parts of an item serve as each other's point of reference. The testee has to discriminate between them.

3) Response biases cannot intervene in the RBVT format. We have defined response bias as a tendency for a given participant to provide more/fewer responses of one type ("Yes") than of the other ("No"). In Chapter 5, we have provided an illustration of the response bias by means of continuous modeling. It showed that when stimuli appear on a learner's continuum of word knowledge, the learner places a Criterion on his internal scale of confidence and answers "Yes" to the stimuli that fall on the left side of the Criterion and "No" to the stimuli that fall on the right side of the continuum. The learner's decision rule is clearly determined by the location of the Criterion. When the Criterion is placed to the right side of the scale (versus the pole "is not a word") the learner will reveal a bias towards the "Yes"-response. When the Criterion is placed to the left (versus the pole "is a word"), the learner will reveal a bias towards the "No"-response. In the new test, the items consist of two stimuli and one of them has to be designated as the existing word. No matter where both stimuli would fall on the learner's internal scale of confidence if they were presented

separately, in the new test format they would in any case be positioned with reference to each other, in other words: one of them will be closer to the pole "this is a word" and subsequently that one will be indicated by the learner. When both stimuli overlap and the learner clearly cannot discriminate between them, he can only resort to a blind guess, which amounts to a 50% chance of getting the correct answer. The formula to correct for guessing is in this case the same as the one used with a True/False format. However, the formats are not the same: whereas a weak learner could for instance demonstrate a preference for the "False" response in the True/false format, exhibiting his lack of confidence in making judgements about the target language, this cannot happen in the new format we propose. The learner cannot demonstrate a preference for a particular response. The format does not allow it.

Two pilot studies were carried out to have a first evaluation of the new test format before submitting it to more thorough scrutiny in a controlled experimental set-up. In these pilot studies the new format was administered to first year French-speaking university students as part of the placement test for Dutch. As usual, the placement test consisted of a grammar MC. The new test was added to assess the students' knowledge of the basic Dutch vocabulary. As was expected, the calculation of the results did not present a problem, but the test as a whole appeared to be easier than the Yes/No test we had used before. This is not surprising given the fact that each item consists of two stimuli (word and pseudoword).

Informally we also obtained information about the surface credibility of the test. The testees found the new format much more appealing than the Yes/No test. They also believed the new format to do a better job at reflecting their vocabulary knowledge. In short, we can say that the new test seemed to have a higher face validity than the Yes/No test. Even though face validity is a controversial concept, generally refuted and severely criticized by language testers, we believe that it is an important factor for obtaining reliable and valid results. It is very important that test takers take a test seriously enough to try their best. Especially when you are dealing with adult learners, the factor whether test users find the test useful or not, cannot be ignored. Another advantage to be noted is that we were able to ascertain that native speakers obtain the maximum score on this new test without exception, which is not the case for the Yes/No test (see Chapter 8).

In its original form (the format we used in the pilot studies), the items of the RBVT consisted of words and pseudowords that were randomly paired (see Appendix 6). In the experiment that will be described in Section 9.3, a slightly different version of the new test will also be tried out. In this test format, which we will call Recognition Based Vocabulary Test II (RBVT 2), the pseudowords are paired up with the words of which they were derived (see Appendix 7). We hypothesize that the use of these minimal pairs could possibly reduce extraneous variables when the learners have to identify the existing

word. A minimal pair like, for instance, "believe – ralieve", could be more univocal to judge than the randomly combined pair "believe – traduce". In the second pair, the familiar morphological form of the pseudoword could prove to be too attractive to the learner. This is not the case in the first pair, where the morphological forms of word and pseudoword are identical except for the first syllable.

9.2 Experiment 7

9.2.1 Aim

The aim of the experiment was to evaluate the two new test formats (hence RBVT 1 and RBVT 2), to compare their results with the Yes/No Vocabulary Test, and to validate them by means of a translation task. The three formats are (1) a Yes/No Vocabulary Test consisting of 60 words and 40 pseudowords; (2) a Recognition Based Vocabulary Test (RBVT 1) consisting of 60 random pairs (each pair is made up of a word and a pseudoword that were arbitrarily put together); (3) a Recognition Based Vocabulary Test (RBVT 2) consisting of 60 minimal pairs (each pair is made up of a pseudoword and the word of which it is derived).

As before, we will consider the translation task as the most straightforward way of verifying the recognition of L2 words. In this context, recognition means that the participant is able to produce one possible meaning of the target word in his or her L1.

The central questions of the experiment are:

1) Will the two new formats (RBVT 1 and RBVT 2) reflect the participants' vocabulary knowledge more accurately than the Yes/No Vocabulary Test? When this question is made operational in the experiment, it gets transformed into: which of the three tests will allow the construct "receptive vocabulary knowledge" as measured by them, to be the same as when we ask the participants to translate the same words into L1? The test format that obtains the strongest evidence of concurrent validity with the translation task, will therefore be considered as the most representative measure of receptive vocabulary knowledge.

2) What is the average time needed to fulfil the tasks presented by the three tests? The average time per test (irrespective of whether the test items consist of one or of two stimuli) will inform us whether the Yes/No decision takes longer than the RBVT 1 or RBVT 2 choice. Given that the relative swiftness of a language test is an important factor in selecting it for a placement test procedure (where practical circumstances such as the number of participants, lack of time and supervisors, etc often compel the test administrators to choose discrete language measures), the amount of time requested for completing a test is a factor of importance.

9.2.2 Method

Material

For this experiment new material was assembled and constructed (see Appendix 8a, 8b, 8c). The criteria for word selection were revised, the word frequencies were adapted to the supposed proficiency of the population, and the rules for creating pseudowords were altered.

a) Word selection
It was decided to continue selecting the words and pseudowords from the corpus of Dutch basic vocabulary that we have been using so far (Woorden in Context I, Dieltjens et al., 1995, Woorden in Context II, Dieltjens et al., 1997). This is a manual with which the students are well-acquainted because they are advised to brush up their vocabulary knowledge by means of it. Because the experiment was going to be conducted with first level students, who should be more proficient in Dutch than the first year students that take the placement test before they enter the Dutch courses, the selection of words and pseudowords was made exclusively from the second manual, Woorden in Context II, which adds about 1700 words to the most frequent and useful 2000 words from the first book (Woorden in Context I).

In keeping with the DIALANG system of restricting the selection of words to a particular grammatical category (in the case of DIALANG, only verbs), our selection consisted of nouns and verbs. Both categories were considered best suited for use in a Yes/No or RBVT context because it seems easier to retrieve the meaning of nouns and verbs from memory when encountered in isolation than the meaning of, for instance, conjunctions or prepositions. For the same reasons we considered them easier to translate. In an attempt to make the sample as representative as possible of the corpus, the number of nouns and verbs in the test were selected according to the proportion in which they figure in the corpus. A classification of the corpus revealed the following proportions: Woorden in Context I contains 2016 items, of which 1066 nouns and 874 verbs. Woorden in Context II contains 1683 items, of which 841 nouns and 462 verbs. Globally, we can conclude that the corpus contains twice as many nouns as verbs. Therefore, we saw to it that the test sample contained 1/3 verbs and 2/3 nouns, and this proportion was respected for the words as well as the pseudowords. It was also decided to respect the presence of cognates in the corpus and consequently in the test sample. In view of the universal properties of a test format, the exclusion of words because they share certain characteristics with words of a particular L1 or group of L1's can hardly be defended.

b) Pseudoword formation

The formation of pseudowords was slightly altered in comparison with previous experiments because the formation principle that consists in changing the affixes of existing words was abandoned. In Dutch, like in most Indo-European languages, words tend to be made up of a relatively stable root, and a system of affixes that are added on to this stem (Ryan 1997). The morphological system of the language is intrinsically dynamic in the sense that it consists of a finite set of (phonological, morphological and syntactical) rules with which an infinite number of words can be formed. As a result, the borderline between existing words and potential words is hazy and although a lot of the potentiality of the morphological structures of Dutch is not used in every day life, a native speaker can choose to make use of them when it serves his or her communication needs.

These operations of forming new words or language units do not happen arbitrarily of course, hence the important notion of productivity. It refers to the opportunity of creating new words by means of a certain language intrinsic word formation principle. Affixation is such a language intrinsic formation principle and it is a very productive process in Dutch, which is why learners of Dutch are often stimulated to study word-building processes and to infer the meaning of words by deconstructing them. With reference to the use of pseudowords in the vocabulary tests under question, it was decided that it is unfitting to ask students to reject pseudowords that are formed through affixation when, at the same time, these students are encouraged during their language courses to discern the meaning of words by making inferences about a word's prefix, suffix or stem. Moreover, experience with the correction formulae of the Yes/No Vocabulary Test has taught us that making word recognition decisions based on partial knowledge gets rebuked heavily in the calculation of a Yes/No test score (because false alarms reduce the test score severely). For these reasons we have opted to exclusively apply the "substitution principle" for the formation of pseudowords in this experiment: words are transformed into pseudowords by changing one to four consecutive letters, depending on the length of the word in question. Vowels are not altered into consonants or vice versa.

Participants

The participants were French-speaking university students of Economics and Business Administration taking Dutch first level courses. A total of 177 students took part in the experiment. They were told they were being tested on their knowledge of the core vocabulary of Dutch. To ensure a motivated participation they were informed that the test results would be shared with their respective teachers who would consequently advise them on how to fill the gaps in their vocabulary knowledge.

Design

The experimental design was set up around three test formats (Yes/No, RBVT 1and RBVT 2) and three different materials (X, Y and Z). Every test format was made into three versions (version with material X, version with material Y and version with material Z). This makes a total of nine different tests. Every participant was presented with a Yes/No test, a RBVT 1 and a RBVT 2, each containing different materials. Afterwards the participants took a translation test that consisted of 60 (existing) words: 20 from each material. They were asked to translate the Dutch words into their mother tongue (French). The translation task was the same for the entire population.

Because each material (X, Y, Z) had to figure in the three test formats, all the selected words had to be transformed into pseudowords (for the RBVT 2 consists of minimal pairs, which means that all the pseudowords had to be derived from the words with which they are paired up). This had as a consequence that:

(1) The formation of pseudowords was the same in the three test formats.

(2) Words that can be considered cognates were also changed into pseudowords (*financiën* [finances] - **finantaan*; *symbool* [symbol] - **symbaat*; *dirigeren* [conduct] - **dirivaren*; *departement* [department] - **minortement*; etc.). This could have the advantage for the Yes/No data that if cognates should elicit mock "hit"-responses among the words of the Yes/No test, this effect could be compensated by the presence of "pseudocognates" among the pseudowords.

(3) A ceiling effect might be imminent in the RBVT 1 and RBVT 2 results (as was the case in the pilot studies) because using the same material as in the Yes/No but presenting it in a discrimination task instead of a detection task might make the test much easier.

a) Construction of the Recognition Based Vocabulary Tests

In order to obtain 60 pairs of words and pseudowords for one test version, 120 nouns and verbs were randomly selected from the corpus (1/3rd verbs and 2/3rd nouns). This procedure was repeated for the Y- and Z-material. Half of the selected words were turned into pseudoword-items according to the "substitution principle" (see Table 9.1). For the RBVT 2, the pseudowords were matched with the words from which they had been derived so as to make up minimal pairs. For the RBVT 1, the pseudowords were randomly paired with the words.

Table 9.1: Construction of both Recognition Based Vocabulary Tests (RBVT 1 and RBVT 2).

120 items (for each material X, Y, Z)	1/3 verbs = 40	50% real verbs = 20
		50% pseudoverbs = 20
	2/3 nouns = 80	50% real nouns = 40
		50% pseudonouns = 40

b) Construction of the Yes/No Vocabulary Test

In conformity with the previous experiments, we opted for a Yes/No Vocabulary Test of 100 items, consisting of 60 words and 40 pseudowords. Therefore, 20 pseudowords had to be deleted from each material (X,Y,Z). The pseudowords were deleted at random but with respect to the proportional presence of nouns and verbs in the sample (1/3rd verbs and 2/3rd nouns).

Table 9.2: Construction of the Yes/No Vocabulary Test

100 items (for each material X, Y, Z)	1/3 verbs = 33	60% real verbs = 20
		40% pseudoverbs = 13
	2/3 nouns = 67	60% real nouns = 40
		40% pseudonouns = 27

c) Construction of the Translation Test

In the Translation Test the participants were presented with 20 existing words from the three test materials (X, Y and Z), which makes a total of 60 items. Again, the proportion of 1/3rd verbs versus 2/3rd nouns was respected. Apart from this restriction, the items were selected randomly.

Table 9.3 illustrates the proportions of word- versus pseudoword-items for all the formats that were administered in the course of the experiment.

Table 9.3: Proportions of word- and pseudoword-items for each test.

Test format	N° of words	N° of pseudowords
Yes/No Vocabulary Test	60	40
Recognition Based Voc. Test I	60	60
Recognition Based Voc. Test II	60	60
Translation Test	60	0

All tests were computerized because this allowed for an automatic scoring (with the exception of the Translation test for which a "human-assisted" computer

scoring was performed, see Section 6.2.3), and because it helped to control several variables: (1) for all three formats, the items were presented one at a time, which prevented the participants from skipping items or keeping track of their response behaviour; (2) the response time per item was automatically recorded for each of the tests, so that the average response time per test will be revealed in the data analysis; (3) it was not possible to omit responses, which turned the test into a forced decision task; (4) in order to control for sequencing effects the computers were programmed into the 36 possible combinations the order of the three formats allow; the order in which the items of the Yes/No test appeared, changed for every participant; and the word- and pseudoword-items that made up the pairs in the RBVT 1 and 2 swopped places (left-right, right-left) with every new participant.

There was no time limit imposed for any of the tested formats or for the translation. We assumed that the entire assignment should not take them longer than about 35 minutes. In reality, all participants performed the four tests in less than 25 minutes. For all the tests; we opted for instructions that avoided any mention of "knowing" or "knowing the meaning of" a word. We thought it more straightforward and less ambiguous to ask the participants if the presented items existed or not. Theoretically one could argue that given the fact that the pseudowords obey the phonological and morphological rules of Dutch, the only way to distinguish between a pseudoword and a word is to know whether it carries a meaning in the target language.
The instruction for the RBVTs read:

Parmi les paires d'items qui vont apparaître à l'écran, un et un seul mot existe en néerlandais. Cliquez sur le mot.
(The item-pairs that will be presented on the screen contain only one existing Dutch word. Click the word.)

The instruction of the Yes/No Vocabulary Test read:

Les items qui vont apparaître à l'écran sont
-soit des mots existants en néerlandais
-soit des mots qui n'existent pas en néerlandais
Cliquez sur le bouton JA si vous pensez que le mot existe en néerlandais.
Cliquez sur le bouton NEE si vous pensez que le mot n'existe pas en néerlandais.
(The items that will be presented on the screen are
-words that exist in the Dutch language
-words that do not exist in the Dutch language
Click the Yes-button if you think the word exists.
Click the No-button if you think the word does not exist.)

9.2.3 Results

In order to test the materials' equivalence an analysis of variance was performed for the Yes/No data. The factor "material" turned out to be significant for all correction formulae (cfg: F=7.016, df=2,174, p=.001, eta²=.075; I_{SDT}: F=4.526, df=2,174, p=.012, eta²=.049; Hcfb: F=4.673, df=2,174, p=.011, eta²=.051; raw: F=6.003, df=2,174, p=.003, eta²=.065). A posthoc analysis demonstrated that the Z-material was slightly easier than the X and Y materials in the case of the Yes/No Vocabulary Test. The same procedure was repeated for the RBVT 1 and RBVT 2 data. For both formats, the difference in materials was not significant (RBVT 1: F=.286, df=2,174, p=.752; RBVT 2: F=1.629, df=2,174, p=.199). Because the explained variance only occurred in the Yes/No test and because it was small, it was decided not to pay any further notice to it.

The raw scores of the Yes/No Vocabulary Test were rather high (See Table 9.4), but they got reduced severely when they were transformed into corrected scores. This reduction pointed to the presence of a substantial false alarm rate in the data (see Figure 9.2).

Table 9.4: Mean scores for the Yes/No Vocabulary Test according to the different correction formula.

Yes/No Test	Formula	Mean /60	SD	%
Mat X (N=57)	Raw	43.90	4.74	73.17
	Cfg	37.36	6.62	62.27
	I_{SDT}	27.56	7.36	45.93
	Hcfb	23.18	6.32	38.63
Mat Y (N=60)	Raw	43.02	4.85	71.70
	Cfg	36.02	7.59	60.03
	I_{SDT}	27.08	8.26	45.13
	Hcfb	22.77	7.00	37.95
Mat Z (N=60)	Raw	46.02	5.01	76.70
	Cfg	40.78	7.24	67.97
	I_{SDT}	31.13	8.42	51.88
	Hcfb	26.29	7.24	43.82
Total (N= 177)	Raw	44.32	5.01	73.87
	Cfg	38.06	7.41	63.43
	I_{SDT}	28.61	8.20	47.68
	Hcfb	24.09	7.01	40.15

The reliabilities of the versions of the Yes/No Test were low (see Table 9.5) and together with the high mean scores this might be an indication of a ceiling effect. As before, the reliabilities of the pseudoword-items were disturbingly high, they even exceeded the reliabilities of the word-items.

*Table 9.5: Reliabilities of the Yes/No Vocabulary Test calculated by means of Cronbach's Alpha and correlations between the scores on words and the scores on pseudowords for the three materials. Significant correlations are marked with * (p<.05), ** (p<.01) and *** (p<.001).*

Yes/No Test	Rel. (100 items)	Rel. W	Rel. PW	Correlation W/PW
Material X (N=57)	.554	.614	.694	-.199
Material Y (N=60)	.627	.629	.757	-.158
Material Z (N=60)	.656	.665	.750	-.123
Total (N=177)				-.134

The correlations between the performances on word-items versus the performances on pseudoword-items were not significant for the three different materials. However, when this correlation was considered for the totality of the population, it bordered on significance (-.134, p=.076). These correlations are illustrated by means of a scatter plot in Figure 9.1.

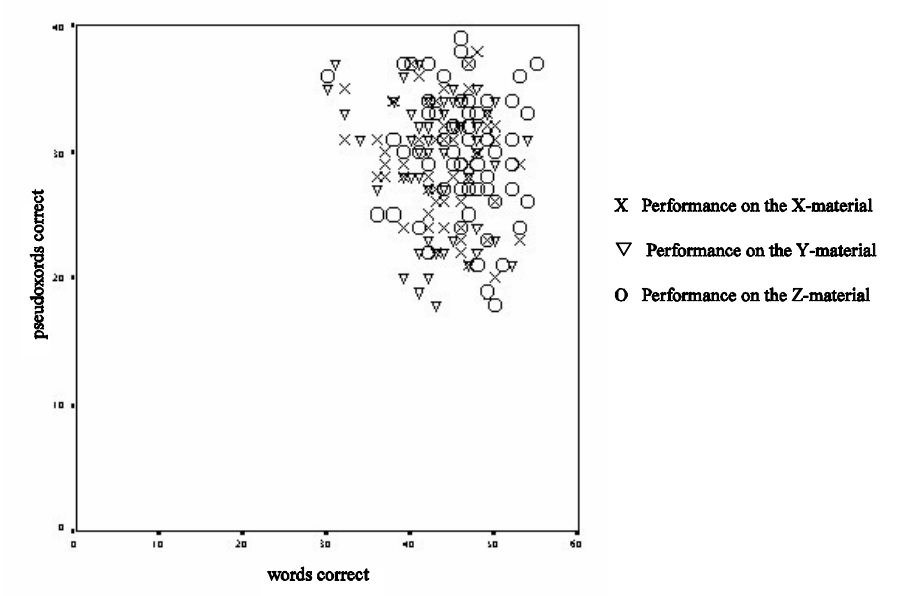


Figure 9.1: Scatter plot of the word-score and the pseudoword-score for the three materials of the Yes/No Vocabulary Test.

When the Yes/No results were split up in a score on the word-items and a score on the pseudoword-items (see Table 9.6), the data showed that the high score on the identification of words (44.32 on a total of 60) was not matched by a near flawless score in rejecting the pseudowords (29.23 on a total of 40).

Table 9.6: Mean score for word- and pseudoword-items on the Yes/No Vocabulary Test.

	Words			Pseudowords		
	Mean	SD	%	Mean	SD	%
N= 177	44.32	5.01	73.87	29.23	4.83	73.08

The stimulus-response matrix in Figure 9.2 confirmed the presence of a bias towards the "Yes"-response. The false alarm rate of 26.9% (which meant that, on average, 11 out of 40 pseudowords were identified as existing words) explained why the raw scores were so severely corrected by the I_{SDT} and Hcfb formulae (see Table 9.4). With reference to the properties of the Yes/No test in this experiment, it could only be concluded that the modification of the test material (in this case: selection from a lower frequency band, restriction to the grammatical categories of the noun and the verb, and different rules for pseudoword formation) did not suffice to resolve the inherent bias problem of the Yes/No format.

		Response alternative: Yes	Response alternative: No
Item alternative	Word	Hit 73.9%	Miss 26.1%
	Pseudoword	False alarm 26.9%	Correct rejection 73.1%

Figure 9.2: Stimulus-response matrix of the Yes/No Vocabulary Test. Percentages are calculated within each response alternative.

The scores for the RBVT 1 and the RBVT 2 were calculated by means of the "correction for blind guessing"-formula[22] because blind guessing was the only thing participants could resort to when they could not distinguish the real word from the pseudoword in the item-pairs. Since the computer applications did not allow for omitted responses the corrected scores were in fact linear transformations of the raw scores.

[22] In the case of the RBVT the "correction for blind guessing"-formula that was used, read: y = x - (60 - x). Because there are no omitted responses, the correlation between x and y is 1.

The mean scores on the RBVT 1 (48.18, 48.08, 48.64; see Table 9.7) were higher than the average score on the Yes/No Test (43.90, 43.02, 46.02; see Table 9.4). This difference could be attributed to the fact that a discrimination task is easier than a detection task considering both tests are infused with the same material because both parts of the RBVT items serve as each other's point of reference. One could also argue that an item in the RBVT contains twice as much information (word and pseudoword) as an item in the Yes/No Test, which is bound to make the decision easier. The reliabilities of the three versions of the RBVT 1 were low. This was probably due to a lack of discriminating items in the tests (see Figure 9.3).

Table 9.7: Mean scores for the RBVT 1. Reliabilities were calculated by means of Cronbach's Alpha.

RBVT 1	Formula	Mean /60	SD	%	Reliability
Mat X (N=62)	Raw	48.18	4.02	80.30	.541
	Corr.	36.36	8.05	60.60	
Mat Y (N=59)	Raw	48.08	4.58	80.13	.670
	Corr.	36.17	9.16	60.28	
Mat Z (N=56)	Raw	48.64	4.08	81.07	.553
	Corr.	37.29	8.17	62.15	
Total (N=177)	Raw	48.29	4.22	80.48	
	Corr.	36.59	8.44	60.98	

With an average score of 51.47 (see Table 9.8), the RBVT 2 (the test version that consists of minimal pairs), appeared to be even easier than the RBVT 1 (mean score of 48.29). This could indicate that although the test tasks of both tests were essentially the same, the skills required to resolve the tests are somehow different. When confronted with item pairs that share the same stem and differ only slightly, the participants succeeded better in distinguishing the word from the pseudoword than when confronted with item pairs that differ fundamentally. The reliabilities of the RBVTs 2 are in the same order as those of the RBVT 1 and, again, these low reliabilities were probably caused by the fact that the tests consisted of many non-discriminating items that did not contribute to the global test reliability.

Table 9.8: Mean scores for the RBVT 2. Reliabilities were calculated by means of Cronbach's Alpha.

RBVT 2	Formula	Mean /60	SD	%	Reliability
Mat X (N=58)	Raw	50.81	3.74	84.68	.554
	Corr.	41.62	7.48	69.37	
Mat Y (N=58)	Raw	51.53	3.41	85.88	.648
	Corr.	43.07	6.82	71.78	
Mat Z (N=61)	Raw	52.03	3.95	86.72	.546
	Corr.	44.07	7.89	73.45	
Total (N=177)	Raw	51.47	3.72	85.78	
	Corr.	42.94	7.44	71.57	

The results of the Translation Test revealed that when it came to actually knowing what the recognized words mean and producing an L1 equivalent of the L2 word, the scores diminished greatly (see Table 9.9). Producing an L1 equivalent of an L2 word proved to be much more challenging than identifying existing words. The reliability of the translation measure was satisfactory (.804).

Table 9.9: Mean score for the Translation Test. Reliabilities were calculated by means of Cronbach's Alpha.

Translation	Mean /60	SD	%	Reliability
Total (N=177)	20.85	5.16	34.75	.804

An item analysis for both the RBVT 1 and RBVT 2 showed that there was a blatant lack of discriminating items in both tests, and this for all test versions. About half of the items of the RBVT 1 (27 in material X, 30 in material Y, and 25 in material Z) were correctly answered by more than 90% of the participants. In the RBVT 2, the ceiling effect was even worse. More than half of the items (35 in material X, 37 in material Y, and 36 in material Z) were correctly answered by more than 90% of the participants. Both ceiling effects are illustrated in Figure 9.3.

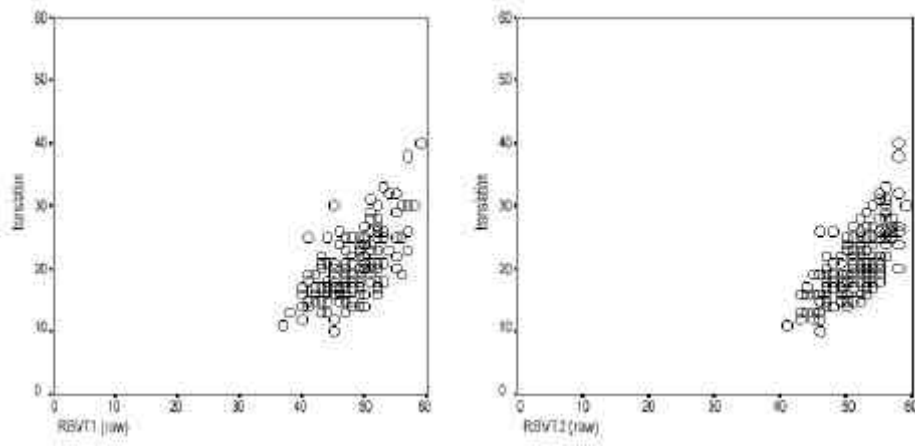


Figure 9.3: Scatter plots of the results on the Translation Test and the results on the RBVT 1 and the RBVT 2.

In Table 9.10, the correlations between the scores on the Translation Test and the scores on the three different vocabulary recognition measures are shown. At first sight, the correlations might seem to be hardly satisfactory but it needs to be remarked that there was no exact overlap of the material this time.

In previous experiments, the translation task contained the same words as the Yes/No Vocabulary Test, whereas in this experiment, the translation task consisted of 20 words from the three materials X, Y and Z, which enlarges the inferential factor. Furthermore, it is clear that although the Translation Test was used in order to check and validate the participants responses in this experiment, it is evident that it essentially measures a different skill than the Yes/No Test and the RBVTs. In view of these two arguments, the correlations with the Translation task were reasonable.

Table 9.10: Correlations between the scores on the Yes/No and RBVT measures and the scores on the Translation Test.

	Yes/No formula				RBVT I	RBVT II
	Raw	cfg	I_{SDT}	Hcfb		
Translation	.384	.538	.594	.589	.620	.689

It can be observed that the correlation between the RBVT measures and the Translation Test (.620 for the RBVT 1 and .689 for the RBVT 2) outperformed the correlation between the Yes/No measure and the Translation Test (.538 [cfg] for the discrete model and .594 [I_{SDT}] and .589 [Hcfb] for the continuous models). However, the difference in correlations between the Yes/No scores that have been corrected with formulae based on continuous models and the new test formats was not significant. However, the lack of significance in difference of correlation was probably caused by the low reliabilities of the RBVT measures, which suppressed the correlation. An RBVT that consists of more difficult items would enhance the test reliability, which in turn might result in a much stronger correlation with the Translation scores.

To conclude this section, the average time it took the participants to complete each of the three tests is considered. On average, the participants needed 4 minutes and 45 seconds to finish the Yes/No Vocabulary Test of 100 items. The RBVT 1 took 3 minutes and 58 seconds for 60 item-pairs. The RBVT 2 was accomplished in 3 minutes and 36 seconds for 60 item-pairs. This means that the RBVTs took up less time than the Yes/No Vocabulary Test although they contained a slightly larger language sample. However, what counts from a vocabulary size perspective is that both test formats consisted of the same number of words to be tested (60). It is clear from these results that the RBVT formats provided a measure of vocabulary recognition for the 60 words that was at least 15% quicker than the Yes/No measure. In view of the fact that the time allocated for placement test procedures is usually relatively short, the gain in time of the RBVTs pleads in their defence.

9.3 Conclusion

On the basis of the results of this experiment, we cannot claim that the RBVT format is a more reliable and valid measure of vocabulary size than the Yes/No Vocabulary Test. However, the low reliabilities and consequently rather disappointing correlations with the Translation task are to be attributed to a lack of discriminating items in the test, and not to a structural problem that is inherent in the format. Conversely, the Yes/No test displayed the same troublesome data as it has always done throughout our use of it. Therefore, the RBVT remains a promising alternative to be explored further.

Obviously, the Recognition Based Vocabulary Test measures a partial kind of word knowledge. It is clear that the demands of the Translation test are much more challenging and rigid. Therefore, the process of extrapolating the RBVT score to a vocabulary size word count has to be approached with caution. It has in any case been established that there is no response bias involved in this vocabulary measure. In Signal Detection terms, we can say that the test task of the RBVT is more univocal than the Yes/No task and it also does not involve the intervention of a "criterion" what so ever. It follows that a simple count of the correct responses will lead to exactly the same results for two learners with the same d' (learners who are of the same proficiency), irrespective of the differences they might exhibit in the Yes/No format due to differences in their criterion placement. Very prudent participants will obtain the same test scores as participants who have a tendency to overestimate their vocabulary knowledge, so that the lengthy discussion of correction formulae is avoided.

Concerning user-friendliness, it can be remarked that the format retains the commodities of the Yes/No Test in terms of ease of construction and it was shown that the RBVT takes up even less time than a Yes/No Test, which is an important factor in language testing in general and estimating vocabulary size in particular.

Future research should direct itself towards making the RBVT more challenging. This could be achieved by infusing the test with a more difficult test content (e.g. selecting words and pseudowords from a lower frequency range), by constructing a longer test, or by increasing the difficulty of the task through pairing words and pseudowords according to an index of difficulty. Since the RBVT 2 suffered from a bigger lack to discriminate than the RBVT 1, it seems appropriate to dismiss the idea of minimal pairs and use an index of difficulty for matching words and pseudowords in the RBVT 1 format. In doing so, one should keep in mind that the relative facility of one of the items of a pair (word or pseudoword) could make the item too easy, which is exactly why random pairing should be abandoned. It is clear that in order to construct reliable and valid measures of vocabulary size some of the universal properties of test construction that made the Yes/No Vocabulary Test so attractive -

random selection of words and pseudowords, use of the format throughout different languages and different language contexts, etc. - have to be sacrificed. The need for test calibration seems to re-assert itself.

Chapter 10

Conclusions and discussion

The aim of this study was to evaluate the use of the Yes/No Vocabulary Test for measuring the vocabulary size of French-speaking learners of Dutch, more specifically, their relative vocabulary size since only the core vocabulary of Dutch (approximately 4000 words) was targeted. It concentrated on handling and suppressing the response bias that the test format seemed to provoke in the test takers which threatened the test's validity. Various experiments were set up to investigate several variables that could contribute to the validity of the Yes/No test in an attempt to recognize the validation process of a language test as an integrated and multifaceted evaluation (Messick 1992, Bachman and Palmer 1996, Chapelle 1994, Chapelle, Jamieson & Hegelheimer 2003).

In this final chapter, we will sum up the most important findings of the seven experiments that have been described and analysed in the preceding chapters of this book (Section 10.1). In Section 10.2, our final conclusion concerning the appropriateness of the use of the Yes/No Test as a measure for receptive vocabulary size will be substantiated and in Section 10.3, the constraints of the study and further research options will be considered. In Section 10.4, we will carefully promote the further development of an alternative format, the Recognition Based Vocabulary Test. To end with, Section 10.5 discusses some reservations and recommendations with reference to the further development of standardized (vocabulary) tests.

10.1 The outcome of the experiments

The data analysis of the first experiment evolved around the contamination of the test scores by a substantial response bias and an exploration of the different ways in which the raw scores could be corrected. It was shown that the different methods of transforming raw scores into corrected scores diverged as false alarm rates increased. The correction schemes that are based on discrete models caused confusion about the difference between guessing and response bias and the correction schemes that are based on continuous models were better equipped to deal with the response bias but their underlying theoretical assumptions still need to be further validated. It was demonstrated that the psychometric qualities of the test suffered when handling this bias because the reliability was overestimated and dropped severely when an attempt was made to extract the bias from the raw score. The biggest drawback constituted the lack of overriding evidence in choosing the most appropriate correction formula. This was even more problematic when considering the large

differences in the participants' rank orders when applying one or the other formula. It was decided that in view of the high false alarm rate and the resulting correction problems, empirical evidence had to be gathered concerning the test's validity.

From then on, a series of experiments was set up in order to determine the factors that were responsible for the high false alarm rate and to evaluate the influence of these factors on the validity of the test. The validity of the Yes/No test scores was verified by means of a translation task in which the participants were asked to translate the words of the Yes/No Test into their L1. Experiment 2 exemplified a first attempt at such validation. The data confirmed the main trends of the results of Experiment 1 (high rate of false alarms and the presence of a response bias in the data). It was clear that when confronted with a rather weak student population (big overlap of distributions) that displayed a considerable variability in their response behaviour (which means that the Criterion covers a wide range), the differences between the correction formulae became unacceptable. When dealing with testees who moved the Criterion to the right (fewer false alarms), the differences slightly converged, but when dealing with testees who moved the Criterion to the left (more false alarms), the differences between the formulae grew so large that the situation became unmanageable from a psychometric point of view. The choice of the most suitable formula was at its most problematic in this case. As could be expected, the lack of validity of the Yes/No responses was established through the low correlations between the Yes/No scores and the translation scores. Clearly, several variables that were not related to the measured construct, but instead to the test takers' profile, interacted with the lexical knowledge that the test claims to measure.

In Experiment 3, an attempt was made to reduce the false alarm rate and consequently the response bias through a careful wording of the instruction. The use of a rigorous instruction did lead to a significant drop of the false alarm rate (the participants grew more careful and shifted their Criterion to the right). Unfortunately, this did not result in a more valid response behaviour. It appeared that we had interfered on the level of the participants' response style, making them ever so vulnerable for exhibiting biased responses that are not related to their lexical knowledge. Influencing participants' response behaviour in the Yes/No Vocabulary Test does not automatically result in a more valid measure of vocabulary size.

In Experiment 4, we took a different tack and sought to investigate the claim that one of the possible advantages of good interface design is that it can reduce the construct-irrelevant variance that is to be attributed to test method (Messick 1989). The experiment centered around the hypothesis that a controlled and sequentially programmed computer design might render the participants' responses more valid. This was not the case. The false alarm rate was not diminished in the experimental group and the concurrent validation

was weak. The bias that seems to be inherent in the Yes/No task manifested itself even more when we tried to intervene in the way participants should accomplish the task.

Experiments 5 and 6 targeted an evaluation of the properties of the Yes/No Test with the European Dialang test content for Dutch. In a first experimental set-up the test was administered to French-speaking participants. The data revealed that the psychometric qualities of the test had not improved: a high false alarm rate tainted the raw scores and although the current calibrated version of the test was an improvement, it was shown that it still remained unsatisfactory with regard to reliability and validity. When the test was administered to native speakers (Experiment 6), the doubtful quality of some of the words and pseudowords was revealed, as the native speakers had a lot of trouble in identifying the existing words in the test. After it had been established that neither the instruction, the computer design nor the Dialang test content had attributed to the validity of the Yes/No format, it was decided that the Yes/No task itself, more specifically the decision criterion that is at the heart of the task, had to be abandoned.

In Experiment 7 a new test was proposed, the Recognition Based Vocabulary Test, in which the testees were presented with pairs of words and pseudowords and had to indicate which item of the pair was an existing word in Dutch. The new test retained the attractive features of the Yes/No Test but replaced the detection task by a discrimination task. Two variants of this new format were compared with a Yes/No Test that consisted of the same material and all test scores were compared with the scores on a translation task. Regretfully, a lack of discriminating items in the test sample and the resulting low reliability and unsatisfactory correlation with the translation scores, prevented us from foregrounding the RBVT as a more reliable and valid measure of vocabulary recognition than the Yes/No Vocabulary Test. The data of the Yes/No test were again characterized by a substantial response bias which undermined the test's validity. It was concluded that the RBVT format could be a more promising measure of vocabulary recognition provided that further language testing research succeeds in making the test more challenging and that the construct validity of this new test is further examined.

Throughout the experiments, it was demonstrated that cognacy between languages was not responsible for the false alarm rate. In fact, it appeared that cognates can work in both directions: they can entice the participants to "Yes" responses which are too tentative but it also happens that cognates get rejected because of their resemblance with the native tongue of the participants. This is also an illustration of the fact that the existence of cognates between languages does not always seem to give the expected advantage in learning second language vocabulary (Ryan 1997). In Experiments 2 and 3, it was shown that the cognates were not responsible for the overestimation of the word knowledge and the only experiment that contained a cognate-free sample

(Experiment 4) revealed that the qualities of the test did not improve when cognates were banned from the test content. The overestimation of vocabulary knowledge remained imminent when there was no confusion caused by the lexical resemblance between the participants' L1 (or L2 and L3) and the target language.

10.2 Use of the Yes/No Vocabulary Test: Yes or No?

The several uses of the Yes/No Vocabulary Tests with French-speaking learners of Dutch have demonstrated that considerable false alarm rates are an indication of a response bias in the data and cause problems in deciding which correction formulae to use. The presence of the response bias in the data indicates that the test is measuring variables that relate to the learners' psychological, cognitive or socio-cultural profile rather than their vocabulary knowledge. The bias constitutes a source of variability that can attain large proportions as a result of which the psychometric qualities of the test in terms of reliability are no longer guaranteed and the validity of the test is endangered. Alderson and Banerjee (2002) point out that in discussing variability, it is important to know to what this variability is to be attributed. They claim that within a unified view of validity, making a distinction between reliability and validity is subsidiary to explaining the sources of that variability. Is the variability relevant to the construct that is being measured? Is it due to error (lack of reliability) or to constructs that should not have been measured like test-wiseness or particular test method effects? In the case of the experiments that have been reported in this study, it has been demonstrated that the bias constitutes construct-irrelevant variability. The test scores were shown to be strongly affected by abilities other than the one we wanted to measure. Therefore the scores did not constitute meaningful indicators of the Dutch core vocabulary size.

It is evident that on the basis of the results we have presented in this study, we have no other option than to argue against the use of the Yes/No Vocabulary Tests for French-speaking learners of Dutch, even in a low-stake situation such as a placement test procedure. Although the Yes/No Vocabulary Test is recommended specifically for placement purposes, we have produced convincing evidence that the format displays no robustness vis-à-vis the many variables that are characteristic of a language testing situation, particularly the circumstances of a placement test procedure.

In today's language testing practice, there is a consensus that determining what degree of relative reliability or validity is required for a particular test context involves a value judgement on the part of the test user (Bachman 1990). Elaborating a validity argument begins with criteria developed on the basis of values in applied linguistics and language testing (Bachman & Palmer 1996). Our use of the Yes/No Vocabulary Test at the Université Libre

de Bruxelles may have had a positive impact on the language centre's daily activity in the sense that it has confirmed the existence of a systematic vocabulary problem and has served as a rationale for incorporating a more deliberate focus on vocabulary within language courses. But, the Yes/No scores themselves were never used to place the students in the appropriate classes. It was clear that too many students could have ended up in the wrong language class. In final analysis a placement test is valid if it allows students to be assigned to classes with only a minimal number of misplacements. According to Messick (1992) the responsibility that is connected to testing is an inherent feature of test validation, and validity is not a feature of a test, but concerns the uses and interpretations of tests and their scores. A language test, as any test, should be evaluated in relation to its purpose and it is our experience as test users that the Yes/No format is too susceptible to the interference of construct-irrelevant variables and therefore falls short as a placement test indicator. The interaction between the effects the pseudowords may exert on learners, the dichotomic character of the Yes/No decision process and the possibility of socio-cultural interference undermines the test's construct validity.

One could argue that low-stakes assessments require less rigorous validation than high-stakes ones because they have smaller effects. The DIALANG organization, for instance, does not supply certificates and refers testees to officially recognized testing boards to obtain these. It is emphasized that the DIALANG system merely serves to inform language learners about their level. Still, when low-stakes assessments like the Yes/No Vocabulary Test are published on the web and countless test takers will invest time and effort in them, proper validation seems a matter of respect and rectitude. Chapelle, Jamieson and Hegelheimer (2003) argue that validation remains essential for tests on the web, even when the examinees are the only recipients of the test results.

10.3 Constraints of the study and further research options

An important question regards the constraints of this study and the generalization of our conclusions. Is the response bias problem of the Yes/No Vocabulary Test confined to the particularities of Belgian French-speaking learners of Dutch? After all, previous accounts of Yes/No Test use did not mention a lack of validity. Unfortunately, there is hardly any mention of the statistical properties of the Yes/No data in the literature. The response data that are being assembled within the Dialang project could give valuable insights into the applicability of the Yes/No Vocabulary Test for different languages and in different language contexts. In the absence of such research data, we would dare to suggest that it is very naïve to presume that high false alarm rates and response biases would not crop up in many other language contexts. In a

recent MA thesis in which the Yes/No Vocabulary test was used with Dutch learners of French, Hommersom (2003) was also confronted with considerable false alarm rates and consequently rather low and unreliable scores. In the literature, studies about self-rating activities and self-assessment tests that do not concern French-speaking learners report tendencies to overestimate abilities as well as acquiescence effects (Heilenman 1990), or reveal that the reliability of learners' self-assessments is affected by their experience of the skill that is being assessed, which seems to indicate that learners resort to recollections of their general proficiency in order to make their judgements (Ross 1998). Even if the response bias problem of the Yes/No format were culturally specific for French-speaking learners, it would still have to be concluded that a format that is susceptible to these kinds of cultural and meta-cognitive variation is not suitable as standardized vocabulary test.

Further research should reconsider the problem of scoring in accordance with current views of the validation process. This means that the several aspects of the testing situation have to be taken into account - and improved when necessary - if one aims to resolve one particular question. This illustrates, once more, that language testing is so complex a matter that a strict Cartesian approach aiming to distinguish the narrow point under interest fails in its purpose because it is unable to capture the linkage between all relevant variables. Too holistic an approach also has potential pitfalls, for we continue to believe that the methodological mistake of selecting a certain scoring method on the basis of a higher reliability (see. Chapter 5), is the consequence of underestimating the sound distinction between what is measured and the accuracy of measurement. It is important to keep looking for ways of reducing the response bias while at the same time improving the test validity. For an investigation of the nature and workings of the bias itself, case studies involving think-aloud procedures when participants are taking the test, would certainly be insightful. We have never taken that course since the bias was considered external to the construct that we aimed to measure. Therefore, it was more important to eliminate or constrain the possibility of a bias appearing in the data than to plumb the depths of it.

If other language centres should consider incorporating the test in their placement procedure, we strongly recommend that language testers revise the role of the pseudowords in the format. If the Yes/No format is to be applied as it has been so far (with the pseudowords functioning as a control measure and false alarm rates that serve to negatively adjust the hit-scores) one should bear in mind that whenever a substantial false alarm rate is encountered in the data, this is evidence of the fact that the participants perform a different task than is expected of them[23], hence the validity of the test becomes doubtful. In these

[23] Given the original instructions in which the participants are asked to indicate if they know the meaning of the presented words.

cases, the competence that is measured with the Yes/No Test does not correlate well with word recognition as measured by a translation task. Therefore, language testers should also clarify the way false alarms should be treated in the calculation of a test score. In our opinion, there are two possible approaches to this discussion:

1) either the pseudowords are considered as a control measure but not a part of the measured construct: for every time the false alarm boundary of 15% is surpassed it is decided that the test's validity is no longer guaranteed and the generated data are dismissed since they do not provide a basis for establishing representative test scores.

2) or the pseudowords are seen as an integral part of the test, which means that being able to distinguish between words and pseudowords is at the heart of the measured construct: the false alarms are considered adjustments of the test score. In this case, it has to be decided how this adjustment needs to be executed. This should preferably be done on the basis of the theoretical model of Signal Detection Theory which would have to be further refined in order to grasp the complexity of the Yes/No task.

In any case, one should continue to be watchful about the presence and the magnitude of a response bias since the Yes/No task is so strongly dependent on a decision criterion, which entails that the format will always be susceptible to biases of whatever kind, depending on the interplay of several factors (language, political and socio-cultural context, meta-cognitive profile, etc).

10.4 Is the Recognition Based Vocabulary Test a valuable alternative?

As was explained in Chapter 9, we think it wiser to explore the possibilities of a derivative test that escapes the ambivalence of the Yes/No task and reduces the undesirable variability between test takers by restraining their response styles. In this perspective the Recognition Based Vocabulary Test could prove to be a valuable alternative. This test retains the advantages of the Yes/No format but succeeds in circumventing the bias problem. When a test taker does not know the answer to a particular item or when he is not sure, the RBVT format offers a 50% chance of getting the correct answer by guessing, and -more importantly - this holds for every test taker (whereas in the Yes/No format, the chances of arriving at a correct answer are determined by the individual's own choice of response style). The built-in constraints of the RBVT format have as a consequence that the test is much more univocal to score (only statistical noise in the data, no possibility of having to correct for bias) and this turns it into a much more user-friendly instrument than the Yes/No Vocabulary Test.

Furthermore, the RBVT task does not rely on self-assessment and it was shown that the face validity of the test is superior to that of the Yes/No Test. Unfortunately, the experiment in which both formats were compared

suffered from a ceiling effect as a result of which the format could not be labelled as superior to the Yes/No Test in terms of psychometric qualities. Therefore, we emphasize that although the format seems very promising, its validity still needs to be established. Does the test provide a measure of the construct word recognition or the construct word knowledge? And in what way do measures of word recognition correlate with measures of word knowledge?

Also, one of the most important conclusions with reference to the construction of standardized vocabulary size test, is that test calibration seems primordial. One of the most appealing features of the Yes/No test was the opportunity to set up a new version by following a set of straightforward rules, with the assurance of getting a reliable and valid measurement of a participant's receptive vocabulary without any preliminary analysis. The experiments have illustrated that a random sampling of words from frequency bands yields a substantial number of items that do not contribute to the qualities of tests (items that are too easy or too difficult).

10.5 A plea for test "robustness"…

Current language testing theory (Bachman 1990, 1991) considers language ability to be multi-componential and acknowledges the influence of the test method and test taker characteristics on test performance. The findings that we have presented in this study confirm the impact of those factors and reinforce the claim that we should investigate the nature and the scope of the tasks we present to test takers and we should be aware of how these tasks may interact with the characteristics of different individuals and with the testing context. The way individuals approach and solve test items can vary enormously, even to the extent that what a particular item may be testing for one test taker is not the same as what it might test for another test taker (even if they share the same level of proficiency). Messick calls the fact that a test score does not reflect a single construct interpretation and that test's construct interpretation might vary from one individual to another "a major current conundrum in educational and psychological measurement" (Messick, 1989: 55). Alderson and Banerjee (2002: 100-101) make a similar claim when they state that:

> "[…] strategies, and presumably traits, can vary across persons and tasks, even when the same scores are achieved. The same test score may represent different abilities, or different combinations of abilities, or different interactions between traits and contexts, and it is currently impossible to say exactly what a score might mean. This we might term The Black Hole of language testing."

Considering the inscrutable nature of language testing, gaining as much control as possible of the processes and strategies that learners engage in when

responding to test items, or trying to minimize the likelihood that tasks could interact with the characteristics of individuals, seems imperative. It is clear that the Yes/No task, due to a combination of self-assessment and the enforced Yes/No dichotomy, is very susceptible to different interactions for different individuals and even to test takers' interpretations of what is expected of them. The confusion about which language task the test takers should exactly perform is probably also caused by the ambivalence of the presence of pseudowords in the format. The fact that the test takers should refrain from aggressively applying the lexical processing strategies that they have been encouraged to develop is much better operationalized in a pair-wise format. Pseudowords as isolated items in a test are a threat to the format's validity because the way learners or groups of learners from various setting and backgrounds are going to react to them is unpredictable and therefore constitutes a big source of variability.

In the pursuit of reliable and valid instruments for measuring receptive vocabulary size and according to the conviction that test evaluation should be determined by pragmatic considerations, we would like to make a plea for the development of tests that are "robust". With this term we refer to the characteristic that a standardized test should interact as little as possible with the learner on the level of format and culture. Nevertheless, the development, use and validation of language measures for populations whose profiles may differ in linguistic as well as in many other ways will entail a continuous monitoring and updating of relevant information, which makes standardization within the language testing domain a tremendous European challenge.

Given that test validation (and science as a whole for that matter) is essentially a process of raising doubts concerning the inferences that are made on the basis of the results of a particular test format, this study can in retrospect be looked upon as a "Popperian" exercise. It represents a counterexample to the effectiveness of the Yes/No format and in its turn awaits to be countered by new research.

Appendix 1

Yes/No Vocabulary Test

Indiquez à l'aide d'une croix les mots que vous connaissez.

Certains mots repris dans la liste n'existent pas en néerlandais !

bijeenkomst		zouwen		architectuur		moeite	
sluk		chauffeur		sportief		ozer	
achtste		regelig		overstellen		leggen	
humor		tegenstander		kwoud		kweul	
touw		getonen		miniman		zind	
welen		pleug		middelbaar		overtreden	
wuide		militair		muurt		schoonzus	
vlakte		nogmeers		weg		draam	
fractie		dakman		evenredig		spel	
meneer		prok		verbod		conclusie	
verzorgen		stif		uiten		doorwerpen	
drukker		industrie		ontspanning		directeur	
erkennen		kapitaan		beha		onderling	
sok		herkomen		toeien		druk	
talent		knie		slim		gevoel	
vernietigen		discriminatie		klanger		ontplooien	
betreuren		kliniek		vanzelf		keit	
moskee		herliezen		verwijken		pestkantoor	
getuigen		binnenkomen		eendom		opereren	
pretachtig		bewegen		profizeren		ook	
ruim		tommerman		beoefenen		zee	
feest		godsachtig		avontaar		camcilie	
schrikken		peper		kraap		ovenal	
herhouden		vrijheid		jals		kaper	
herwijzen		bereiken		scheren		voorbehoedmiddel	

Appendix 2
Yes/No Vocabulary Test

Lisez attentivement la liste reprise ci-dessous, qui comporte des mots (verbes ou noms) existant en néerlandais et des mots qui n'existent pas en néerlandais.
Pour chaque mot : inscrivez **J** (pour "Ja") dans la case si le mot **existe en néerlandais.**
inscrivez **N** (pour "Nee") dans la case si le mot **n'existe pas en néerlandais.**
Donnez une réponse pour chaque mot, même si vous hésitez.

overstellen		verontschuldigen		stand		voek	
top		werf		naaien		legerte	
warenhuis		roeien		verbeteren		gedragen	
overscholen		benaderen		tijder		geloof	
bewonderen		knop		huisvrouw		wolsel	
mantel		samenstelling		pilzen		flent	
naderen		ontvoerder		vok		beweging	
maus		rogen		vuist		braatsen	
lawaai		loren		duisternis		bescheidenheid	
ontraken		bardijn		mat		kring	
sterven		schuilmoeder		uitstellen		toezicht	
schorsheid		arbeid		vreedkamer		nageldom	
hooi		gereedschap		vermoorden		opstopsel	
onderwijs		vork		kroeg		duif	
briem		doodgaan		overzin		stenten	
loods		uitwaarde		overslaan		tussenhalen	
verbinden		uitschillen		achternemen		afwijking	
broodje		kachelaar		appleis		wei	
plek		bewerper		kwaal		beheersen	
aap		vlek		bureelspelen		ontgroting	
pater		grensdom		ouderdom		zeed	
vertrouwen		evenwicht		doorbevelen		fornuis	
kars		bijten		bewerken		afkunst	
aanvaarden		firma		ruilen		kantoor	
goedzenden		burgerij		verlichtsel		halk	

Appendix 3

Material Experiment 1, 2 and 3

Test Version I

Words (60)

avontuur
bevredigen
bewijzen
zuiden
wagen
vouwen
techniek
straffen
stof
steil
voorzitter
voetbal
vervoer
pakken
overal
over
organisme
versieren
verliezen
verdenken
toekomst
zodoende
wind
weide
staart
sociaal
slok
schitterend
richting
rauw
postkantoor
passeren
opsluiten
ontwerpen
onderwerp
oceaan
midden
lichaam
kool
kapitein
goedkoop
godsdienstig
gezicht
gelijk
geit
effect
jack
inhouden
hun
goedvinden
eenheid
droom
kiezen
keuring
doordringen
dijk
demonstreren
dank
boel
biljart

Pseudowords (40)

zulden
wug
vrijlijk
voorbehuidmiddel
uitvaarden
tegenhalen
tegendoen
tandens
spol
schuinzus
schrakken
rommen
reud
practie
plikken
personaal
hool
herreiken
hernietigen
geper
evenradig
directier
compasitie
opverkocht
oop
ondelen
mistuigen
minderachtig
middelachtig
lumor
kwert
kraat
indat
indastrie
chautteur
bijzending
beurderij
benieuw
architectoor
aanplooien

Test Version II

Words (60)

achtste
architectuur
beha
beoefenen
bereiken
betreuren
bewegen
bijeenkomst
binnenkomen
chauffeur
conclusie
directeur
discriminatie
druk
drukker
erkennen
evenredig
feest
fractie
getuigen
gevoel
humor
industrie
kaper
kliniek
knie
leggen
meneer
middelbaar
militair
moeite
moskee
onderling
ontplooien
ontspanning
ook
opereren
overtreden
peper
ruim
scheren
schoonzus
schrikken
slim
sok
spel
sportief
talent
tegenstander
touw
uiten
vanzelf
verbod
vernietigen
verzorgen
vlakte
voorbehoedmiddel
vrijheid
weg
zee

Pseudowords (40)

regelig
avontaar
camcilie
dakman
doorwerpen
draam
eendom
getonen
godsachtig
herhouden
herkomen
herliezen
herwijzen
jals
kapitaan
keit
klanger
kraap
kweul
kwoud
miniman
muurt
nogmeers
ovenal
overstellen
ozer
pestkantoor
pleug
pretachtig
profizeren
prok
sluk
stif
toeien
tommerman
verwijken
welen
wuide
zind
zouwen

Appendix 4

Material Experiment 4

Words (60)

bestuderen
aannemen
bedrijfsleider
verpakking
verwijten
winst
wrijven
zelfmoord
behalen drukwerk
eten
ziekenfonds
leunen
loket
mes
naderen
teleurstellen
toelaten
vakbeweging
nieuws
oplossing
zitting
gedrag
getuige
gooien
grootvader
huisvesting
inslaan
knie
knijpen
koelkast
leer
oppassen
peper
pil
pleidooi
prijzen
proef
rechtspraak
optreden
opvoeding
overeenstemming
overschrijven
regenjas
richting
rollen
ruit
schil
scholier
bezorgen
borstel
cel
doek
aanbeveling
sluipen
spier
student
tegenstander
tekening
verlichten

Pseudowords (40)

afgeverij
angeren
bezekering
bijdeel
bijkomst
bijroep
deuf
doorstellen
echtgelate
bannis
beus
roen
scheiler
scherten
toebetaling
uitstand
uitvraag
veinen
vlaat
vuizen
wulden
uitlatten
bevrijen
hervoeren
erks
gerekenen
grak
herfsting
arstenaar
banighouden
herzakken
herzorgen
hooglering
inkomsting
kroem
navoeren
nawerp
ran
ret
zuiding

Appendix 5

Material Experiment 5 and 6

DIALANG Pilot Set
(150 items)

Words (100)

samenwerken
aanknopen
betreffen
afgescheiden
inschrijven
moeien
opkomen
aanzuiveren
denken
spannen
kenschetsen
examineren
dribbelen
bulderen
nacijferen
laden
losmaken
ondersteunen
pakken
belommeren
camoufleren
degraderen
begroten
lachen
hypnotiseren
krimpen
nakijken
bijeennemen
putten
bevolken
voorafgaan
duwen
toevoegen
ondertekenen
ademen
omspitten
beladen
haken
nestelen
bijmengen
boeten
huishouden
geuren
openslaan
prakkezeren
bezetten
prikken
bijten
onderhouden
aanhitsen
beamen
botsen
muiten
panden
geeuwen
koken
stabiliseren
koteren
poken
glooien
keffen
toestemmen
banen
toelaten
gruizelen
hebben
aanrichten
najagen
brengen
bevallen
hamsteren
detineren
cementen
buitensluiten
instemmen
gadeslaan
strikken
royeren
aandoen
kwijnen
lenzen
avanceren
verhaasten
beweren
exporteren
ploegen
veroorloven
schoren
herinneren
merken
kerven
lakken
brabbelen
overschermen
openbaren
schuwen
aanbelanden
gorden
hekelen
halen

Pseudowords (50)

trechtingen
stremen
blanden
vanoveren
lonsen
deslenversen
onderleunen
winken
groeibaren
vandagen
valteren
krilderen
veradelen
naarmen
maanderen
ontlonen
zwamelen
spotteren
ouderjassen
tochtigingen
bouden
zween
holteren
sloeten
afbreden
inzoeken
vuivelen
verpaarden
tweezamen
broekgaan
verhekken
ovelen
pretten
belegenen
zingbaren
ontkippen
kwaadstoken
dretten
achterslaan
geteren
hulderen
studden
amheden
dolaren
loteren
lichtvollen
vraken
hollieperken
vandaagen
viezelen

Appendix 6

Recognition Based Vocabulary Test I (pairs)

Lisez attentivement la liste de paires de mots reprise ci-dessous.

Chaque paire comporte **un et un seul mot existant en néerlandais** (soit un verbe, soit un nom).

Soulignez le mot existant comme dans les deux exemples.

Donnez une réponse pour chaque paire, même si vous hésitez.

ex.1	vernieuwen	gezetter		concect	kwaal
ex.2	lonnen	pijl		nageldom	verbeteren
	mug	hulpverboding		kring	overzin
	invachten	tennis		overscholen	ouderdom
	beker	wank		sterven	doorbevelen
	verpauwen	afspelen		stenten	beweging
	monopolie	nuil		vertrouwen	belijfte
	tank	vetopolie		vermoorden	kliezen
	vok	afdrogen		ruilen	bureelspelen
	doker	ruilen		fornuis	uitschillen
	bezolpen	kuil		mat	voedheid
	duivel	inkollen		arbeid	opstopsel
	hulpverlening	rielen		maus	onderwijs
	uitlap	zender		werf	bewerper
	detective	lampetitie		gedragen	ergerdom
	omscholen	kannis		aanvaarden	achternemen
	verprogen	slip		doodgaan	zeed
	gemeerte	hooi		voek	bescheidenheid
	bezetter	pensiaul		braatsen	aap
	kleuter	vlommen		beheersen	herdragen
	bezichtigen	intekken		geloof	verbrengen
	inpakken	afspagen		samenstelling	overstellen
	wetsplaats	brommen		burgerij	overzin
	kleuvel	overdrijven		verontschuldigen	uitwaarde
	opinie	eranie		afkunst	ontvoerder
	rollen	breul		achternemen	hooi
	campiste	onthouden		uitstikken	bewerken
	opzacht	zolder		benaderen	kachelaar
	neigils	invullen		uitstikken	plek
	bezorgen	pijr		top	rogen
	mechanisme	verluiping		stand	oprijken

Appendix 7

Recognition Based Vocabulary Test II (minimal pairs)

Lisez attentivement la liste de paires de mots reprise ci-dessous.

Chaque paire comporte **un et un seul mot existant en néerlandais**
(soit un verbe, soit un nom)
Soulignez le mot existant comme dans les deux exemples.

Donnez une réponse pour chaque paire, même si vous hésitez.

ex. 1	Rennen	lonnen	dirigeren	dirivaren
ex. 2	overdrielen	overdrijven	uitwerken	uitwelpen
	bezorgen	bezolpen	bestroeden	bestrijden
	brommen	vlommen	opvoeden	omgoeden
	afspagen	afspelen	toekennen	toezinnen
	preken	prumen	schirmen	schillen
	bewachtigen	bezichtigen	smeren	pleren
	indienen	indeuren	beheersen	betoorsen
	verplegen	verprogen	ontslaan	ontspeen
	intekken	inpakken	afbetalen	afbepelen
	afzetten	afbitten	pegen	zagen
	uitluren	uitkeren	begraven	begropen
	omscholen	omschieren	stelen	stuken
	ontroeden	onthouden	aanrien	aangaan
	inrichten	invachten	knielen	kneupen
	gullen	rollen	stiepen	staken
	inkollen	invullen	aanzetten	aanmatten
	vernieuwen	verpauwen	inleiden	inlouten
	ruilen	rielen	vernietigen	verruitigen
	afdrogen	afdralen	vancurreren	concurreren
	breul	breuk	preek	dreek
	campagne	campiste	das	dap
	ons	ors	biolaar	bioloog
	pla	vla	geheugen	geriegen
	zolder	talder	race	bice
	nadeel	zedeel	fan	fap
	vertainer	container	symbool	symbaat
	mug	mup	roning	lening
	financiën	finantaan	dialoog	dialaan
	kleuter	kleuvel	klinaal	kliniek
	mechanisme	vachanisme	spandoek	speldoek

Appendix 8 a

Yes/No Vocabulary Test

Material X

Words (60)

rennen
overdrijven
bezorgen
brommen
afspelen
preken
bezichtigen
indienen
verplegen
inpakken
afzetten
uitkeren
omscholen
onthouden: zich -
inrichten
rollen
invullen
vernieuwen
ruilen
afdrogen
breuk
campagne
ons
vla
zolder
nadeel
container
mug
financiën
kleuter
mechanisme
slip
uitleg
sok
pensioen
effect
pijl
hooi
kuil
hulpverlening
zender
tank
werkplaats
lijk
optocht
trommel
opinie
ontvoering
neiging
vlinder
verdieping
competitie
monopolie
bezetter

kwaal
tennis
beker
geboorte
duivel
detective

Pseudowords (40)

lonnen
overdrielen
bezolpen
vlommen
afspagen
prumen
bewachtigen
indeuren
verprogen
intekken
afbitten
uitluren
omschieren
breul
campiste
ors
pla
talder
zedeel
vertainer
mup
finantaan
kleuvel
vachanisme
slit
uitlap
vok
pensiaul
effont
pijr
hool
nuil
hulpverboding
bonder
wank
wetsplaats
lijs
opzacht
stammel
eranie

Material Y

Words (60)

dirigeren
uitwerken
bestrijden
opvoeden
toekennen
schillen
smeren
beheersen
ontslaan
afbetalen
zagen
begraven
stelen
aangaan
knielen
staken
aanzetten
inleiden
vernietigen
concurreren
preek
das
bioloog
geheugen
race
fan
symbool
lening
dialoog
kliniek
spandoek
wolk
kandidaat
zuinigheid
kam
roos
wol
schijf
geweten
voer
geding
ontspanning
doelstelling
wieg
oppositie
brievenbus
beha
automatisering
ziekenfonds
advocaat
catalogus
verbetering
bezienswaardigheid
concern

duel
muis
olie
wortel
leverancier
tomaat

Pseudowords (40)

dirivaren
uitwelpen
bestroeden
omgoeden
toezinnen
schirmen
pleren
betoorsen
ontspeen
afbepelen
pegen
begropen
stuken
dreek
dap
biolaar
geriegen
bice
fap
symbaat
roning
dialaan
klinaal
speldoek
wolp
perdidaat
peurigheid
kag
roog
wom
schijp
geraten
voek
gelang
ontstenning
doelspilling
zieg
antositie
brievenlek
bete

Material Z

Words (60)

plukken
remmen
aanbevelen
voorsorteren
uitvoeren
afvaardigen
vriezen
opruimen
bewerken
verwijten
ondertekenen
melken
strooien
bevorderen
verwaarlozen
graven
stofzuigen
opgroeien
bijten
presenteren
weide
tempo
staal
huisvesting
wang
departement
fotograaf
aflevering
grootvader
eigenaar
gehoor
aandeelhouder
interview
laken
ceintuur
indeling
dichter
saus
rechtbank
auteur
maag
cheque
mouw
interpretatie
kuur
voorraad
moraal
bruid
straal
uiting
gymnastiek
humor
bewind
bak
moordenaar
eigendom
boterham
dagtaak
paraplu
datum

Pseudowords (40)

plitten
ressen
aanbatelen
voorsommeren
uitmeuren
afvoeldigen
pliezen
alruimen
bewonken
verwieken
onderbakenen
mersen
sprooien
ruide
temte
staap
reusvesting
wans
minortement
fotogreem
aflopering
blaatvader
autenaar
gepuur
aarmeelhouder
intervaas
lapen
rijntuur
inzuling
dichtel
saup
rechtbalm
autiek
raag
bleque
moul
interpremotie
kuut
voorleed
minaal

Appendix 8 b

Recognition Based Vocabulary Test I (pairs)

Material X

vernieuwen - gezetter
pijl - lonnen
mug - hulpverboding
tennis - invachten
beker - wank
afspelen - verpauwen
monopolie - nuil
tank - vetopolie
afdrogen - vok
ruilen - doker
kuil - bezolpen
duivel - inkollen
hulpverlening - rielen
zender - uitlap
detective - lampetitie
omscholen - kannis
slip - verprogen
hooi - gemeerte
bezetter - pensiaul
kleuter - vlommen
bezichtigen - intekken
inpakken - afspagen
brommen - wetsplaats
overdrijven - kleuvel
opinie - eranie
rollen - breul
onthouden - campiste
zolder - opzacht
invullen - neigils
bezorgen - pijr
mechanisme - verluiping
rennen - pla
competitie - afdralen
kwaal - vertainer
trommel - ontroeden
financiën - slit
effect - omschieren
breuk - finantaan
optocht - rotective
preken - ontleuring
vla - indeuren
lijk - vachanisme
verdieping - prumen
pensioen - zedeel
ons - kwaap
nadeel - bonder
afzetten - ors
verplegen - bewachtigen
uitleg - stammel
indienen - effont
vlinder - hool
neiging - afbitten
sok - talder
uitkeren - overdrielen
geboorte - duigil
campagne - lijs
inrichten - uitluren
werkplaats - sponder
container - mup
ontvoering - gullen

Material Y

wortel - ontstenning
bestrijden - omgoeden
wieg - kneupen
beheersen - wolp
staken - doelspilling
geding - verbakering
vernietigen - roning
advocaat - kag
automatisering - vancurreren
inleiden - muel
concern - schirmen
bezienswaardigheid - pleren
uitwerken - tomeer
aanzetten - dap
kliniek - voek
beha - afbepelen
catalogus - adviraat
voer - fap
dirigeren - aanrien
ziekenfonds - betoorsen
fan - kuikenfonds
afbetalen - stiepen
wolk - antositie
wol - dreek
roos - symbaat
concurreren - bete
knielen - geraten
race - geriegen
dialoog - zuis
muis - ontspeen
verbetering - dirivaren
opvoeden - uitwelpen
kandidaat - koperancier
symbool - bezienspoordigheid
olie - bestroeden
das - schijp
zagen - biolaar
ontspanning - concect
leverancier - verruitigen
kam - klinaal
lening - stuken
begraven - gelang
zuinigheid - inlouten
toekennen - toezinnen
geheugen - perdidaat
ontslaan - dialaan

preek - begropen
spandoek - automelanering
brievenbus - pegen
stelen - zieg
doelstelling - bice
tomaat - aanmatten
bioloog - elie
schillen - speldoek
schijf - peurigheid
duel - wormel
oppositie - wom
aangaan - catametus
geweten - brievenlek
smeren - roog

Material Z
alruimen - moraal
temte - eigenaar
ruide - tempo
afvoeldigen - departement
bevelderen - laken
minortement - bevorderen
stelzuigen - mouw
verwaarlepen - rechtbank
sprooien - opgroeien
rijntuur - bruid
zieten - staal
wans - aandeelhouder
aflopering - remmen
bewonken - grootvader
moul - auteur
auling - verwijten
raag - bewind
fak - kuur
lapen - afvaardigen
bewilt - aanbevelen
moordeleen - eigendom
ressen - uitvoeren
uitmeuren - gehoor
viraplu - opruimen
mersen - bak
boterlep - presenteren
inzuling - verwaarlozen
gepuur - wang
autenaar - bewerken
autiek - moordenaar
aanbatelen - voorraad
saup - fotograaf
mornastiek - datum
verwieken - melken
minaal - huisvesting
rechtbalm - ceintuur
reusvesting - saus
plitten - maag
huker - plukken
gluid - vriezen
omstoeien - uiting
greken - bijten
onderbakenen - gymnastiek
dater - interpretatie
voorleed - dagtaak
interpremotie - boterham
fotogreem - straal
kuut - strooien
straak - weide
plomenteren - humor
staap - ondertekenen
eigendal - graven
bleque - cheque
blaatvader - stofzuigen
kiltaak - voorsorteren
voorsommeren - paraplu
dichtel - interview
intervaas - dichter
pliezen - aflevering
aarmeelhouder - indeling

Appendix 8 c

Recognition Based Vocabulary Test II (minimal pairs)

Material X

rennen - lonnen
overdrijven - overdrielen
bezorgen - bezolpen
brommen - vlommen
afspelen - afspagen
preken - prumen
bezichtigen - bewachtigen
indienen - indeuren
verplegen - verprogen
inpakken - intekken
afzetten - afbitten
uitkeren - uitluren
omscholen - omschieren
onthouden - ontroeden
inrichten - invachten
rollen - gullen
invullen - inkollen
vernieuwen - verpauwen
ruilen - rielen
afdrogen - afdralen
breuk - breul
campagne - campiste
ons - ors
vla - pla
zolder - talder
nadeel - zedeel
container - vertainer
mug - mup
financiën - finantaan
kleuter - kleuvel
mechanisme - vachanisme
slip - slit
uitleg - uitlap
sok - vok
pensioen - pensiaul
effect - effont
pijl - pijr
hooi - hool
kuil - nuil
hulpverlening - hulpverboding
zender - bonder
tank - wank
werkplaats - wetsplaats
lijk - lijs
optocht - opzacht
trommel - stammel
opinie - eranie
ontvoering - ontleuring
neiging - neigils
vlinder - sponder
verdieping - verluiping
competitie - lampetitie
monopolie - vetopolie
bezetter - gezetter
kwaal - kwaap
tennis - kannis
beker - doker
geboorte - gemeerte
duivel - duigil
detective - rotective

Material Y

dirigeren - dirivaren
uitwerken - uitwelpen
bestrijden - bestroeden
opvoeden - omgoeden
toekennen - toezinnen
schillen - schirmen
smeren - pleren
beheersen - betoorsen
ontslaan - ontspeen
afbetalen - afbepelen
zagen - pegen
begraven - begropen
stelen - stuken
aangaan - aanrien
knielen - kneupen
staken - stiepen
aanzetten - aanmatten
inleiden - inlouten
vernietigen - verruitigen
concurreren - vancurreren
preek - dreek
das - dap
bioloog - biolaar
geheugen - geriegen
race - bice
fan - fap
symbool - symbaat
lening - roning
dialoog - dialaan
kliniek - klinaal
spandoek - speldoek
wolk - wolp
kandidaat - perdidaat
zuinigheid - peurigheid
kam - kag
roos - roog
wol - wom
schijf - schijp
geweten - geraten
voer - voek
geding - gelang
ontspanning - ontstenning
doelstelling - doelspilling
wieg - zieg
oppositie - antositie
brievenbus - brievenlek
beha - bete
automatisering - automelanering
ziekenfonds - kuikenfonds
advocaat - adviraat
catalogus - catametus

verbetering - verbakering
bezienswaardigheid - bezienspoordigheid
concern - concect
duel - muel
muis - zuis
olie - elie
wortel - wormel
leverancier - koperancier
tomaat - tomeer

humor - huker
bewind - bewilt
bak - fak
moordenaar - moordeleen
eigendom - eigendal
boterham - boterlep
dagtaak - kiltaak
paraplu - viraplu
datum - dater

Material Z
plukken - plitten
remmen - ressen
aanbevelen - aanbatelen
voorsorteren - voorsommeren
uitvoeren - uitmeuren
afvaardigen - afvoeldigen
vriezen - pliezen
opruimen - alruimen
bewerken - bewonken
verwijten - verwieken
ondertekenen - onderbakenen
melken - mersen
strooien - sprooien
bevorderen - bevelderen
verwaarlozen - verwaarlepen
graven - greken
stofzuigen - stelzuigen
opgroeien - omstoeien
bijten - zieten
presenteren - plomenteren
weide - ruide
tempo - temte
staal - staap
huisvesting - reusvesting
wang - wans
departement - minortement
fotograaf - fotogreem
aflevering - aflopering
grootvader - blaatvader
eigenaar - autenaar
gehoor - gepuur
aandeelhouder - aarmeelhouder
interview - intervaas
laken - lapen
ceintuur - rijntuur
indeling - inzuling
dichter - dichtel
saus - saup
rechtbank - rechtbalm
auteur - autiek
maag - raag
cheque - bleque
mouw - moul
interpretatie - interpremotie
kuur - kuut
voorraad - voorleed
moraal - minaal
bruid - gluid
straal - straak
uiting - auling
gymnastiek - mornastiek

References

Abels, M. (1994) Ken ik dit woord? MA Thesis, University of Nijmegen.

Abu Rabia, S. & Seigel, L.S. (1995) Different orthographies, different context effects: The effects of Arabic sentence context in skilled and poor readers. *Reading Psychology*. 16, 1-19.

Alderson, J.C. & Banerjee, J. (2001) Language testing and assessment (Part 1) State-of-the-Art Review. *Language Testing* 18, 213-236.

Alderson, J.C. & Banerjee, J. (2002) Language testing and assessment (Part 2) State-of-the-Art Review. *Language Testing* 19, 79-113

Al-Hazemi, H. (1993) Low level EFL vocabulary tests for Arabic speakers. MA Thesis, University of Wales, Swansea.

Anckaert, P. & Beeckmans, R. (1992) Le C-Test. Difficulté intrinsèque, pouvoir discriminant et validité de contenu. In Grotjahn, R., editor, *Manuskripte zur Sprachlehrforschung*, (pp. 145-172) Bochum, Germany: Universitätsverlag Brockmeyer.

Anderson, R.C. & Freebody, P. (1981) Vocabulary Knowledge. In J.T. Guthrie (ed.) *Comprehension and Teaching: Research Reviews* (pp.77-117). Newark, DE: International Reading Association.

Anderson, R.C. & Freebody, P. (1983) Reading comprehension and the assessment and acquisition of word knowledge. In B. Huxton (ed.), *Advances in Reading/Language Research*. Volume 2 (pp. 231-256). Greenwich, CT: JAI Press.

Anglin, J.M. (1993) Vocabulary development: a morphological analysis. *Monographs of the Society for Research in Child Development,* 58 (10*, Serial no.238),* 1-165.

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1991) What does language testing have to offer? *TESOL Quarterly*, 25, nr. 4, 671-704.

Bachman, L. F. (2000) Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17, 1-42.

Bachman, L.F. & Palmer A. (1996) *Language Testing in Practice* Oxford : Oxford University Press

Beck, I.& McKeown, M.(1991) Conditions of vocabulary acquisition. In R. Barr. M. Camail, P. Mosenthal & P.D. Pearson (eds), *The Handbook of Reading Research*, Vol. II: (pp.789-814).

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M. & Van de Velde, H. (2001) Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing*, 18, 235-274.

Cameron, L. (2002) Measuring vocabulary size in English an an additional language. *Language Teaching Research.* 6,2, 145-173.

Chall, J.S. & Dale, E. (1950) Familiarity of selected health terms. *Educational Research Bulletin*. 39, 197-206.

Chapelle, C. (1994) Are C-Tests valid measures for L2 vocabulary research? *Second Language Research* 10, 157-187.

Chapelle, C. (2001) *Computer Applications in Second Language Acquisition.* Cambridge: Cambridge University Press.

Chapelle, C., Jamieson, J. & Hegelheimer, V. (2003) Validation of a web-based ESL test. *Language Testing* 20, 409-434.

Coady, J, Magoto, J., Hubbard, P., Graney, J. & Mokhtari, K. (1993) High frequency vocabulary and reading proficiency in ESL readers. In T. Huckin, M. Haynes and J. Coady (Eds.) *Second Language Reading and Vocabulary.*(pp. 3-23) Norwood, NJ.: Ablex.

Cobb, T. (2000) One size fits all? Francophone learners and English vocabulary tests. *The Canadian Modern Language Review/ La Revue canadienne des langues vivantes,* 57, 2, 295-324.

Cohen, A.D. (1994) *Assessing Language Ability in the Classroom*. New York: Heinle and Heinle

de Jong, J.H.A.L. (1992) Assessment of language proficiency in the perspective of the 21th century. *AILA Review*, 9, 39-45.

Dieltjens, L., Vanparijs, J., Baten, L., Claes, M.-T., Alkema, P. & Lodewick, J. (1995) *Woorden in Context Deel 2.* Brussels: De Boeck.

Dieltjens, L., Vanparijs, J., Baten, L., Claes, M.-T., Alkema, P. & Lodewick, J. (1997) *Woorden in Context Deel 1.* Brussels: De Boeck.

Ellis, N.C. (1997) Vocabulary acquisition, word structure, collocation, word-class, and meaning. In N. Schmitt & M. McCarthy (eds.) *Vocabulary: Description, Acquisition and Pedagogy*, (pp.122-139) Cambridge: Cambridge University Press.

Ellis, N.C. (2002) Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.

Eyckmans, J.; Beeckmans, R. and Van de Velde H. (2001) Characteristics and implications of a response bias in the Yes/No Vocabulary Test. Paper presented at the 11th Vocabulary Acquisition Research Group Network Conference, 20th July, University of Wales.

Eyckmans, J.; Beeckmans, R. and Van de Velde H. (2002) The Paired Vocabulary Test. Paper presented at the Second Language Vocabulary Acquisition Colloquium, 16th March, Leiden.

Fulcher, G. (2003) Interface design in computer-based language testing. *Language Testing* 20, 384-408.

Gervais, C. (1997) Computers and language testing; a harmonious relationship? *Francophonie*,16,3-7.

Green, D.M. & Swets, J.A. (1966) *Signal detection theory and psychophysics*. New York: John Wiley.

Groot, P.J.M. (1990) Language testing in research and education: the need for standards. *AILA Review,* 7, 9-23.

Guilford, J. P. (1946) New standards for test evaluation. *Educational and Psychological Measurement* 6, 427-439.

Hacquebord, H. (1999) Lees- en luisterbegrip van studieteksten bij Nederlandse en anderstalige leerlingen en studenten. In E. Huls & B. Weltens (eds.) *Artikelen van de Derde Sociolinguistische Conferentie.*(pp. 161-172) Delft: Eburon.

Hazenberg, S. (1994) *Een Keur van Woorden. De wenselijke en Feitelijke Receptieve Woordenschat van Anderstalige Studenten.* PHD dissertation. Amsterdam: Vrije Universiteit.

Hazenberg, S. & Hulstijn, J.H. (1996) Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics* 17, 145-163.

Heilenman, L.K. (1990) Self-assessment of second language ability: the role of response effects. *Language Testing*, 7, 174-201.

Hermans, D. (2000) Word production in a foreign language. MA Thesis, University of Nijmegen.

Hirsh, D. & Nation, I.S.P. (1992) What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8, 689-696.

Hodos, W. (1970) Nonparametric index of response bias for the use in detection and recognition experiments. *Psychological Bulletin* 74 , 351-354.

Hommersom, M. (2003) Het testen van receptieve woordenschatkennis. MA Thesis. University of Utrecht.

Huibregtse, I. & Admiraal, W. (1999): De score op een ja/nee-woordenschattoets: correctie voor raden en persoonlijke antwoordstijl. *Tijdschrift voor Onderwijsresearch* 24, nr. 2, 110 – 124.

Huibregtse, I, Admiraal, W & Meara, P. (2001) Scores on a yes/no vocabulary test: correction for guessing and response style. *Language testing* 18, 227-245.

Janssens, V. (1999) Over 'slapen' en 'snurken' en de hulp van de context hierbij. *ANBF-nieuwsbrief* 4, 29-45.

Janssen-van Dieten, J.M. (1992) Zelfbeoordeling en tweede- taalleren : een empirisch onderzoek naar zelfbeoordeling bij volwassen leerders van het Nederlands. PHD dissertation, University of Nijmegen.

Laufer, B. (1989) What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (eds.) *Special Language: From Humans Thinking to Thinking Machines.* (pp.316-323) Clevedon: Multilingual Matters.

Laufer, B. (1992) How much lexis is necessary for reading comprehension? In P. Arnaud & H. Béjoint (eds.) *Vocabulary and Applied Linguistics.*(pp.126-132) London: Macmillan.

Laufer, B. (1998) The development of passive and active vocabulary: same or different? *Applied Linguistics*, 19, 255-271

Lewis, M. (1993) *The Lexical Approach. The State of ELT and a way forward.* Hove: Language Teaching Publications.

Lewis, M. (1997) *Implementing the Lexical Approach. Putting Theory into Practice.* Hove: Language Teaching Publications.

Lewis, M. (Ed.)(2000) *Teaching Collocation*. Hove : LTP.

Lord, F.M. (1980) *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Laurence Erlbaum.

Luoma, S. & Tarnanen, M. (2003) Creating a self-rating instrument for second language writing: from idea to implementation. *Language Testing* 20, 440-465.

Meara, P. (1990) Some notes on the Eurocentres vocabulary tests. In J. Tommola (ed.) *Foreign Language Comprehension and Production.* (pp. 103-113) Turku : AFinLa

Meara, P.(1992) *EFL Vocabulary Tests*. Swansea: Centre for Applied Language Studies, University of Wales.

Meara, P.(1993) The bilingual lexicon and the teaching of vocabulary. In R. Schreuder & B. Weltens (eds.) *The Bilingual Lexicon*. (pp. 279-297) Amsterdam, Philadelphia: John Benjamins.

Meara, P. (1995) The importance of early emphasis on L2 vocabulary. *The Language Teacher*, 19,2, 8-11.

Meara, P.(1996) The dimensions of Lexical Competence. In G. Brown, K. Malmkjaer and J. Williams (eds.) *Performance and Competence in Second Language Acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P. (2002) The rediscovery of vocabulary. *Second Language Research*. 18, 4, 393-407.

Meara, P & Buxton, B. (1987) An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-151.

Meara, P; & Jones, G. (1988) Vocabulary size as a placement indicator. In P. Grunwell (ed.) *Applied Linguistics in Society* (pp. 80-87). London: Centre for Information on Language Teaching and Research.

Meara, P; & Jones, G. (1990) *Eurocentres Vocabulary Size Test, Version E1.1/K10.* Zurich: Eurocentres Learning Service.

Meara, P., Lightbown, P.M. & Halter, R.H. (1994) The effect of cognates on the applicability of yes/no vocabulary tests. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes,* 50,2, 296-311.

Messick, S. (1989) Validity. In Linn, R. L., (ed.), *Educational measurement*. 3rd edn. New York: American Council on Education/Macmillan, 13-103.

Messick, S. (1992) Validity of test interpretation and use. In M.C. Alkin (ed.) *Encyclopedia of educational research*. Sixth edition. New York: Macmillan, 1487-1495.

Mettewie, L. (2003) Contacthypothese en taalleermotivatie in Nederlandstalige scholen in Brussel. *Toegepaste Taalwetenschap in Artikelen*, 70, 2, 79-89.

Meunier, L. E. (1994) Computer adaptive language tests (CALT) offer a great potential for functional testing. Yet, why don't they? *CALICO Journal,* 11 (4), 29-39.

Morelli, A., Dierickx, L. en Lesage, D. (1998) *Racisme: een element in het conflict tussen Franstaligen en Vlamingen.* Berchem/Bruxelles: EPO/Labor.

Nation, I.S.P. (1983) Testing and teaching vocabulary. *Guidelines* 5, 12-25.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary.* New York : Heinle and Heinle.

Nation, I.S.P. (1993) Vocabulary size, growth, and use. In Schreuder, R. & Weltens, B. (eds.) *The Bilingual Lexicon.* (pp.115-134) Amsterdam : Benjamins.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation, I.S.P. & Waring, R. (1997) Vocabulary size, text coverage and word lists. In N. Schmitt and M McCarthy (eds.) *Vocabulary: description, acquisition and pedagogy*. (pp.6-19) Cambridge: Cambridge University Press.

Nattinger, J.R. & DeCarrico, J.S. (1992) *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nieuwborg, E. (1992) Tekstdekking en tekstbegrip. Een experimenteel onderzoek. In Halbo, A. (ed.) *Evaluation and Language Teaching. Liber Amicorum Frans van Passel.* Bern: Peter Lang.

Nunally, J.C. & Bernstein, I.H. (1994) *Psychometric Theory*. New York: McGraw-Hill.

Oscarson, M. (1997) Self-Assessment of foreign and second language proficiency. In Clapham, C. & Corson, D. (eds.) *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment.* (pp.175-187) Dordrecht: Kluwer Academic Publishers.

Read, J. (1993) The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10, 355-371.

Read, J. (1997a) Assessing vocabulary in a second language. In Clapham, C. & Corson, D. (eds.) *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment.*(pp. 99-107) Dordrecht: Kluwer.

Read, J. (1997b) Vocabulary and testing. In Schmitt, N. & McCarthy, M. (eds.) *Vocabulary: Description, Acquisition and Pedagogy.* (pp. 303-320) Cambridge: Cambridge University Press.

Read, J. (2000) *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Richards, J.C. (1976) The role of vocabulary teaching. *TESOL Quarterly* 10, 77-89

Rietvelt, T. & Van Hout, R. (1993*) Statistical techniques for the study of language and language behaviour*. Berlin, New York: Mouton de Gruyter.

Ross, S. (1998) Self-assessment in second language testing: a meta-analysis of experimental factors. *Language Testing,* 15, 1-20.

Ryan, A. (1997) Learning the orthographic form of L2 vocabulary – a receptive and a productive process. In Schmitt, M. & McCarthy, M. (eds.) *Vocabulary: Description, Acquisition & Pedagogy* (pp.181-198). Cambridge: Cambridge University Press.

Ryan, A. & Meara, P. (1991) The case of the invisible vowels: Arabic speakers reading English words. *Reading in a Foreign Language*, 7,2, 531-540.

Sasaki, M**.** (2000) Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing* 17, 85-114.

Schmitt, N. (2000) *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Sciarone, A.G. (1979) *Woordjes Leren in het Vreemde-talenonderwijs.* Muiderberg. Netherlands: Coutinho.

Shillaw, J. (1996) The Application of Rasch modelling to yes/no vocabulary tests. Vocabulary Acquisition Research Group discussion document No.js.96a, available over the Internet at www.swan.ac.uk/cals/vlibrary/js96a.htm

Shohamy, E. (2001) *The Power of Tests* : *A Critical Perspective on the Uses of Language Tests.* Essex : Longman, Pearson Education Limited.

Sims, V.M. (1929) The reliability and validity of four types of vocabulary test. *Journal of Educational Research*, 20, 91-96.

Skehan, P. (1998) *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Stanovich, K. (1980) Towards an interactive-compensatory model of individual differences in the development of reading. *Reading Research Quarterly.* 16, 32-71.

Terman, L.M. (1918) Vocabulary tests as a measure of intelligence. *Journal of Educational Psychology.* 9, 452-466.

Thorndike, E.L. & Lorge, I. (1944) *The Teacher's Word Book of 30,000 Words.* New York: Teachers College, Columbia University.

Tilley, H.C. (1936) A technique for determining the relative difficulty of word meanings among elementary school children. *Journal of Experimental Education* 5, 61-64.

Van de Walle, P. (1999) Onderzoek naar de omvang van de receptieve en productieve kennis van de basiswoordenschat van zesdeklassers uit het A.S.O. in het Brussels Hoofdstedelijk Gewest. MA Thesis, Université Libre de Bruxelles.

Van Hout, R. en Knops, U. (1988). *Language Attitudes in the Dutch Language Area.* Dordrecht: Foris Publications.

Vives Boix, G. (1995) The development of a measure of lexical organization: the Association Vocabulary Test. PHD- dissertation, University College of Swansea, University of Wales.

Weir, C.J. (1993) *Understanding and Developing Language Tests*. London: Prentice Hall.

Willis, D. (1990) *The Lexical Syllabus.* Collins: Cobuild.

Wittrock, M.C., Marks, C. & Doctorow, M (1975) Reading as a generative process. *Journal of Educational Psychology*, 67, 484-489.

Zimmerman, J., Broder, P.K., Shaughnessy, J.J. & Underwood, B.J. (1977) A recognition test of vocabulary using signal-detection measures and some correlates of word and non word recognition. *Intelligence* 1, 5-13.

Nederlandse samenvatting

In deze studie worden de betrouwbaarheid en validiteit van de *Yes/No Vocabulary Test* onderzocht. Deze toets werd eind jaren tachtig ontwikkeld en is een gestandaardiseerd formaat om receptieve woordenschatkennis te meten (Meara en Buxton 1987). De taalleerders krijgen een lijst met woorden uit een bepaalde frequentieband voorgelegd en dienen met Ja of Nee aan te duiden of ze de betekenis van de woorden kennen. Om overschatting of oneerlijk antwoordgedrag tegen te gaan, worden er ook pseudowoorden opgenomen in de lijst. Dit zijn niet-bestaande woorden die wel de fonotactische regels van de doeltaal volgen. Aan de hand van het aantal Ja antwoorden bij pseudowoorden (in de literatuur *false alarm*s genoemd) wordt de algemene toetsscore verlaagd.

De *Yes/No Vocabulary Test* is vooral aangewend om de receptieve woordenschatkennis van het Engels te meten van leerders met verschillende taalachtergronden. Ondanks het feit dat deze toets internationaal wordt gebruikt, werd het formaat nauwelijks gevalideerd. In deze studie wordt nagegaan of de toets een betrouwbare en valide meting oplevert van de receptieve woordenschatkennis van de basiswoordenschat Nederlands van Franstalige leerders. Een eerste analyse brengt een *response bias* aan het licht. De participanten vertonen een vertekend antwoordgedrag dat de validiteit van de toets compromiteert. In een reeks experimenten worden de variabelen onderzocht die een rol spelen in de manier waarop de participant met de Ja/Nee taak omgaat. De invloed van de toetsinstructie, de computer interface en de toetsinhoud op het antwoordgedrag van de participanten wordt daarbij onderzocht. Omdat we er niet in slagen om de *response bias* terug te dringen en de toets op een aanvaardbare manier te valideren, stellen we een nieuwe toetsformaat voor dat erop gericht is de *response bias*-problematiek te omzeilen: de *Recognition Based Vocabulary Test*.

De resultaten van dit onderzoek zijn relevant voor toetsontwikkelaars en iedereen die zich bezighoudt met het evalueren van taalvaardigheid. Een grondig gevalideerde woordenschattoets kan gebruikt worden om de algemene woordenschatomvang van leerders te meten of hun kennis van een bepaald woordenschatdomein te bepalen. Het ontwikkelen van een betrouwbaar en valide instrument om woordenschatkennis in kaart te brengen is ook van belang voor taalverwervingsonderzoekers en cursusontwikkelaars omdat het tot inzichten kan leiden in hoe leerders woordenschat verwerven en met welke snelheid deze lexicale ontwikkeling plaatsgrijpt. Als niveautoets kan een dergelijk instrument bovendien aangewend worden om op een snelle en efficiënte manier cursisten in de geschikte taalcursussen te plaatsen.

Na een algemene inleiding in Hoofdstuk 1 waarin het onderzoek wordt gesitueerd, beschrijft Hoofdstuk 2 de centrale rol die het lexicon speelt bij vreemdetaalverwerving. Er wordt ingegaan op het belang van het meten van woordenschatkennis, meerbepaald het bepalen van de woordenschatomvang van vreemdetaalleerders. In Hoofdstuk 3 wordt de *Yes/No Vocabulary Test* voorgesteld. De onstaansgeschiedenis van de toets komt aan bod en er wordt verslag uitgebracht van de in de literatuur gerapporteerde voor- en nadelen en de factoren die een rol spelen bij het maken en afleggen van de toets. In Hoofdstuk 4 wordt het reilen en zeilen van het taleninstituut van de Université Libre de Bruxelles beschreven, waar de *Yes/No Vocabulary Test* wordt gebruikt in de plaatsingsprocedure als aanvulling op een meerkeuze grammaticatoets en waar alle data uit dit proefschrift werden vergaard. Er wordt bijzonder belang gehecht aan het pragmatische kader waaruit dit onderzoek ontstond en de studentenpopulatie wordt gekarakteriseerd.

Hoofdstuk 5 brengt verslag uit van het eerste gebruik van de *Yes/No Vocabulary Test*. Het grote aantal *false alarm*- antwoorden dat we aantreffen in bij de verwerking van de data zorgt voor problemen bij het kiezen van de meest geschikte correctieformule en doet vragen rijzen bij de validiteit van de toets. Een aanzienlijk deel van de studenten beweert immers de betekenis te kennen van een groot aantal pseudowoorden in de toets. Een herevaluatie van de psychometrische kwaliteiten van de toets dringt zich dan ook op. De verschillende correctieformules die in de literatuur werden voorgesteld worden toegelicht en de betrouwbaarheid van de toets wordt beschouwd op zowel theoretische als empirische gronden. We laten zien dat de resultaten vertekend zijn en dat de verschillende correctieformules uiteenlopende resultaten opleveren naarmate het aantal *false alarm*s stijgt. Wanneer de *response bias* uit de data wordt geëxtraheerd, daalt de betrouwbaarheid van de toets zienderogen, wat het vermoeden doet rijzen dat de betrouwbaarheden die worden vermeld in andere studies overschat zijn omwille van de aanwezigheid van een *response bias* in de data.

In een tweede experiment (Hoofdstuk 6) wordt gepoogd de validiteit van de *Yes/No Vocabulary Test* te onderzoeken om op basis daarvan de meest geschikte correctieformule te selecteren. Aan de hand van een vertaaltaak wordt nagegaan hoe valide de Ja/Nee antwoorden van de participanten zijn: kunnen ze een equivalent verstrekken in hun moedertaal van de woorden die ze met Ja hebben beantwoord in de *Yes/No Vocabulary Test*? De centrale doelstelling van het experiment is de invloed te onderzoeken van de verschillende correctieformules op de correlatie tussen de Ja/Nee toetsscores en de Vertaalscores. De empirische gegevens bevestigen de resultaten van het eerste experiment: opnieuw worden er veel *false alarm*-antwoorden aangetroffen en is er sprake van een *response bias* in de data. Als gevolg hiervan blijft de beoogde validering uit: er is een lage correlatie tussen de Ja/Nee scores en de Vertaalscores. Het wordt duidelijk dat de *Yes/No Vocabulary Test* iets meet dat

geen verband heeft met de woordenschatkkennis van de participanten maar eerder de interactie betreft van hun metacognitief en/of sociocultureel profiel met de taak die afgelegd dient te worden.

In Hoofdstuk 7 wordt verslag uitgebracht van twee experimenten die tot doel hebben de variabelen die verantwoordelijk zouden kunnen zijn voor de hoge *false alarm*-score en de *response bias* te isoleren en te onderzoeken wat de invloed is van deze variabelen op de validiteit van de *Yes/No Vocabulary Test*. Het eerste experiment bestudeert de invloed van de opgave op het antwoordgedrag van de participanten in de *Yes/No Vocabulary Test*. Twee groepen krijgen dezelfde *Yes/No Vocabulary Test* aangeboden. Bij de experimentele groep wordt de toets vergezeld van een zeer rigoureuze instructie waarbij vooraf reeds wordt aangekondigd dat de waarachtigheid van de antwoorden achteraf zal worden nagegaan. De controlegroep daarentegen krijgt een *Yes/No Vocabulary Test* met een vage, neutrale opgave. Beide groepen worden achteraf gevraagd een Vertaaltoets te maken die bestaat uit dezelfde woorden als de *Yes/No Vocabulary Test*. De resultaten tonen aan dat het gebruik van een rigoureuze instructie de participanten aanzet tot voorzichtiger antwoordgedrag: de *false alarm*-score is significant lager bij de experimentele groep. Dit resulteert echter niet in een betere validering: de correlatie tussen Ja/Nee scores en Vertaalscores is niet hoger bij de experimentele groep dan bij de controlegroep. Het beïnvloeden van het antwoordgedrag van participanten leidt niet noodzakelijkerwijs tot een meer valide meting van hun woordenschatkennis.

Een volgend experiment is gericht op de mogelijke invloed van de computerinterface op het antwoordgedrag van de participanten. Twee radicaal verschillende computerinterfaces worden geprogrammeerd, waarbij de ene zo getrouw mogelijk een "pen-en-papier" toets weerspiegelt (de items staan in een lijst, de participanten kunnen kiezen in welke volgorde ze de items beantwoorden, enz.) en de andere gebruik maakt van de mogelijkheden van de computer om zoveel mogelijk controle in te bouwen bij het afleggen van de toets (de items verschijnen één voor één, de participanten kunnen niet terugkeren naar een reeds beantwoord item enz.). De hypothese dat een meer gecontroleerde interface de *response bias* van de participanten zou reduceren of elimineren wordt echter ontkracht door de resultaten.

Om de mogelijke kritiek te weerleggen dat de verontrustende data die we verzamelden met de *Yes/No Vocabulary Test* te wijten zouden zijn aan de toetsinhoud die we zelf hadden opgesteld, worden er in Hoofdstuk 8 twee experimenten uitgevoerd met de toetsinhoud van het Europese DIALANG project. In een eerste experiment wordt de toets voorgelegd aan Franstalige participanten. De psychometrische kwaliteiten van de toets blijken niet te zijn verbeterd: opnieuw zorgen een hoge *false alarm*-score en de aanwezigheid van een *response bias* voor een problematische correctie van de ruwe score en een manifest validiteitsprobleem. Uit het experiment met moedertaalsprekers

Nederlands wordt de twijfelachtige kwaliteit van de geselecteerde woorden en pseudowoorden duidelijk. Voor de Nederlandstalige participanten is het nog moeilijker de bestaande woorden aan te duiden dan de pseudowoorden te verwerpen. Tot slot wordt besloten dat geen van de pogingen om de *Yes/No Vocabulary Test* te verbeteren vruchten heeft afgeworpen omdat de Ja/Nee taak zelf te gevoelig is aan interactie met andere factoren die de te maken hebben met de taallleerder of met de context waarin de toets wordt afgenomen. Het is bijgevolg geen geschikt formaat om doorheen talen en culturen te gebruiken.

In Hoofdstuk 9 wordt een alternatief toetsformaat voorgesteld, de *Recognition Based Vocabulary Test* (RBVT). In deze toets kan geen *response bias* optreden omdat de detectietaak van de Ja/Nee Toets wordt omgezet in een discriminatietaak waarbij de participant in een item-paar dat bestaat uit een woord en een pseudowoord het bestaande woord moet aanduiden. In de RBVT 1 zijn de woorden en pseudowoorden willekeurig aan elkaar gepaard, in de RBVT 2 zijn de item-paren minimale paren, wat wil zeggen dat het pseudowoord wordt afgeleid van het woord waarmee het wordt gecombineerd. Beide formaten worden in een experiment vergeleken met de *Yes/No Vocabulary Test* en gevalideerd aan de hand van een Vertaaltoets. Door een gebrek aan discriminerende items in beide RBVT-formaten is een gedegen validering van de RBVT-toetsen niet mogelijk omdat de correlatie tussen de RBVT-formaten en de Vertaaltaak nadelig beïnvloed wordt door de matige betrouwbaarheid van de toetsen. De *Yes/No Vocabulary Test* wordt opnieuw gekenmerkt door een hoge *false alarm*-score en de nefaste gevolgen hiervan voor de betrouwbaarheid en validiteit van de toets.

Tot slot wordt in Hoofdstuk 10 besloten dat de *Recognition Based Vocabulary Test* een veelbelovender toetsformaat is dan de *Yes/No Vocabulary Test* om de receptieve woordenschatkennis op een betrouwbare en valide manier te meten. In de experimenten is immers keer op keer aangetoond dat in het gebruik van de *Yes/No Vocabulary Test* met Franstalige leerders van het Nederlands de hoge *false alarm*-scores een indicatie zijn van het bestaan van een *response bias* in de data, wat problemen veroorzaakt bij het toekennen van een representatieve score. Deze *response bias* wijst erop dat de toets andere variabelen meet dan de woordenschatkennis van de participanten. Deze variabelen betreffen vermoedelijk de metacognitieve en socioculturele kenmerken van de taalleerder en kunnen erg verschillen naargelang de context of het land waar de toets wordt afgenomen. Het is duidelijk dat we hier te maken hebben met construct-irrelevante variabelen die een representatieve meting van de woordenschatkennis verstoren. Bovendien neemt deze vertekening van de antwoorden zulke proporties aan dat de psychometrische kwaliteiten van de toets in termen van betrouwbaarheid niet gegarandeerd zijn en dat de validiteit van de toets twijfelachtig wordt. Het toetsformaat is ons inziens te weinig robust om gebruikt te worden als gestandaardiseerde woordenschattoets. Dit is vooral te wijten aan de Ja/Nee dichotomie die aan de basis ligt van *The Yes/No*

Vocabulary Test en die een *response bias* ontlokt. In de *Recognition Based Vocabulary Test* is dit probleem niet aanwezig, maar ook dit toetsformaat dient verder gevalideerd te worden om het te kunnen gebruiken voor het meten van receptieve woordenschatkennis.

Curriculum vitae

June Eyckmans was born in Mechelen (Belgium) on February 28th, 1972. She took her Master's degree in Germanic Languages at the Vrije Universiteit Brussel. She started her professional career at the Institut de Langues Vivantes et de Phonétique of the Université Libre de Bruxelles where she taught Dutch to French-speaking students for seven years. It was there that her research interest in vocabulary acquisition and testing originated. In 2001, she started to work in the department of applied linguistics of the Erasmuscollege of Brussels where she continued her research. She has published in the field of language testing and applied linguistics. Thanks to a grant of the Katholieke Universiteit Nijmegen she was able to finalize her PHD-dissertation.